Loss Functions for Top-k Error: Analysis and Insights

Maksim Lapin,[1] Matthias Hein[2] and Bernt Schiele[1]
[1]Max Planck Institute for Informatics, Saarbrücken, Germany
[2]Saarland University, Saarbrücken, Germany

Abstract

In order to push the performance on realistic computer vision tasks, the number of classes in modern benchmark datasets has significantly increased in recent years. This increase in the number of classes comes along with increased ambiguity between the class labels, raising the question if top-1 error is the right performance measure. In this paper, we provide an extensive comparison and evaluation of established multiclass methods comparing their top-k performance both from a practical as well as from a theoretical perspective. Moreover, we introduce novel top-k loss functions as modifications of the softmax and the multiclass SVM losses and provide efficient optimization schemes for them. In the experiments, we compare on various datasets all of the proposed and established methods for top-k error optimization. An interesting insight of this paper is that the softmax loss yields competitive top-k performance for all k simultaneously. For a specific top-k error, our new top-k losses lead typically to further improvements while being faster to train than the softmax.

1. Introduction

The number of classes is rapidly growing in modern computer vision benchmarks [47, 62]. Typically, this also leads to ambiguity in the labels as classes start to overlap. Even for humans, the error rates in top-1 performance are often quite high ($\approx 30\%$ on SUN 397 [60]). While previous research focuses on minimizing the top-1 error, we address top-k error optimization in this paper. We are interested in two cases: a) achieving small top-k error for *all* reasonably small k; and b) minimization of a specific top-k error.

While it is argued in [2] that the one-versus-all (OVA) SVM scheme performs on par in top-1 and top-5 accuracy with the other SVM variations based on ranking losses, we have recently shown in [28] that minimization of the *top-k hinge loss* leads to improvements in top-k performance compared to OVA SVM, multiclass SVM, and other ranking-based formulations. In this paper, we study top-k error optimization from a wider perspective. On the one hand, we compare OVA schemes and direct multiclass losses in extensive experiments, and on the other, we present theoretical discussion regarding their calibration for the top-k error. Based on these insights, we suggest 4 new families of loss functions for the top-k error. Two are smoothed versions of the top-k hinge losses [28], and the other two are top-k versions of the softmax loss. We discuss their advantages and disadvantages, and for the convex losses provide an efficient implementation based on stochastic dual coordinate ascent (SDCA) [48].

We evaluate a battery of loss functions on 11 datasets of different tasks ranging from text classification to large scale vision benchmarks, including fine-grained and scene classification. We systematically optimize and report results separately for each top-k accuracy. One interesting message that we would like to highlight is that the softmax loss is able to optimize *all top-k error measures simultaneously*. This is in contrast to multiclass SVM and is also reflected in our experiments. Finally, we show that our new top-k variants of smooth multiclass SVM and the softmax loss can further improve top-k performance for a specific k.

Related work. Top-k optimization has recently received revived attention with the advent of large scale problems [20, 28, 30, 31]. The top-k error in multiclass classification, which promotes good ranking of class *labels* for each example, is closely related to the precision@k metric in information retrieval, which counts the fraction of positive instances among the top-k ranked *examples*. In essence, both approaches enforce a desirable *ranking* of items [28].

The classic approaches optimize pairwise ranking with $\mathrm{SVM}^{\mathrm{struct}}$ [25, 53], RankNet [11], and LaRank [7]. An alternative direction was proposed by Usunier *et al*. [54], who described a general family of convex loss functions for ranking and classification. One of the loss functions that we consider (top-k SVM^{β} [28]) also falls into that family. Weston *et al*. [59] then introduced Wsabie, which optimizes an approximation of a ranking-based loss from [54]. A Bayesian approach was suggested by [51].

Recent works focus on the top of the ranked list [1, 9, 39, 46], scalability to large datasets [20, 28, 30], explore transductive learning [31] and prediction of tuples [45].

Method	Name	Loss function	Conjugate	SDCA update	Top-k calibrated	Convex
SVM^{OVA}	One-vs-all (OVA) SVM	$\max\{0, 1 - a\}$	[48]	[48]	no[1] (Prop. 1)	yes
LR^{OVA}	OVA logistic regression	$\log(1 + e^{-a})$			yes (Prop. 2)	
$\text{SVM}^{\text{Multi}}$	Multiclass SVM	$\max\left\{0, (a + c)_{\pi_1}\right\}$	[28, 48]	[28, 48]	no (Prop. 3)	
LR^{Multi}	Softmax (maximum entropy)	$\log\left(\sum_{j \in \mathcal{Y}} \exp(a_j)\right)$	Prop. 8	Prop. 12	yes (Prop. 4)	
top-k SVM^{α}	Top-k hinge (α)	$\max\left\{0, \frac{1}{k} \sum_{j=1}^{k} (a + c)_{\pi_j}\right\}$	[28]	[28]	open question for $k > 1$	
top-k SVM^{β}	Top-k hinge (β)	$\frac{1}{k} \sum_{j=1}^{k} \max\left\{0, (a + c)_{\pi_j}\right\}$				
top-k $\text{SVM}^{\alpha}_{\gamma}$	Smooth top-k hinge (α) *	Eq. (12) w/ Δ^{α}_{k}	Prop. 7	Prop. 11		
top-k $\text{SVM}^{\beta}_{\gamma}$	Smooth top-k hinge (β) *	Eq. (12) w/ Δ^{β}_{k}				
top-k Ent	Top-k entropy *	Prop. 9	Eq. (14)	Prop. 12		
top-k Ent_{tr}	Truncated top-k entropy *	Eq. (22)	-	-	yes (Prop. 10)	no

Note that $\text{SVM}^{\text{Multi}} \equiv$ top-1 $\text{SVM}^{\alpha} \equiv$ top-1 SVM^{β} and $\text{LR}^{\text{Multi}} \equiv$ top-1 Ent $\equiv$ top-1 Ent_{tr}.

We let $a \triangleq y f(x)$ (binary one-vs-all); $a \triangleq (f_j(x) - f_y(x))_{j \in \mathcal{Y}}$, $c \triangleq \mathbf{1} - e_y$ (multiclass); $\pi : a_{\pi_1} \geq \ldots \geq a_{\pi_m}$.

Table 1: Overview of the methods we consider and our contributions. *Novel loss. [1]But *smoothed* one is (Prop. 5).

Contributions. We study the problem of top-k error optimization on a diverse range of learning tasks. We consider existing methods as well as propose 4 novel loss functions for minimizing the top-k error. A brief overview of the methods is given in Table 1. For the proposed convex top-k losses, we develop an efficient optimization scheme based on SDCA[1], which can also be used for training with the softmax loss. All methods are evaluated empirically in terms of the top-k error and, whenever possible, in terms of classification calibration. We discover that the softmax loss and the proposed smooth top-1 SVM are astonishingly competitive in all top-k errors. Further small improvements can be obtained with the new top-k losses.

2. Loss Functions for Top-k Error

We consider multiclass problems with m classes where the training set $(x_i, y_i)_{i=1}^n$ consists of n examples $x_i \in \mathbb{R}^d$ along with the corresponding labels $y_i \in \mathcal{Y} \triangleq \{1, \ldots, m\}$. We use π and τ to denote a permutation of (indexes) $\mathcal{Y}$. Unless stated otherwise, a_π reorders components of a vector $a \in \mathbb{R}^m$ in descending order, *i.e.* $a_{\pi_1} \geq a_{\pi_2} \geq \ldots \geq a_{\pi_m}$. While we consider linear classifiers in our experiments, all loss functions below are formulated in the general setting where a function $f : \mathcal{X} \to \mathbb{R}^m$ is learned and prediction at test time is done via $\arg\max_{y \in \mathcal{Y}} f_y(x)$, resp. the top-$k$ predictions. For the linear case, all predictors f_y have the form $f_y(x) = \langle w_y, x \rangle$. Let $W \in \mathbb{R}^{d \times m}$ be the stacked weight matrix, $L : \mathcal{Y} \times \mathbb{R}^m \to \mathbb{R}$ be a convex loss function, and $\lambda > 0$ be a regularization parameter. We consider the following multiclass optimization problem $\min_W \frac{1}{n} \sum_{i=1}^n L(y_i, W^\top x_i) + \lambda \|W\|_F^2$.

[1] Code available at: https://github.com/mlapin/libsdca

We use the Iverson bracket notation $[\![P]\!]$, defined as $[\![P]\!] = 1$ if P is true, 0 otherwise; and introduce a shorthand $p_y(x) \triangleq \Pr(Y = y \,|\, X = x)$. We generalize the standard zero-one error and allow k guesses instead of one. Formally, the **top-k zero-one loss** (**top-k error**) is

$$\text{err}_k(y, f(x)) \triangleq [\![f_{\pi_k}(x) > f_y(x)]\!]. \tag{1}$$

Note that for $k = 1$ we recover the standard zero-one error. **Top-k accuracy** is defined as 1 minus the top-k error.

2.1. Bayes Optimality and Top-k Calibration

In this section, we establish the best achievable top-k error, determine when a classifier achieves it, and define a notion of top-k calibration.

Lemma 1. *The Bayes optimal top-k error at x is*

$$\min_{g \in \mathbb{R}^m} \mathbb{E}_{Y|X}[\text{err}_k(Y, g) \mid X = x] = 1 - \textstyle\sum_{j=1}^k p_{\tau_j}(x),$$

*where $p_{\tau_1}(x) \geq p_{\tau_2}(x) \geq \ldots \geq p_{\tau_m}(x)$. A classifier f is **top-k Bayes optimal at** x if and only if*

$$\left\{y \,|\, f_y(x) \geq f_{\pi_k}(x)\right\} \subset \left\{y \,|\, p_y(x) \geq p_{\tau_k}(x)\right\},$$

where $f_{\pi_1}(x) \geq f_{\pi_2}(x) \geq \ldots \geq f_{\pi_m}(x)$.

Proof. Let $g \in \mathbb{R}^m$ and π be a permutation such that $g_{\pi_1} \geq g_{\pi_2} \geq \ldots \geq g_{\pi_m}$. The expected top-$k$ error at x is

$$\begin{aligned} &\mathbb{E}_{Y|X}[\text{err}_k(Y, g) \mid X = x] = \textstyle\sum_{y \in \mathcal{Y}} [\![g_{\pi_k} > g_y]\!] p_y(x) \\ &= \textstyle\sum_{y \in \mathcal{Y}} [\![g_{\pi_k} > g_{\pi_y}]\!] p_{\pi_y}(x) = \sum_{j=k+1}^m p_{\pi_j}(x) \\ &= 1 - \textstyle\sum_{j=1}^k p_{\pi_j}(x). \end{aligned}$$

The error is minimal when $\sum_{j=1}^{k} p_{\pi_j}(x)$ is maximal, which corresponds to taking the k largest conditional probabilities $\sum_{j=1}^{k} p_{\tau_j}(x)$ and yields the Bayes optimal top-k error at x.

Since the relative order within $\{p_{\tau_j}(x)\}_{j=1}^{k}$ is irrelevant for the top-k error, any classifier $f(x)$, for which the sets $\{\pi_1, \ldots, \pi_k\}$ and $\{\tau_1, \ldots, \tau_k\}$ coincide, is Bayes optimal.

Note that we assumed w.l.o.g. that there is a clear cut $p_{\tau_k}(x) > p_{\tau_{k+1}}(x)$ between the k most likely classes and the rest. In general, ties can be resolved arbitrarily as long as we can guarantee that the k largest components of $f(x)$ correspond to the classes (indexes) that yield the maximal sum $\sum_{j=1}^{k} p_{\pi_j}(x)$ and lead to top-k Bayes optimality. □

Optimization of the zero-one loss (and, by extension, the top-k error) leads to hard combinatorial problems. Instead, a standard approach is to use a convex surrogate loss which upper bounds the zero-one error. Under mild conditions on the loss function [3, 52], the optimal classifier w.r.t. the surrogate yields a Bayes optimal solution for the zero-one loss. Such loss is called *classification calibrated*, which is known in statistical learning theory as a necessary condition for a classifier to be universally Bayes consistent [3]. We introduce now the notion of calibration for the top-k error.

Definition 1. *A loss function $L : \mathcal{Y} \times \mathbb{R}^m \to \mathbb{R}$ (or a reduction scheme) is called **top-k calibrated** if for all possible data generating measures on $\mathbb{R}^d \times \mathcal{Y}$ and all $x \in \mathbb{R}^d$*

$$\begin{aligned} &\arg\min\nolimits_{g \in \mathbb{R}^m} \mathbb{E}_{Y|X}[L(Y, g) \,|\, X = x] \\ &\subseteq \arg\min\nolimits_{g \in \mathbb{R}^m} \mathbb{E}_{Y|X}[\mathrm{err}_k(Y, g) \,|\, X = x]. \end{aligned}$$

If a loss is *not* top-k calibrated, it implies that even in the limit of infinite data, one does not obtain a classifier with the Bayes optimal top-k error from Lemma 1.

2.2. OVA and Direct Multiclass Approaches

The standard multiclass problem is often solved using the one-vs-all (OVA) reduction into a set of m binary classification problems. Every class is trained versus the rest which yields m classifiers $\{f_y\}_{y \in \mathcal{Y}}$.

Typically, the binary classification problems are formulated with a convex margin-based loss function $L(yf(x))$, where $L : \mathbb{R} \to \mathbb{R}$ and $y = \pm 1$. We consider in this paper:

$$L(yf(x)) = \max\{0,\, 1 - yf(x)\}, \tag{2}$$

$$L(yf(x)) = \log(1 + e^{-yf(x)}). \tag{3}$$

The **hinge** (2) and **logistic** (3) losses correspond to the SVM and logistic regression respectively. We now show when the OVA schemes are top-k calibrated, not only for $k = 1$ (standard multiclass loss) but *for all k* simultaneously.

Lemma 2. *The OVA reduction is top-k calibrated for any $1 \leq k \leq m$ if the Bayes optimal function of the convex margin-based loss $L(yf(x))$ is a strictly monotonically increasing function of $\Pr(Y = 1 \,|\, X = x)$.*

Proof. For every class $y \in \mathcal{Y}$, the Bayes optimal classifier for the corresponding binary problem has the form

$$f_y(x) = g\big(\Pr(Y = y \,|\, X = x)\big),$$

where g is a strictly monotonically increasing function. The ranking of f_y corresponds to the ranking of $\Pr(Y = y \,|\, X = x)$ and hence the OVA reduction is top-k calibrated for any $k = 1, \ldots, m$. □

Next, we check if the one-vs-all schemes employing hinge and logistic regression losses are top-k calibrated.

Proposition 1. *OVA SVM is not top-k calibrated.*

Proof. First, we show that the Bayes optimal function for the binary hinge loss is

$$f^*(x) = 2[\![\Pr(Y = 1 \,|\, X = x) > \tfrac{1}{2}]\!] - 1.$$

We decompose the expected loss as

$$\mathbb{E}_{X,Y}[L(Y, f(X))] = \mathbb{E}_X[\mathbb{E}_{Y|X}[L(Y, f(x)) \,|\, X = x]].$$

Thus, one can compute the Bayes optimal classifier f^* pointwise by solving

$$\arg\min_{\alpha \in \mathbb{R}} \mathbb{E}_{Y|X}[L(Y, \alpha) \,|\, X = x],$$

for every $x \in \mathbb{R}^d$, which leads to the following problem

$$\arg\min_{\alpha \in \mathbb{R}} \; \max\{0, 1 - \alpha\} p_1(x) + \max\{0, 1 + \alpha\} p_{-1}(x),$$

where $p_y(x) \triangleq \Pr(Y = y \,|\, X = x)$. It is obvious that the optimal α^* is contained in $[-1, 1]$. We get

$$\arg\min_{-1 \leq \alpha \leq 1} \; (1 - \alpha) p_1(x) + (1 + \alpha) p_{-1}(x).$$

The minimum is attained at the boundary and we get

$$f^*(x) = \begin{cases} +1 & \text{if } p_1(x) > \frac{1}{2}, \\ -1 & \text{if } p_1(x) \leq \frac{1}{2}. \end{cases}$$

Therefore, the Bayes optimal classifier for the hinge loss is not a strictly monotonically increasing function of $p_1(x)$.

To show that OVA hinge is not top-k calibrated, we construct an example problem with 3 classes and $p_1(x) = 0.4$, $p_2(x) = p_3(x) = 0.3$. Note that for every class $y = 1, 2, 3$, the Bayes optimal binary classifier is -1, hence the predicted ranking of labels is arbitrary and may not produce the Bayes optimal top-k error. □

In contrast, logistic regression is top-k calibrated.

Proposition 2. *OVA logistic regression is top-k calibrated.*

Proof. First, we show that the Bayes optimal function for the binary logistic loss is

$$f^*(x) = \log\Big(\frac{p_x(1)}{1 - p_x(1)}\Big).$$

As above, the pointwise optimization problem is

$$\arg\min_{\alpha \in \mathbb{R}} \; \log(1+\exp(-\alpha))p_1(x) + \log(1+\exp(\alpha))p_{-1}(x).$$

The logistic loss is known to be convex and differentiable and thus the optimum can be computed via

$$\frac{-\exp(-\alpha)}{1+\exp(-\alpha)} p_1(x) + \frac{\exp(\alpha)}{1+\exp(\alpha)} p_{-1}(x) = 0.$$

Re-writing the first fraction we get

$$\frac{-1}{1+\exp(\alpha)} p_1(x) + \frac{\exp(\alpha)}{1+\exp(\alpha)} p_{-1}(x) = 0,$$

which can be solved as $\alpha^* = \log\Big(\frac{p_1(x)}{p_{-1}(x)}\Big)$ and leads to the formula for the Bayes optimal classifier stated above.

We check now that the function $\phi : (0,1) \to \mathbb{R}$ defined as $\phi(x) = \log(\frac{x}{1-x})$ is strictly monotonically increasing.

$$\begin{aligned}\phi'(x) &= \frac{1-x}{x}\Big(\frac{1}{1-x} + \frac{x}{(1-x)^2}\Big) \\ &= \frac{1-x}{x}\frac{1}{(1-x)^2} = \frac{1}{x(1-x)} > 0, \quad \forall x \in (0,1).\end{aligned}$$

The derivative is strictly positive on $(0,1)$, which implies that ϕ is strictly monotonically increasing. The logistic loss, therefore, fulfills the conditions of Lemma 2 and is top-k calibrated for any $1 \le k \le m$. ☐

An alternative to the OVA scheme with binary losses is to use a *multiclass* loss $L : \mathcal{Y} \times \mathbb{R}^m \to \mathbb{R}$ directly. We consider two generalizations of the hinge and logistic losses below:

$$L(y, f(x)) = \max_{j \in \mathcal{Y}} \big\{ [\![j \neq y]\!] + f_j(x) - f_y(x) \big\}, \tag{4}$$

$$L(y, f(x)) = \log\Big(\textstyle\sum_{j \in \mathcal{Y}} \exp(f_j(x) - f_y(x)) \Big). \tag{5}$$

Both the **multiclass hinge loss** (4) of Crammer & Singer [15] and the **softmax loss** (5) are popular losses for multiclass problems. The latter is also known as the cross-entropy or multiclass logistic loss and is often used as the last layer in deep architectures [6, 26, 50]. The multiclass hinge loss has been shown to be competitive in large-scale image classification [2], however, it is known to be not calibrated [52] for the top-1 error. Next, we show that it is not top-k calibrated for any k.

Proposition 3. *Multiclass SVM is not top-k calibrated.*

Proof. First, we derive the Bayes optimal function.

Let $y \in \arg\max_{j \in \mathcal{Y}} p_j(x)$. Given any $c \in \mathbb{R}$, a Bayes optimal function $f^* : \mathbb{R}^d \to \mathbb{R}^m$ for the loss (4) is

$$\begin{aligned} f^*_y(x) &= \begin{cases} c+1 & \text{if } \max_{j \in \mathcal{Y}} p_j(x) \ge \frac{1}{2}, \\ c & \text{otherwise,} \end{cases} \\ f^*_j(x) &= c, \; j \in \mathcal{Y} \setminus \{y\}. \end{aligned}$$

Let $g = f(x) \in \mathbb{R}^m$, then

$$\mathbb{E}_{Y|X}[L(Y,g) \,|\, X] = \sum_{l \in \mathcal{Y}} \max_{j \in \mathcal{Y}} \big\{ [\![j \neq l]\!] + g_j - g_l \big\} p_l(x).$$

Suppose that the maximum of $(g_j)_{j \in \mathcal{Y}}$ is not unique. In this case, we have

$$\max_{j \in \mathcal{Y}} \big\{ [\![j \neq l]\!] + g_j - g_l \big\} \ge 1, \; \forall\, l \in \mathcal{Y}$$

as the term $[\![j \neq l]\!]$ is always active. The best possible loss is obtained by setting $g_j = c$ for all $j \in \mathcal{Y}$, which yields an expected loss of 1. On the other hand, if the maximum is unique and is achieved by g_y, then

$$\begin{aligned} &\max_{j \in \mathcal{Y}} \big\{ [\![j \neq l]\!] + g_j - g_l \big\} \\ &= \begin{cases} 1 + g_y - g_l & \text{if } l \neq y, \\ \max\big\{0, \max_{j \neq y}\{1 + g_j - g_y\}\big\} & \text{if } l = y. \end{cases} \end{aligned}$$

As the loss only depends on the gap $g_y - g_l$, we can optimize this with $\beta_l = g_y - g_l$.

$$\begin{aligned} &\mathbb{E}_{Y|X}[L(Y,g) \,|\, X = x] \\ &= \sum_{l \neq y} (1 + g_y - g_l) p_l(x) \\ &+ \max\big\{0, \max_{l \neq y}\{1 + g_l - g_y\}\big\} p_y(x) \\ &= \sum_{l \neq y} (1 + \beta_l) p_l(x) + \max\big\{0, \max_{l \neq y}\{1 - \beta_l\}\big\} p_y(x) \\ &= \sum_{l \neq y} (1 + \beta_l) p_l(x) + \max\{0, 1 - \min_{l \neq y} \beta_l\} p_y(x). \end{aligned}$$

As only the minimal β_l enters the last term, the optimum is achieved if all β_l are equal for $l \neq y$ (otherwise it is possible to reduce the first term without affecting the last term). Let $\alpha \triangleq \beta_l$ for all $l \neq y$. The problem becomes

$$\begin{aligned} &\min_{\alpha \ge 0} \sum_{l \neq y} (1 + \alpha) p_l(x) + \max\{0, 1 - \alpha\} p_y(x) \\ &\equiv \min_{0 \le \alpha \le 1} \alpha(1 - 2p_y(x)) \end{aligned}$$

Let $p \triangleq p_y(x) = \Pr(Y = y \,|\, X = x)$. The solution is

$$\alpha^* = \begin{cases} 0 & \text{if } p < \frac{1}{2}, \\ 1 & \text{if } p \ge \frac{1}{2}, \end{cases}$$

and the associated risk is

$$\mathbb{E}_{Y|X}[L(Y,g) \mid X=x] = \begin{cases} 1 & \text{if } p < \frac{1}{2}, \\ 2(1-p) & \text{if } p \geq \frac{1}{2}. \end{cases}$$

If $p < \frac{1}{2}$, then the Bayes optimal classifier $f_j^*(x) = c$ for all $j \in \mathcal{Y}$ and any $c \in \mathbb{R}$. Otherwise, $p \geq \frac{1}{2}$ and

$$f_j^*(x) = \begin{cases} c+1 & \text{if } j = y, \\ c & \text{if } j \in \mathcal{Y} \setminus \{y\}. \end{cases}$$

Moreover, we have that the Bayes risk at x is

$$\mathbb{E}_{Y|X}[L(Y, f^*(x)) \mid X = x] = \min\{1, 2(1-p)\} \leq 1.$$

It follows, that the multiclass hinge loss is not (top-1) classification calibrated at any x where $\max_{y \in \mathcal{Y}} p_y(x) < \frac{1}{2}$ as its Bayes optimal classifier reduces to a constant. Moreover, even if $p_y(x) \geq \frac{1}{2}$ for some y, the loss is not top-k calibrated for $k \geq 2$ as the predicted order of the remaining classes need not be optimal. □

Again, a contrast between the hinge and logistic losses.

Proposition 4. *The softmax loss is top-k calibrated.*

Proof. The multiclass logistic loss is (top-1) calibrated for the zero-one error in the following sense. If

$$f^*(x) \in \operatorname*{arg\,min}_{g \in \mathbb{R}^m} \mathbb{E}_{Y|X}[L(Y,g) \mid X = x],$$

then for some $\alpha > 0$ and all $y \in \mathcal{Y}$

$$f_y^*(x) = \begin{cases} \log(\alpha\, p_y(x)) & \text{if } p_y(x) > 0, \\ -\infty & \text{otherwise,} \end{cases}$$

which implies

$$\operatorname*{arg\,max}_{y \in \mathcal{Y}} f_y^*(x) = \operatorname*{arg\,max}_{y \in \mathcal{Y}} \Pr(Y = y \mid X = x).$$

We now prove this result and show that it also generalizes to top-k calibration for $k > 1$. Using the identity

$$L(y,g) = \log\big(\textstyle\sum_{j \in \mathcal{Y}} e^{g_j - g_y}\big) = \log\big(\textstyle\sum_{j \in \mathcal{Y}} e^{g_j}\big) - g_y$$

and the fact that $\sum_{y \in \mathcal{Y}} p_y(x) = 1$, we write for a $g \in \mathbb{R}^m$

$$\begin{aligned} &\mathbb{E}_{Y|X}[L(Y,g) \mid X = x] \\ &\quad = \sum_{y \in \mathcal{Y}} L(y,g) p_y(x) = \log\big(\sum_{y \in \mathcal{Y}} e^{g_y}\big) - \sum_{y \in \mathcal{Y}} g_y p_x(y). \end{aligned}$$

As the loss is convex and differentiable, we get the global optimum by computing a critical point. We have

$$\frac{\partial}{\partial g_j} \mathbb{E}_{Y|X}[L(Y,g) \mid X = x] = \frac{e^{g_j}}{\sum_{y \in \mathcal{Y}} e^{g_y}} - p_j(x) = 0$$

for $j \in \mathcal{Y}$. We note that the critical point is not unique as multiplication $g \to \kappa g$ leaves the equation invariant for any $\kappa > 0$. One can verify that $e^{g_j} = \alpha p_j(x)$ satisfies the equations for any $\alpha > 0$. This yields a solution

$$f_y^*(x) = \begin{cases} \log(\alpha p_y(x)) & \text{if } p_y(x) > 0, \\ -\infty & \text{otherwise,} \end{cases}$$

for any fixed $\alpha > 0$. We note that f_y^* is a strictly monotonically increasing function of the conditional class probabilities. Therefore, it preserves the ranking of $p_y(x)$ and implies that f^* is top-k calibrated for any $1 \leq k \leq m$. □

The implicit reason for top-k calibration of the OVA schemes and the softmax loss is that one can estimate the probabilities $p_y(x)$ from the Bayes optimal classifier. Loss functions which allow this are called *proper*. We refer to [41] and references therein for a detailed discussion.

We have established that the OVA logistic regression and the softmax loss are top-k calibrated for any k, so why should we be interested in defining new loss functions for the top-k error? The reason is that calibration is an asymptotic property as the Bayes optimal functions are obtained pointwise. The picture changes if we use linear classifiers, since they obviously cannot be minimized independently at each point. Indeed, most of the Bayes optimal classifiers cannot be realized by linear functions.

In particular, convexity of the softmax and multiclass hinge losses leads to phenomena where $\mathrm{err}_k(y, f(x)) = 0$, but $L(y, f(x)) \gg 0$. This happens if $f_{\pi_1}(x) \gg f_y(x) \geq f_{\pi_k}(x)$ and adds a bias when working with "rigid" function classes such as linear ones. The loss functions which we introduce in the following are modifications of the above losses with the goal of alleviating that phenomenon.

2.3. Smooth Top-k Hinge Loss

Recently, we introduced two top-k versions of the multiclass hinge loss (4) in [28], where the second version is based on the family of ranking losses introduced earlier by [54]. We use our notation from [28] for direct comparison and refer to the first version as α and the second one as β. Let $c = \mathbf{1} - e_y$, where $\mathbf{1}$ is the all ones vector, e_y is the y-th basis vector, and let $a \in \mathbb{R}^m$ be defined componentwise as $a_j \triangleq \langle w_j, x \rangle - \langle w_y, x \rangle$. The two **top-$k$ hinge losses** are

$$L(a) = \max\big\{0, \tfrac{1}{k} \textstyle\sum_{j=1}^k (a+c)_{\pi_j}\big\} \;\big(\text{top-}k \text{ SVM}^{\alpha}\big), \quad (6)$$

$$L(a) = \tfrac{1}{k} \textstyle\sum_{j=1}^k \max\big\{0, (a+c)_{\pi_j}\big\} \;\big(\text{top-}k \text{ SVM}^{\beta}\big), \quad (7)$$

where $(a)_{\pi_j}$ is the j-th largest component of a. It was shown in [28] that (6) is a tighter upper bound on the top-k error than (7), however, both losses performed similarly in our experiments. In the following, we simply refer to them as the top-k hinge or the top-k SVM loss.

Both losses reduce to the multiclass hinge loss (4) for $k = 1$. Therefore, they are unlikely to be top-k calibrated, even though we can currently neither prove nor disprove this for $k > 1$. The multiclass hinge loss is not calibrated as it is non-smooth and does not allow to estimate the class conditional probabilities $p_y(x)$. Our new family of *smooth* top-k hinge losses is based on the Moreau-Yosida regularization [5, 34]. This technique has been used in [48] to smooth the binary hinge loss (2). Interestingly, smooth binary hinge loss fulfills the conditions of Lemma 2 and leads to a top-k calibrated OVA scheme. The hope is that the smooth top-k hinge loss becomes top-k calibrated as well.

Smoothing works by adding a quadratic term to the conjugate function[2], which then becomes strongly convex. Smoothness of the loss, among other things, typically leads to much faster optimization as we discuss in Section 3.

Proposition 5. *OVA smooth hinge is top-k calibrated.*

Proof. In order to derive the smooth hinge loss, we first compute the **conjugate** of the standard **binary hinge loss**,

$$\begin{aligned} L(\alpha) &= \max\{0, 1-\alpha\}, \\ L^*(\beta) &= \sup_{\alpha \in \mathbb{R}} \{\alpha\beta - \max\{0, 1-\alpha\}\} \\ &= \begin{cases} \beta & \text{if } -1 \leq \beta \leq 0, \\ \infty & \text{otherwise.} \end{cases} \end{aligned} \tag{8}$$

The smoothed conjugate is

$$L^*_\gamma(\beta) = L^*(\beta) + \frac{\gamma}{2}\beta^2.$$

The corresponding primal **smooth hinge loss** is given by

$$\begin{aligned} L_\gamma(\alpha) &= \sup_{-1 \leq \beta \leq 0} \{\alpha\beta - \beta - \tfrac{\gamma}{2}\beta^2\} \\ &= \begin{cases} 1 - \alpha - \frac{\gamma}{2} & \text{if } \alpha < 1 - \gamma, \\ \frac{(\alpha-1)^2}{2\gamma} & \text{if } 1 - \gamma \leq \alpha \leq 1, \\ 0, & \text{if } \alpha > 1. \end{cases} \end{aligned} \tag{9}$$

$L_\gamma(\alpha)$ is convex and differentiable with the derivative

$$L'_\gamma(\alpha) = \begin{cases} -1 & \text{if } \alpha < 1 - \gamma, \\ \frac{\alpha-1}{\gamma} & \text{if } 1 - \gamma \leq \alpha \leq 1, \\ 0, & \text{if } \alpha > 1. \end{cases}$$

We compute the Bayes optimal classifier pointwise.

$$f^*(x) = \arg\min_{\alpha \in \mathbb{R}} \; L(\alpha)p_1(x) + L(-\alpha)p_{-1}(x).$$

Let $p \triangleq p_1(x)$, the optimal α^* is found by solving

$$L'(\alpha)p - L'(-\alpha)(1-p) = 0.$$

[2] The **convex conjugate** of f is $f^*(x^*) = \sup_x\{\langle x^*, x\rangle - f(x)\}$.

Case $0 < \gamma \leq 1$**.** Consider the case $1 - \gamma \leq \alpha \leq 1$,

$$\frac{\alpha - 1}{\gamma}p + (1-p) = 0 \quad \Longrightarrow \quad \alpha^* = 1 - \gamma\frac{1-p}{p}.$$

This case corresponds to $p \geq \frac{1}{2}$, which follows from the constraint $\alpha^* \geq 1 - \gamma$. Next, consider $\gamma - 1 \leq \alpha \leq 1 - \gamma$,

$$-p + (1-p) = 1 - 2p \neq 0,$$

unless $p = \frac{1}{2}$, which is already captured by the first case. Finally, consider $-1 \leq \alpha \leq \gamma - 1 \leq 1 - \gamma$. Then

$$-p - \frac{-\alpha - 1}{\gamma}(1-p) = 0 \quad \Longrightarrow \quad \alpha^* = -1 + \gamma\frac{p}{1-p},$$

where we have $-1 \leq \alpha^* \leq \gamma - 1$ if $p \leq \frac{1}{2}$. We obtain the Bayes optimal classifier for $0 < \gamma \leq 1$ as follows:

$$f^*(x) = \begin{cases} 1 - \gamma\frac{1-p}{p} & \text{if } p \geq \frac{1}{2}, \\ -1 + \gamma\frac{p}{1-p} & \text{if } p < \frac{1}{2}. \end{cases}$$

Note that while $f^*(x)$ is not a continuous function of $p = p_1(x)$ for $\gamma < 1$, it is still a strictly monotonically increasing function of p for any $0 < \gamma \leq 1$.

Case $\gamma > 1$**.** First, consider $\gamma - 1 \leq \alpha \leq 1$,

$$\frac{\alpha - 1}{\gamma}p + (1-p) = 0 \quad \Longrightarrow \quad \alpha^* = 1 - \gamma\frac{1-p}{p}.$$

From $\alpha^* \geq \gamma - 1$, we get the condition $p \geq \frac{\gamma}{2}$. Next, consider $1 - \gamma \leq \alpha \leq \gamma - 1$,

$$\frac{\alpha - 1}{\gamma}p - \frac{-\alpha - 1}{\gamma}(1-p) = 0 \quad \Longrightarrow \quad \alpha^* = 2p - 1,$$

which is in the range $[1 - \gamma, \gamma - 1]$ if $1 - \frac{\gamma}{2} \leq p \leq \frac{\gamma}{2}$. Finally, consider $-1 \leq \alpha \leq 1 - \gamma$,

$$-p - \frac{-\alpha - 1}{\gamma}(1-p) = 0 \quad \Longrightarrow \quad \alpha^* = -1 + \gamma\frac{p}{1-p},$$

where we have $-1 \leq \alpha^* \leq 1 - \gamma$ if $p \leq 1 - \frac{\gamma}{2}$. Overall, the Bayes optimal classifier for $\gamma > 1$ is

$$f^*(x) = \begin{cases} 1 - \gamma\frac{1-p}{p} & \text{if } p \geq \frac{\gamma}{2}, \\ 2p - 1 & \text{if } 1 - \frac{\gamma}{2} \leq p \leq \frac{\gamma}{2}, \\ -1 + \gamma\frac{p}{1-p} & \text{if } p < 1 - \frac{\gamma}{2}. \end{cases}$$

Note that f^* is again a strictly monotonically increasing function of $p = p_1(x)$. Therefore, for any $\gamma > 0$, the one-vs-all scheme with the smooth hinge loss (9) is top-k calibrated for all $1 \leq k \leq m$ by Lemma 2. ☐

Next, we introduce the *multiclass* smooth top-k hinge losses, which extend the top-k hinge losses (6) and (7). We define the **top-k simplex** (α and β) of radius r as

$$\Delta^\alpha_k(r) \triangleq \left\{x \mid \langle \mathbf{1}, x\rangle \leq r,\; 0 \leq x_i \leq \tfrac{1}{k}\langle \mathbf{1}, x\rangle,\; \forall i\right\}, \tag{10}$$

$$\Delta^\beta_k(r) \triangleq \left\{x \mid \langle \mathbf{1}, x\rangle \leq r,\; 0 \leq x_i \leq \tfrac{1}{k}r,\; \forall i\right\}. \tag{11}$$

We also let $\Delta^\alpha_k \triangleq \Delta^\alpha_k(1)$ and $\Delta^\beta_k \triangleq \Delta^\beta_k(1)$.

Proposition 6 ([28]). *The convex conjugates of (6) and (7) are respectively* $L^*(b) = -\langle c, b\rangle$*, if* $b \in \Delta_k^\alpha$*,* $+\infty$ *otherwise; and* $L^*(b) = -\langle c, b\rangle$*, if* $b \in \Delta_k^\beta$*,* $+\infty$ *otherwise.*

Smoothing applied to the top-k hinge loss (6) yields the following **smooth top-k hinge loss** (α). Smoothing of (7) is done similarly, but the set $\Delta_k^\alpha(r)$ is replaced with $\Delta_k^\beta(r)$.

Proposition 7. *Let* $\gamma > 0$ *be the smoothing parameter. The smooth top-k hinge loss (α) and its conjugate are*

$$L_\gamma(a) = \tfrac{1}{\gamma}\big(\langle a + c, p\rangle - \tfrac{1}{2}\langle p, p\rangle \big), \tag{12}$$

$$L_\gamma^*(b) = \tfrac{\gamma}{2}\langle b, b\rangle - \langle c, b\rangle, \text{ if } b \in \Delta_k^\alpha, +\infty \text{ o/w}, \tag{13}$$

where $p = \operatorname{proj}_{\Delta_k^\alpha(\gamma)}(a + c)$ *is the Euclidean projection of* $(a + c)$ *on* $\Delta_k^\alpha(\gamma)$*. Moreover,* $L_\gamma(a)$ *is* $1/\gamma$*-smooth.*

Proof. We take the convex conjugate of the top-k hinge loss, which was derived in [28, Proposition 2],

$$L^*(b) = \begin{cases} -\langle c, b\rangle & \text{if } b \in \Delta_k^\alpha(1), \\ +\infty & \text{otherwise,} \end{cases}$$

and add the regularizer $\frac{\gamma}{2}\langle b, b\rangle$ to obtain the γ-strongly convex conjugate loss $L_\gamma^*(b)$ as stated in the proposition. As mentioned above [21] (see also [48, Lemma 2]), the primal smooth top-k hinge loss $L_\gamma(a)$, obtained as the convex conjugate of $L_\gamma^*(b)$, is $1/\gamma$-smooth. We now obtain a formula to compute it based on the Euclidean projection onto the top-k simplex. By definition,

$$\begin{aligned} L_\gamma(a) &= \sup_{b \in \mathbb{R}^m} \{\langle a, b\rangle - L_\gamma^*(b)\} \\ &= \max_{b \in \Delta_k^\alpha(1)} \Big\{ \langle a, b\rangle - \tfrac{\gamma}{2}\langle b, b\rangle + \langle c, b\rangle \Big\} \\ &= -\min_{b \in \Delta_k^\alpha(1)} \Big\{ \tfrac{\gamma}{2}\langle b, b\rangle - \langle a + c, b\rangle \Big\} \\ &= -\tfrac{1}{\gamma} \min_{b \in \Delta_k^\alpha(1)} \Big\{ \tfrac{1}{2}\langle \gamma b, \gamma b\rangle - \langle a + c, \gamma b\rangle \Big\} \\ &= -\tfrac{1}{\gamma} \min_{\frac{b}{\gamma} \in \Delta_k^\alpha(1)} \Big\{ \tfrac{1}{2}\langle b, b\rangle - \langle a + c, b\rangle \Big\}. \end{aligned}$$

For the constraint $\frac{b}{\gamma} \in \Delta_k^\alpha(1)$, we have

$$\begin{aligned} &\langle \mathbf{1}, b/\gamma\rangle \le 1, \quad 0 \le b_i/\gamma \le \tfrac{1}{k}\langle \mathbf{1}, b/\gamma\rangle \iff \\ &\quad \langle \mathbf{1}, b\rangle \le \gamma, \quad 0 \le b_i \le \tfrac{1}{k}\langle \mathbf{1}, b\rangle \iff b \in \Delta_k^\alpha(\gamma). \end{aligned}$$

The final expression follows from the fact that

$$\begin{aligned} &\operatorname*{arg\,min}_{b \in \Delta_k^\alpha(\gamma)} \big\{ \tfrac{1}{2}\langle b, b\rangle - \langle a + c, b\rangle \big\} \\ &\equiv \operatorname*{arg\,min}_{b \in \Delta_k^\alpha(\gamma)} \|(a + c) - b\|^2 \equiv \operatorname{proj}_{\Delta_k^\alpha(\gamma)}(a + c). \end{aligned}$$

□

There is no analytic expression for (12) and evaluation requires computing a projection onto the top-k simplex $\Delta_k^\alpha(\gamma)$, which can be done in $O(m \log m)$ time as shown in [28]. The non-analytic nature of smooth top-k hinge losses currently prevents us from proving their top-k calibration.

2.4. Top-k Entropy Loss

As shown in § 4 on synthetic data, top-1 and top-2 error optimization, when limited to linear classifiers, lead to completely different solutions. The softmax loss, primarily aiming at top-1 performance, produces a solution that is reasonably good in top-1 error, but is far from what can be achieved in top-2 error. That reasoning motivated us to adapt the softmax loss to top-k error optimization. Inspired by the conjugate of the top-k hinge loss, we introduce in this section the top-k entropy loss.

Recall that the conjugate functions of multiclass SVM [15] and the top-k SVM [28] differ only in their effective domain[3] while the conjugate function is the same. Instead of the standard simplex, the conjugate of the top-k hinge loss is defined on a subset, the top-k simplex.

This suggests a way to *construct novel losses* with specific properties by taking the conjugate of an existing loss function, and modifying its essential domain in a way that enforces the desired properties. The motivation for doing so comes from the interpretation of the dual variables as forces with which every training example pushes the decision surface in the direction given by the ground truth label. The absolute value of the dual variables determines the magnitude of these forces and the optimal values are often attained at the boundary of the feasible set (which coincides with the essential domain of the loss). Therefore, by reducing the feasible set we can limit the maximal contribution of a given training example.

We begin with the **conjugate** of the **softmax loss**. Let $a^{\setminus y}$ be obtained by removing the y-th coordinate from a.

Proposition 8. *The convex conjugate of (5) is*

$$L^*(v) = \begin{cases} \sum_{j \ne y} v_j \log v_j + (1 + v_y)\log(1 + v_y), \\ \qquad \text{if } \langle \mathbf{1}, v\rangle = 0 \text{ and } v^{\setminus y} \in \Delta, \\ +\infty \quad \text{otherwise,} \end{cases} \tag{14}$$

where $\Delta \triangleq \{x \mid \langle \mathbf{1}, x\rangle \le 1,\ 0 \le x_j \le 1,\ \forall j\}$.

Proof. We provide a derivation for the convex conjugate of the softmax loss which was already given in [32, Appendix D.2.3] without a proof. We also highlight the constraint $\langle \mathbf{1}, v\rangle = 0$ which can be easily missed when computing the conjugate and is re-stated explicitly in Lemma 3.

Let $u \triangleq f(x) \in \mathbb{R}^m$. The softmax loss on example x is

$$L(u) = \log\Big(\sum_{j \in \mathcal{Y}} \exp(u_j - u_y)\Big) = \log\Big(\sum_{j \in \mathcal{Y}} \exp(u'_j)\Big),$$

[3] The **effective domain** of f is $\operatorname{dom} f = \{x \in X \mid f(x) < +\infty\}$.

where we let $u' \triangleq H_y u$ and $H_y \triangleq \mathbf{I} - \mathbf{1} e_y^\top$. Let

$$\phi(u) \triangleq \log \Big(\textstyle\sum_{j \in \mathcal{Y}} \exp(u_j) \Big),$$

then $L(u) = \phi(H_y u)$ and the convex conjugate is computed similar to [28, Lemma 2] as follows.

$$\begin{aligned} L^*(v) &= \sup\{ \langle u, v \rangle - L(u) \,|\, u \in \mathbb{R}^m \} \\ &= \sup\{ \langle u, v \rangle - \phi(H_y u) \,|\, u \in \mathbb{R}^m \} \\ &= \sup\{ \langle u^\parallel, v \rangle + \langle u^\perp, v \rangle - \phi(H_y u^\perp) \,| \\ &\qquad u^\parallel \in \operatorname{Ker} H_y, u^\perp \in \operatorname{Ker}^\perp H_y \}, \end{aligned}$$

where $\operatorname{Ker} H_y = \{ u \,|\, H_y u = 0 \} = \{ t\mathbf{1} \,|\, t \in \mathbb{R} \}$ and $\operatorname{Ker}^\perp H_y = \{ u \,|\, \langle \mathbf{1}, u \rangle = 0 \}$. It follows that $L^*(v)$ can only be finite if $\langle u^\parallel, v \rangle = 0$, which implies $v \in \operatorname{Ker}^\perp H_y \iff \langle \mathbf{1}, v \rangle = 0$. Let $H_y^\dagger$ be the Moore-Penrose pseudoinverse of H_y. For a $v \in \operatorname{Ker}^\perp H_y$, we write

$$\begin{aligned} L^*(v) &= \sup\{ \langle H_y^\dagger H_y u^\perp, v \rangle - \phi(H_y u^\perp) \,|\, u^\perp \} \\ &= \sup\{ \langle z, (H_y^\dagger)^\top v \rangle - \phi(z) \,|\, z \in \operatorname{Im} H_y \}, \end{aligned}$$

where $\operatorname{Im} H_y = \{ H_y u \,|\, u \in \mathbb{R}^m \} = \{ u \,|\, u_y = 0 \}$. Using rank-1 update of the pseudoinverse [37, § 3.2.7], we have

$$(H_y^\dagger)^\top = \mathbf{I} - e_y e_y^\top - \frac{1}{m} (\mathbf{1} - e_y) \mathbf{1}^\top,$$

which together with $\langle \mathbf{1}, v \rangle = 0$ implies

$$(H_y^\dagger)^\top v = v - v_y e_y.$$

Therefore,

$$\begin{aligned} L^*(v) &= \sup\{ \langle u, v - v_y e_y \rangle - \phi(u) \,|\, u_y = 0 \} \\ &= \sup \Big\{ \langle u^{\setminus y}, v^{\setminus y} \rangle - \log \Big(1 + \sum_{j \neq y} \exp(u_j) \Big) \Big\}. \end{aligned}$$

The function inside sup is concave and differentiable, hence the global optimum is at the critical point [10]. Setting the partial derivatives to zero yields

$$v_j = \exp(u_j) / \big(1 + \textstyle\sum_{j \neq y} \exp(u_j) \big)$$

for $j \neq y$, from which we conclude, similar to [48, § 5.1], that $\langle \mathbf{1}, v \rangle \leq 1$ and $0 \leq v_j \leq 1$ for all $j \neq y$, *i.e.* $v^{\setminus y} \in \Delta$. Let $Z \triangleq \sum_{j \neq y} \exp(u_j)$, we have at the optimum

$$u_j = \log(v_j) + \log(1 + Z), \quad \forall j \neq y.$$

Since $\langle \mathbf{1}, v \rangle = 0$, we also have that $v_y = -\sum_{j \neq y} v_j$, hence

$$\begin{aligned} L^*(v) &= \textstyle\sum_{j \neq y} u_j v_j - \log(1 + Z) \\ &= \textstyle\sum_{j \neq y} v_j \log(v_j) + \log(1 + Z) \Big(\sum_{j \neq y} v_j - 1 \Big) \\ &= \textstyle\sum_{j \neq y} v_j \log(v_j) - \log(1 + Z)(1 + v_y). \end{aligned}$$

Summing v_j and using the definition of Z,

$$\sum_{j \neq y} v_j = \sum_{j \neq y} \exp(u_j) / \big(1 + \sum_{j \neq y} \exp(u_j) \big) = Z / (1 + Z).$$

Therefore,

$$1 + Z = 1 / \big(1 - \textstyle\sum_{j \neq y} v_j \big) = 1 / (1 + v_y),$$

which finally yields

$$L^*(v) = \textstyle\sum_{j \neq y} v_j \log(v_j) + \log(1 + v_y)(1 + v_y),$$

if $\langle \mathbf{1}, v \rangle = 0$ and $v^{\setminus y} \in \Delta$ as stated in the proposition. □

The **conjugate** of the **top-k entropy loss** is obtained by replacing Δ in (14) with Δ_k^α. A β version could be obtained using the Δ_k^β instead, which defer to future work. There is no closed-form solution for the primal top-k entropy loss for $k > 1$, but we can evaluate it as follows.

Proposition 9. *Let $u_j \triangleq f_j(x) - f_y(x)$ for all $j \in \mathcal{Y}$. The **top-k entropy loss** is defined as*

$$\begin{aligned} L(u) = \max \big\{ & \langle u^{\setminus y}, x \rangle - (1 - s) \log(1 - s) \\ & - \langle x, \log x \rangle \,|\, x \in \Delta_k^\alpha, \ \langle \mathbf{1}, x \rangle = s \big\}. \end{aligned} \tag{15}$$

Moreover, we recover the softmax loss (5) if $k = 1$.

Proof. The convex conjugate of the top-k entropy loss is

$$L^*(v) \triangleq \begin{cases} \sum_{j \neq y} v_j \log v_j + (1 + v_y) \log(1 + v_y), \\ \qquad \text{if } \langle \mathbf{1}, v \rangle = 0 \text{ and } v^{\setminus y} \in \Delta_k^\alpha, \\ +\infty \quad \text{otherwise}, \end{cases}$$

where the setting is the same as in Proposition 8. The (primal) top-k entropy loss is defined as the convex conjugate of the $L^*(v)$ above. We have

$$\begin{aligned} L(u) &= \sup\{ \langle u, v \rangle - L^*(v) \,|\, v \in \mathbb{R}^m \} \\ &= \sup\{ \langle u, v \rangle - \sum_{j \neq y} v_j \log v_j - (1 + v_y) \log(1 + v_y) \\ &\qquad | \ \langle \mathbf{1}, v \rangle = 0, \ v^{\setminus y} \in \Delta_k^\alpha \} \\ &= \sup\{ \langle u^{\setminus y}, v^{\setminus y} \rangle - u_y \sum_{j \neq y} v_j - \sum_{j \neq y} v_j \log v_j \\ &\qquad - (1 - \sum_{j \neq y} v_j) \log(1 - \sum_{j \neq y} v_j) \,|\, v^{\setminus y} \in \Delta_k^\alpha \}. \end{aligned}$$

Note that $u_y = 0$, and hence the corresponding term vanishes. Finally, we let $x \triangleq v^{\setminus y}$ and $s \triangleq \sum_{j \neq y} v_j = \langle \mathbf{1}, x \rangle$ and obtain (15).

Next, we discuss how this problem can be solved and show that it reduces to the softmax loss for $k = 1$. Let $a \triangleq u^{\setminus y}$ and consider an equivalent problem below.

$$L(u) = -\min \big\{ \langle x, \log x \rangle + (1-s)\log(1-s) \\ - \langle a, x \rangle \,|\, x \in \Delta_k^\alpha,\ \langle \mathbf{1}, x \rangle = s \big\}. \tag{16}$$

The Lagrangian for (16) is

$$\mathcal{L} = \langle x, \log x \rangle + (1-s)\log(1-s) - \langle a, x \rangle \\ + t(\langle \mathbf{1}, x \rangle - s) + \lambda(s-1) - \langle \mu, x \rangle + \langle \nu, x - \tfrac{s}{k}\mathbf{1} \rangle,$$

where $t \in \mathbb{R}$ and $\lambda, \mu, \nu \geq 0$ are the dual variables. Computing the partial derivatives of $\mathcal{L}$ w.r.t. x_j and s, and setting them to zero, we obtain

$$\log x_j = a_j - 1 - t + \mu_j - \nu_j, \quad \forall j \\ \log(1-s) = -1 - t - \tfrac{1}{k}\langle \mathbf{1}, \nu \rangle + \lambda.$$

Note that $x_j = 0$ and $s = 1$ cannot satisfy the above conditions for any choice of the dual variables in $\mathbb{R}$. Therefore, $x_j > 0$ and $s < 1$, which implies $\mu_j = 0$ and $\lambda = 0$. The only constraint that might be active is $x_j \leq \frac{s}{k}$. Note, however, that in view of $x_j > 0$ it can only be active if either $k > 1$ or we have a one dimensional problem. We consider the case when this constraint is active below.

Consider x_j's for which $0 < x_j < \frac{s}{k}$ holds at the optimum. The complementary slackness conditions imply that the corresponding $\mu_j = \nu_j = 0$. Let $p \triangleq \langle \mathbf{1}, \nu \rangle$ and redefine t as $t \leftarrow 1 + t$. We obtain the simplified equations

$$\log x_j = a_j - t, \\ \log(1-s) = -t - \tfrac{p}{k}.$$

If $k = 1$, then $0 < x_j < s$ for all j in a multiclass problem as discussed above, hence also $p = 0$. We have

$$x_j = e^{a_j - t}, \qquad 1 - s = e^{-t},$$

where $t \in \mathbb{R}$ is to be found. Plugging that into the objective,

$$\textstyle\sum_j (a_j - t)e^{a_j - t} - te^{-t} - \sum_j a_j e^{a_j - t} \\ = e^{-t}\Big[\sum_j (a_j - t)e^{a_j} - t - \sum_j a_j e^{a_j}\Big] \\ = -te^{-t}\big[1 + \sum_j e^{a_j}\big] = -t\big[e^{-t} + \sum_j e^{a_j - t}\big] \\ = -t\big[1 - s + s\big] = -t.$$

To compute t, we note that

$$\textstyle\sum_j e^{a_j - t} = \langle \mathbf{1}, x \rangle = s = 1 - e^{-t},$$

from which we conclude

$$1 = \big(1 + \sum_j e^{a_j}\big)e^{-t} \implies -t = -\log(1 + \textstyle\sum_j e^{a_j}).$$

Taking into account the minus in front of the min in (16) and the definition of a, we finally recover the softmax loss

$$L(y, f(x)) = \log\big(1 + \textstyle\sum_{j \neq y} \exp(f_j(x) - f_y(x))\big).$$

□

The non-analytic nature of the loss for $k > 1$ does not allow us to check if it is top-k calibrated. We now show how this problem can be solved efficiently.

How to solve (15). We continue the derivation started in the proof of Propostion 9. First, we write the system that follows directly from the KKT [10] optimality conditions.

$$x_j = \min\{\exp(a_j - t), \tfrac{s}{k}\}, \quad \forall j, \\ \nu_j = \max\{0, a_j - t - \log(\tfrac{s}{k})\}, \quad \forall j, \\ 1 - s = \exp(-t - \tfrac{p}{k}), \\ s = \langle \mathbf{1}, x \rangle, \quad p = \langle \mathbf{1}, \nu \rangle. \tag{17}$$

Next, we define the two index sets U and M as follows

$$U \triangleq \{j \,|\, x_j = \tfrac{s}{k}\}, \qquad M \triangleq \{j \,|\, x_j < \tfrac{s}{k}\}.$$

Note that the set U contains at most k indexes corresponding to the largest components of a_j. Now, we proceed with finding a t that solves (17). Let $\rho \triangleq \frac{|U|}{k}$. We eliminate p as

$$p = \sum_j \nu_j = \sum_U a_j - |U|\big(t + \log(\tfrac{s}{k})\big) \implies \\ \tfrac{p}{k} = \tfrac{1}{k}\sum_U a_j - \rho\big(t + \log(\tfrac{s}{k})\big).$$

Let $Z \triangleq \sum_M \exp a_j$, we write for s

$$s = \sum_j x_j = \sum_U \tfrac{s}{k} + \sum_M \exp(a_j - t) \\ = \rho s + \exp(-t)\sum_M \exp a_j = \rho s + \exp(-t)Z.$$

We conclude that

$$(1-\rho)s = \exp(-t)Z \implies \\ t = \log Z - \log\big((1-\rho)s\big).$$

Let $\alpha \triangleq \frac{1}{k}\sum_U a_j$. We further write

$$\log(1-s) = -t - \tfrac{p}{k} \\ = -t - \alpha + \rho\big(t + \log(\tfrac{s}{k})\big) \\ = \rho\log(\tfrac{s}{k}) - (1-\rho)t - \alpha \\ = \rho\log(\tfrac{s}{k}) - \alpha \\ - (1-\rho)\big[\log Z - \log\big((1-\rho)s\big)\big],$$

which yields the following equation for s

$$\log(1-s) - \rho(\log s - \log k) + \alpha$$

$$+ (1-\rho)\big[\log Z - \log(1-\rho) - \log s\big] = 0.$$

Therefore,

$$\begin{aligned}
&\log(1-s) - \log s + \rho \log k + \alpha \\
&+ (1-\rho)\log Z - (1-\rho)\log(1-\rho) = 0, \\
&\log\left(\frac{1-s}{s}\right) = \log\left(\frac{(1-\rho)^{(1-\rho)}\exp(-\alpha)}{k^\rho Z^{(1-\rho)}}\right).
\end{aligned}$$

We finally get

$$\begin{aligned}
&s = 1/(1+Q), \\
&Q \triangleq (1-\rho)^{(1-\rho)}/(k^\rho Z^{(1-\rho)} e^\alpha).
\end{aligned} \tag{18}$$

We note that: *a)* Q is readily computable once the sets U and M are fixed; and *b)* $Q = 1/Z$ if $k = 1$ since $\rho = \alpha = 0$ in that case. This yields the formula for t as

$$t = \log Z + \log(1+Q) - \log(1-\rho). \tag{19}$$

As a sanity check, we note that we again recover the softmax loss for $k = 1$, since $t = \log Z + \log(1 + 1/Z) = \log(1+Z) = \log(1 + \sum_j \exp a_j)$.

To verify that the computed s and t are compatible with the choice of the sets U and M, we check if this holds:

$$\begin{aligned}
&\exp(a_j - t) \geq \tfrac{s}{k}, \quad \forall j \in U, \\
&\exp(a_j - t) \leq \tfrac{s}{k}, \quad \forall j \in M,
\end{aligned}$$

which is equivalent to

$$\max_M a_j \leq \log(\tfrac{s}{k}) + t \leq \min_U a_j. \tag{20}$$

Computation of the top-k entropy loss (15). The above derivation suggests a simple and efficient algorithm to compute s and t that solve the KKT system (17) and, therefore, the original problem (15).

1. Initialization: $U \leftarrow \{\}$, $M \leftarrow \{1, \ldots, m\}$.
2. $Z \leftarrow \sum_M \exp a_j$, $\alpha \leftarrow \frac{1}{k}\sum_U a_j$, $\rho \leftarrow \frac{|U|}{k}$.
3. Compute s and t using (18) and (19).
4. If (20) holds, stop; otherwise, $j \leftarrow \arg\max_M a_j$.
5. $U \leftarrow U \cup \{j\}$, $M \leftarrow M \setminus \{j\}$ and go to step 2.

Note that the algorithm terminates after at most k iterations since $|U| \leq k$. The overall complexity is therefore $O(km)$.

To compute the actual loss (15), we note that if U is empty, *i.e.* there were no violated constraints, then the top-k entropy loss coincides with the softmax loss and is directly given by t. Otherwise, we have

$$\begin{aligned}
&\langle a, x\rangle - \langle x, \log x\rangle - (1-s)\log(1-s) \\
&= \sum_U a_j \tfrac{s}{k} + \sum_M a_j \exp(a_j - t) - \sum_U \tfrac{s}{k}\log(\tfrac{s}{k}) \\
&- \sum_M (a_j - t)\exp(a_j - t) - (1-s)\log(1-s) \\
&= \alpha s - \rho s \log(\tfrac{s}{k}) + t\exp(-t) Z - (1-s)\log(1-s) \\
&= \alpha s - \rho s \log(\tfrac{s}{k}) + (1-\rho) s t - (1-s)\log(1-s).
\end{aligned}$$

Therefore, the top-k entropy loss is readily computed once the optimal s and t are found.

2.5. Truncated Top-k Entropy Loss

A major limitation of the softmax loss for top-k error optimization is that it cannot ignore the highest scoring predictions, which yields a high loss even if the top-k error is zero. This can be seen by rewriting (5) as

$$L(y, f(x)) = \log\big(1 + \sum_{j \neq y} \exp(f_j(x) - f_y(x))\big). \tag{21}$$

If there is only a *single* j such that $f_j(x) - f_y(x) \gg 0$, then $L(y, f(x)) \gg 0$ even though $\mathrm{err}_2(y, f(x)) = 0$.

This problem is is also present in all top-k hinge losses considered above and is an inherent limitation due to their convexity. The origin of the problem is the fact that ranking based losses [54] are based on functions such as

$$\phi(f(x)) = \tfrac{1}{m}\sum_{j \in \mathcal{Y}} \alpha_j f_{\pi_j}(x) - f_y(x).$$

The function ϕ is convex if the sequence (α_j) is monotonically non-increasing [10]. This implies that convex ranking based losses have to put *more* weight on the highest scoring classifiers, while we would like to put *less* weight on them. To that end, we drop the first $(k-1)$ highest scoring predictions from the sum in (21), sacrificing convexity of the loss, and define the **truncated top-k entropy loss** as follows

$$L(y, f(x)) = \log\big(1 + \sum_{j \in \mathcal{J}_y} \exp(f_j(x) - f_y(x))\big), \tag{22}$$

where $\mathcal{J}_y$ are the indexes corresponding to the $(m-k)$ *smallest* components of $(f_j(x))_{j \neq y}$. This loss can be seen as a smooth version of the top-k error (1), as it is small whenever the top-k error is zero. Below, we show that this loss is top-k calibrated.

Proposition 10. *The truncated top-k entropy loss is top-s calibrated for any $k \leq s \leq m$.*

Proof. Given a $g \in \mathbb{R}^m$, let π be a permutation such that $g_{\pi_1} \geq g_{\pi_2} \geq \ldots \geq g_{\pi_m}$. Then, we have

$$\mathcal{J}_y = \begin{cases} \{\pi_{k+1}, \ldots, \pi_m\} & \text{if } y \in \{\pi_1, \ldots, \pi_{k-1}\}, \\ \{\pi_k, \ldots, \pi_m\} \setminus \{y\} & \text{if } y \in \{\pi_k, \ldots, \pi_m\}. \end{cases}$$

Therefore, the expected loss at x can be written as

$$\begin{aligned}\mathbb{E}_{Y|X}[L(Y,g)\,|\,X=x] &= \textstyle\sum_{y\in\mathcal{Y}} L(y,g)\,p_y(x) \\ &= \textstyle\sum_{r=1}^{k-1} \log\left(1+\sum_{j=k+1}^{m} e^{g_{\pi_j}-g_{\pi_r}}\right) p_{\pi_r}(x) \\ &\quad + \textstyle\sum_{r=k}^{m} \log\left(\sum_{j=k}^{m} e^{g_{\pi_j}-g_{\pi_r}}\right) p_{\pi_r}(x).\end{aligned}$$

Note that the sum inside the logarithm does not depend on g_{π_r} for $r < k$. Therefore, a Bayes optimal classifier will have $g_{\pi_r} = +\infty$ for all $r < k$ as then the first sum vanishes.

Let $p \triangleq (p_y(x))_{y\in\mathcal{Y}}$ and $q \triangleq (L(y,g))_{y\in\mathcal{Y}}$, then

$$q_{\pi_1} = \ldots = q_{\pi_{k-1}} = 0 \leq q_{\pi_k} \leq \ldots \leq q_{\pi_m}$$

and we can re-write the expected loss as

$$\mathbb{E}_{Y|X}[L(Y,g)\,|\,X=x] = \langle p, q\rangle = \langle p_\pi, q_\pi\rangle \geq \langle p_\tau, q_\pi\rangle\,,$$

where $p_{\tau_1} \geq p_{\tau_2} \geq \ldots \geq p_{\tau_m}$ and we used the rearrangement inequality. Therefore, the expected loss is minimized when π and τ coincide (up to a permutation of the first $k-1$ elements, since they correspond to zero loss).

We can also derive a Bayes optimal classifier following the proof of Proposition 4. We have

$$\begin{aligned}&\mathbb{E}_{Y|X}[L(Y,g)\,|\,X=x] \\ &= \textstyle\sum_{r=k}^{m} \log\left(\sum_{j=k}^{m} e^{g_{\tau_j}-g_{\tau_r}}\right) p_{\tau_r}(x) \\ &= \textstyle\sum_{r=k}^{m} \left(\log\left(\sum_{j=k}^{m} e^{g_{\tau_j}}\right) - g_{\tau_r}\right) p_{\tau_r}(x).\end{aligned}$$

A critical point is found by setting partial derivatives to zero for all $y \in \{\tau_k, \ldots, \tau_m\}$, which leads to

$$\frac{e^{g_y}}{\sum_{j=k}^{m} e^{g_{\tau_j}}} \textstyle\sum_{r=k}^{m} p_{\tau_r}(x) = p_y(x).$$

We let $g_y = -\infty$ if $p_y(x) = 0$, and obtain finally

$$g^*_{\tau_j} = \begin{cases} +\infty & \text{if } j < k, \\ \log\left(\alpha p_{\tau_j}(x)\right) & \text{if } j \geq k \text{ and } p_{\tau_j}(x) > 0, \\ -\infty & \text{if } j \geq k \text{ and } p_{\tau_j}(x) = 0, \end{cases}$$

as a Bayes optimal classifier for any $\alpha > 0$.

Note that g^* preserves the ranking of $p_y(x)$ for all y in $\{\tau_k, \ldots, \tau_m\}$, hence, it is top-$s$ calibrated for all $s \geq k$. $\square$

As the loss (22) is nonconvex, we use solutions obtained with the softmax loss (5) as initial points and optimize them further via gradient descent. However, the resulting optimization problem seems to be "mildly nonconvex" as the same-quality solutions are obtained from different initializations. In Section 4, we show a synthetic experiment, where the advantage of discarding the highest scoring classifier in the loss becomes apparent.

3. Optimization Method

In this section, we briefly discuss how the proposed smooth top-k hinge losses and the top-k entropy loss can be optimized efficiently within the SDCA framework of [48].

The primal and dual problems. Let $X \in \mathbb{R}^{d\times n}$ be the matrix of training examples $x_i \in \mathbb{R}^d$, $K = X^\top X$ the corresponding Gram matrix, $W \in \mathbb{R}^{d\times m}$ the matrix of primal variables, $A \in \mathbb{R}^{m\times n}$ the matrix of dual variables, and $\lambda > 0$ the regularization parameter. The primal and Fenchel dual [8] objective functions are given as

$$\begin{aligned}P(W) &= +\frac{1}{n}\sum_{i=1}^{n} L\left(y_i, W^\top x_i\right) + \frac{\lambda}{2}\operatorname{tr}\left(W^\top W\right), \\ D(A) &= -\frac{1}{n}\sum_{i=1}^{n} L^*\left(y_i, -\lambda n a_i\right) - \frac{\lambda}{2}\operatorname{tr}\left(AKA^\top\right),\end{aligned} \tag{23}$$

where L^* is the convex conjugate of L. SDCA proceeds by randomly picking a variable a_i (which in our case is a vector of dual variables over all m classes for a sample x_i) and modifying it to achieve maximal increase in the dual objective $D(A)$. It turns out that this update step is equivalent to a proximal problem, which can be seen as a regularized projection onto the essential domain of L^*.

The convex conjugate. An important ingredient in the SDCA framework is the convex conjugate L^*. We show that for all multiclass loss functions that we consider the fact that they depend on the differences $f_j(x) - f_y(x)$ enforces a certain constraint on the conjugate function.

Lemma 3. *Let $H_y = \mathbf{I} - \mathbf{1}e_y^\top$ and let $\Phi(u) = \phi(H_y u)$. $\Phi^*(v) = +\infty$ unless $\langle \mathbf{1}, v\rangle = 0$.*

Proof. The proof follows directly from [28, Lemma 2] and was already reproduced in the proof of Proposition 8 for the softmax loss. We have formulated this simplified lemma since [28, Lemma 2] additionally required y-compatibility to show that if $\langle \mathbf{1}, v\rangle = 0$, then $\Phi^*(v) = \phi^*(v - v_y e_y)$, which does not hold *e.g.* for the softmax loss. $\square$

Lemma 3 tells us that we need to enforce $\langle \mathbf{1}, a_i\rangle = 0$ at all times, which translates into $a_{y_i} = -\sum_{j\neq y_i} a_j$. The update steps are performed on the $(m-1)$-dimensional vector obtained by removing the coordinate a_{y_i}.

The update step for top-k SVM$^\alpha_\gamma$. Let $a^{\setminus y}$ be obtained by removing the y-th coordinate from vector a. We show that performing an update step for the smooth top-k hinge loss is equivalent to projecting a certain vector b, computed from the prediction scores $W^\top x_i$, onto the essential domain of L^*, the top-k simplex, with an added regularization $\rho\,\langle \mathbf{1}, x\rangle^2$, which biases the solution to be orthogonal to $\mathbf{1}$.

Proposition 11. *Let L and L^* in (23) be respectively the top-k SVM$^\alpha_\gamma$ loss and its conjugate as in Proposition 7.*

The update $\max_{a_i}\{D(A) \mid \langle \mathbf{1}, a_i\rangle = 0\}$ *is equivalent with the change of variables* $x \leftrightarrow -a_i^{\setminus y_i}$ *to solving*

$$\min_x \{\|x - b\|^2 + \rho \langle \mathbf{1}, x\rangle^2 \mid x \in \Delta_k^{\alpha}(\tfrac{1}{\lambda n})\}, \tag{24}$$

where $b = \frac{1}{\langle x_i, x_i\rangle + \gamma n\lambda}\left(q^{\setminus y_i} + (1 - q_{y_i})\mathbf{1}\right)$*,*
$q = W^\top x_i - \langle x_i, x_i\rangle a_i$*, and* $\rho = \frac{\langle x_i, x_i\rangle}{\langle x_i, x_i\rangle + \gamma n\lambda}$*.*

Proof. We follow the proof of [28, Proposition 4]. We choose $i \in \{1, \ldots, n\}$ and, having all other variables fixed, update a_i to maximize

$$-\tfrac{1}{n} L^*(y_i, -\lambda n a_i) - \tfrac{\lambda}{2} \operatorname{tr}\left(AKA^\top\right).$$

For the nonsmooth top-k hinge loss, it was shown [28] that

$$L^*(y_i, -\lambda n a_i) = \langle c, \lambda n(a_i - a_{y_i,i} e_{y_i})\rangle$$

if $-\lambda n(a_i - a_{y_i,i} e_{y_i}) \in \Delta_k^{\alpha}$ and $+\infty$ otherwise. Now, for the smoothed loss, we add regularization and obtain

$$-\tfrac{1}{n}\left(\tfrac{\gamma}{2}\|-\lambda n(a_i - a_{y_i,i} e_{y_i})\|^2 + \langle c, \lambda n(a_i - a_{y_i,i} e_{y_i})\rangle\right)$$

with $-\lambda n(a_i - a_{y_i,i} e_{y_i}) \in \Delta_k^{\alpha}$. Using $c = \mathbf{1} - e_{y_i}$ and $\langle \mathbf{1}, a_i\rangle = 0$, one can simplify it to

$$-\frac{\gamma n\lambda^2}{2}\left\|a_i^{\setminus y_i}\right\|^2 + \lambda a_{y_i,i},$$

and the feasibility constraint can be re-written as

$$-a_i^{\setminus y_i} \in \Delta_k^{\alpha}(\tfrac{1}{\lambda n}), \qquad a_{y_i,i} = \langle \mathbf{1}, -a_i^{\setminus y_i}\rangle.$$

For the regularization term $\operatorname{tr}\left(AKA^\top\right)$, we have

$$\operatorname{tr}\left(AKA^\top\right) = K_{ii}\langle a_i, a_i\rangle + 2\sum_{j \neq i} K_{ij}\langle a_i, a_j\rangle + \text{const}.$$

We let $q = \sum_{j \neq i} K_{ij} a_j = AK_i - K_{ii} a_i$ and $x = -a_i^{\setminus y_i}$:

$$\langle a_i, a_i\rangle = \langle \mathbf{1}, x\rangle^2 + \langle x, x\rangle,$$
$$\langle q, a_i\rangle = q_{y_i}\langle \mathbf{1}, x\rangle - \langle q^{\setminus y_i}, x\rangle.$$

Now, we plug everything together and multiply with $-2/\lambda$.

$$\min_{x \in \Delta_k^{\alpha}(\frac{1}{\lambda n})} \gamma n\lambda \|x\|^2 - 2\langle \mathbf{1}, x\rangle + 2\big(q_{y_i}\langle \mathbf{1}, x\rangle - \langle q^{\setminus y_i}, x\rangle\big) + K_{ii}\big(\langle \mathbf{1}, x\rangle^2 + \langle x, x\rangle\big).$$

Collecting the corresponding terms finishes the proof. ☐

Note that setting $\gamma = 0$, we recover the update step for the non-smooth top-k hinge loss [28]. It turns out that we can employ their projection procedure for solving (24) with only a minor modification of b and ρ.

The update step for the top-k SVM_γ^β loss is derived similarly using the set Δ_k^β in (24) instead of Δ_k^α. The resulting projection problem is a biased continuous quadratic knapsack problem, which is discussed in the supplement of [28].

Smooth top-k hinge losses converge significantly faster than their nonsmooth variants as we show in the scaling experiments below. This can be explained by the theoretical results of [48] on the convergence rate of SDCA. They also had similar observations for the smoothed binary hinge loss.

The update step for top-k Ent. We now discuss the optimization of the proposed top-k entropy loss in the SDCA framework. Note that the top-k entropy loss reduces to the softmax loss for $k = 1$. Thus, our SDCA approach can be used for *gradient-free* optimization of the softmax loss without having to tune step sizes or learning rates.

Proposition 12. *Let* L *in (23) be the* top-k Ent *loss (15) and* L^* *be its convex conjugate as in (14) with* Δ *replaced by* Δ_k^α*. The update* $\max_{a_i}\{D(A) \mid \langle \mathbf{1}, a_i\rangle = 0\}$ *is equivalent with the change of variables* $x \leftrightarrow -\lambda n a_i^{\setminus y_i}$ *to solving*

$$\begin{aligned} \min_{x \in \Delta_k^\alpha} \ & \tfrac{\alpha}{2}(\langle x, x\rangle + \langle \mathbf{1}, x\rangle^2) - \langle b, x\rangle + \\ & \langle x, \log x\rangle + (1 - \langle \mathbf{1}, x\rangle)\log(1 - \langle \mathbf{1}, x\rangle) \end{aligned} \tag{25}$$

where $\alpha = \frac{\langle x_i, x_i\rangle}{\lambda n}$*,* $b = q^{\setminus y_i} - q_{y_i}\mathbf{1}$*,* $q = W^\top x_i - \langle x_i, x_i\rangle a_i$*.*

Proof. Let $v \triangleq -\lambda n a_i$ and $y = y_i$. Using Proposition 8,

$$L^*(y, v) = \textstyle\sum_{j \neq y} v_j \log v_j + (1 + v_y)\log(1 + v_y),$$

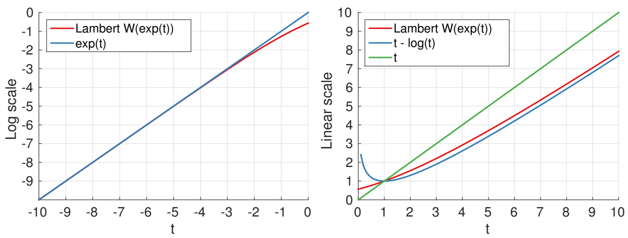
where $\langle \mathbf{1}, v\rangle = 0$ and $v^{\setminus y} \in \Delta_k^\alpha$. Let $x \triangleq v^{\setminus y}$ and $s \triangleq -v_y$. It follows that $s = \langle \mathbf{1}, x\rangle$ and from $\operatorname{tr}\left(AKA^\top\right)$ we get

$$K_{ii}(\langle x, x\rangle + s^2)/(\lambda n)^2 - 2\big\langle q^{\setminus y} - q_y \mathbf{1}, x\big\rangle/(\lambda n),$$

where $q = \sum_{j \neq i} K_{ij} a_j = AK_i - K_{ii} a_i$ as before. Finally, we plug everything together as in Proposition 11. ☐

Note that this optimization problem is similar to (24), but is more difficult to solve due to the presence of logarithms in the objective. We propose to tackle this problem using the Lambert W function introduced below.

Lambert W function. The Lambert W function is defined to be the inverse of the function $w \mapsto we^w$ and is widely used in many fields [14, 18, 55]. Taking logarithms on both sides of the defining equation $z = We^W$, we obtain $\log z = W(z) + \log W(z)$. Therefore, if we are given an equation of the form $x + \log x = t$ for some $t \in \mathbb{R}$, we can directly "solve" it in closed-form as $x = W(e^t)$. The crux of the problem is that the function $V(t) \triangleq W(e^t)$ is transcendental [18] just like the logarithm and the exponent. There exist highly optimized implementations for the latter and we argue that the same can be done for the Lambert W



(a) $V(t) \approx e^t$ for $t \ll 0$. (b) $V(t) \approx t - \log t$ for $t \gg 0$.

Figure 1: Behavior of the Lambert W function of the exponent ($V(t) = W(e^t)$). **(a)** Log scale plot with $t \in (-10, 0)$. **(b)** Linear scale plot with $t \in (0, 10)$.

function. In fact, there is already some work on this topic [18, 55], which we also employ in our implementation.

To develop intuition concerning the Lambert W function of the exponent, we now briefly discuss how the function $V(t) = W(e^t)$ behaves for different values of t. An illustration is provided in Figure 1. One can see directly from the equation $x + \log x = t$ that the behavior of $x = V(t)$ changes dramatically depending on whether t is a large positive or a large negative number. In the first case, the linear part dominates the logarithm and the function is approximately linear; a better approximation is $x(t) \approx t - \log t$, when $t \gg 1$. In the second case, the function behaves like an exponent e^t. To see this, we write $x = e^t e^{-x}$ and note that $e^{-x} \approx 1$ when $t \ll 0$, therefore, $x(t) \approx e^t$, if $t \ll 0$. We use these approximations as initial points for a 5-th order Householder method [22], which was also used in [18]. A *single* iteration is already sufficient to get full `float` precision and at most two iterations are needed for `double`.

How to solve (25). We present a similar derivation as was already done for the problem (15) above. The main difference is that we now encounter the Lambert W function in the optimality conditions. We re-write the problem as

$$\min \left\{ \tfrac{\alpha}{2}(\langle x, x\rangle + s^2) - \langle a, x\rangle + \langle x, \log x\rangle \right. \\ \left. + (1-s)\log(1-s) \mid s = \langle \mathbf{1}, x\rangle \,,\; x \in \Delta_k^\alpha \right\}.$$

The Lagrangian is given by

$$\begin{aligned}\mathcal{L} =& \tfrac{\alpha}{2}(\langle x, x\rangle + s^2) - \langle a, x\rangle + \langle x, \log x\rangle \\ &+ (1-s)\log(1-s) + t(\langle \mathbf{1}, x\rangle - s) \\ &+ \lambda(s-1) - \langle \mu, x\rangle + \left\langle \nu, x - \tfrac{s}{k}\mathbf{1}\right\rangle,\end{aligned}$$

where $t \in \mathbb{R}$, $\lambda, \mu, \nu \geq 0$ are the dual variables. Computing partial derivatives of $\mathcal{L}$ w.r.t. x_i and s, and setting them to zero, we obtain

$$\begin{aligned}\alpha x_i + \log x_i &= a_i - 1 - t + \mu_i - \nu_i, \quad \forall i, \\ \alpha(1-s) + \log(1-s) &= \alpha - 1 - t - \lambda - \tfrac{1}{k}\langle \mathbf{1}, \nu\rangle\,, \quad \forall i.\end{aligned}$$

Note that $x_i > 0$ and $s < 1$ as before, which implies $\mu_i = 0$ and $\lambda = 0$. We re-write the above as

$$\begin{aligned}&\alpha x_i + \log(\alpha x_i) = a_i - 1 - t + \log\alpha - \nu_i, \\ &\alpha(1-s) + \log\big(\alpha(1-s)\big) = \alpha - 1 - t + \log\alpha - \tfrac{\langle \mathbf{1}, \nu\rangle}{k}.\end{aligned}$$

Note that these equations correspond to the Lambert W function of the exponent, *i.e.* $V(t) = W(e^t)$ discussed above. Let $p \triangleq \langle \mathbf{1}, \nu\rangle$ and re-define $t \leftarrow 1 + t - \log\alpha$.

$$\begin{aligned}\alpha x_i &= W\big(\exp(a_i - t - \nu_i)\big), \\ \alpha(1-s) &= W\big(\exp(\alpha - t - \tfrac{p}{k})\big).\end{aligned}$$

Finally, we obtain the following system:

$$\begin{aligned}x_i &= \min\{\tfrac{1}{\alpha}V(a_i - t), \tfrac{s}{k}\}, \quad \forall i, \\ \alpha x_i &= V(a_i - t - \nu_i), \quad \forall i, \\ \alpha(1-s) &= V(\alpha - t - \tfrac{p}{k}), \\ s &= \langle \mathbf{1}, x\rangle\,, \quad p = \langle \mathbf{1}, \nu\rangle\,.\end{aligned}$$

Note that $V(t)$ is a strictly monotonically increasing function, therefore, it is invertible and we can write

$$\begin{aligned}a_i - t - \nu_i &= V^{-1}(\alpha x_i), \\ \alpha - t - \tfrac{p}{k} &= V^{-1}\big(\alpha(1-s)\big).\end{aligned}$$

Next, we defined the sets U and M as before and write

$$\begin{aligned}s &= \langle \mathbf{1}, x\rangle = \textstyle\sum_U \tfrac{s}{k} + \sum_M \tfrac{1}{\alpha}V(a_i - t), \\ p &= \langle \mathbf{1}, \nu\rangle = \textstyle\sum_U a_i - |U|\left(t + V^{-1}(\tfrac{\alpha s}{k})\right).\end{aligned}$$

Let $\rho \triangleq \frac{|U|}{k}$ as before and $A \triangleq \frac{1}{k}\sum_U a_i$, we get

$$\begin{aligned}(1-\rho)s &= \tfrac{1}{\alpha}\textstyle\sum_M V(a_i - t), \\ \tfrac{p}{k} &= A - \rho\big(t + V^{-1}(\tfrac{\alpha s}{k})\big).\end{aligned}$$

Finally, we eliminate p and obtain a system in *two* variables,

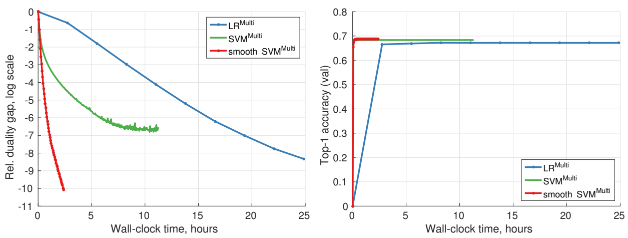
$$\begin{aligned}&\alpha(1-\rho)s - \textstyle\sum_M V(a_i - t) = 0, \\ &(1-\rho)t + V^{-1}\big(\alpha(1-s)\big) - \rho V^{-1}(\tfrac{\alpha s}{k}) + A - \alpha = 0,\end{aligned}$$

which can be solved using the Newton's method [36]. Moreover, when U is empty, the system above simplifies into a single equation in *one* variable t

$$V(\alpha - t) + \textstyle\sum_M V(a_i - t) = \alpha,$$

which can be solved efficiently using the Householder's method [22]. As both methods require derivatives of $V(t)$, we note that $\partial_t V(t) = V(t)/(1 + V(t))$ [14]. Therefore, $V(a_i - t)$ is only computed *once* for each $a_i - t$ and then re-used to also compute the derivatives.

The efficiency of this approach crucially depends on fast computation of $V(t)$. Our implementation was able to scale the training procedure to large datasets as we show next.



(a) Relative duality gap vs. time (b) Top-1 accuracy vs. time

Figure 2: SDCA convergence with $\mathrm{LR}^{\mathrm{Multi}}$, $\mathrm{SVM}^{\mathrm{Multi}}$, and top-1 SVM_1^α objectives on ILSVRC 2012.

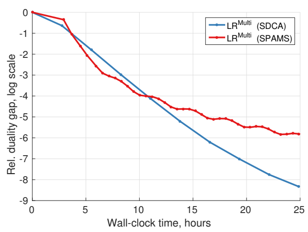


Figure 3: Convergence rate of SDCA (ours) and the SPAMS toolbox [33].

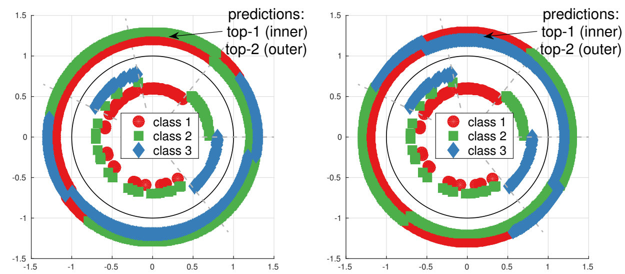
Runtime. We compare the wall-clock runtime of the top-1 multiclass SVM [28] ($\mathrm{SVM}^{\mathrm{Multi}}$) with our smooth multiclass SVM (smooth $\mathrm{SVM}^{\mathrm{Multi}}$) and the softmax loss ($\mathrm{LR}^{\mathrm{Multi}}$) objectives in Figure 2. We plot the relative duality gap $(P(W) - D(A))/P(W)$ and the validation accuracy versus time for the best performing models on ILSVRC 2012. We obtain substantial improvement of the convergence rate for smooth top-1 SVM compared to the non-smooth baseline. Moreover, top-1 accuracy saturates after a few passes over the training data, which justifies the use of a fairly loose stopping criterion (we used 10^{-3}). For $\mathrm{LR}^{\mathrm{Multi}}$, the cost of each epoch is significantly higher compared to the top-1 SVMs, which is due to the difficulty of solving (25). This suggests that one can use the smooth top-1 SVM_1^α and obtain competitive performance (see § 5) at a lower training cost.

We also compare our implementation $\mathrm{LR}^{\mathrm{Multi}}$ (SDCA) with the SPAMS optimization toolbox [33], denoted $\mathrm{LR}^{\mathrm{Multi}}$ (SPAMS), which provides an efficient implementation of FISTA [4]. We note that the rate of convergence of SDCA is competitive with FISTA for $\epsilon \geq 10^{-4}$ and is noticeably better for $\epsilon < 10^{-4}$. We conclude that our approach is competitive with the state-of-the-art, and faster computation of $V(t)$ would lead to a further speedup.

Gradient-based optimization. Finally, we note that the proposed smooth top-k hinge and the truncated top-k entropy losses are easily amenable to gradient-based optimization, in particular, for training deep architectures (see § 5). The computation of the gradient of (22) is straightforward, while for the smooth top-k hinge loss (12) we have

$$\nabla L_\gamma(a) = \tfrac{1}{\gamma} \operatorname{proj}_{\Delta_k^\alpha(\gamma)}(a + c),$$

which follows from the fact that $L_\gamma(a)$ can be written as $\frac{1}{2\gamma}(\|x\|^2 - \|x - p\|^2)$ for $x = a + c$ and $p = \operatorname{proj}_{\Delta_k^\alpha(\gamma)}(x)$, and a known result from convex analysis [8, § 3, Ex. 12.d] which postulates that $\nabla_x \frac{1}{2}\|x - P_C(x)\|^2 = x - P_C(x)$.



(a) top-1 SVM_1 test accuracy (top-1 / top-2): 65.7% / 81.3% (b) top-2 $\mathrm{Ent}_{\mathrm{tr}}$ test accuracy (top-1 / top-2): 29.4%, 96.1%

Figure 4: Synthetic data on the unit circle in $\mathbb{R}^2$ (inside black circle) and visualization of top-1 and top-2 predictions (outside black circle). **(a)** Smooth top-1 SVM_1 optimizes top-1 error which impedes its top-2 error. **(b)** Trunc. top-2 entropy loss ignores top-1 scores and optimizes directly top-2 errors leading to a much better top-2 result.

4. Synthetic Example

In this section, we demonstrate in a synthetic experiment that our proposed top-2 losses outperform the top-1 losses when one aims at optimal top-2 performance. The dataset with three classes is shown in the inner circle of Figure 4.

Sampling. First, we generate samples in $[0, 7]$ which is subdivided into 5 segments. All segments have unit length, except for the 4-th segment which has length 3. We sample uniformly at random in each of the 5 segments according to the following class-conditional probabilities: $(0, 1, .4, .3, 0)$ for class 1, $(1, 0, .1, .7, 0)$ for class 2, and $(0, 0, .5, 0, 1)$ for class 3. Finally, the data is rescaled to $[0, 1]$ and mapped onto the unit circle.

Circle (synthetic)

Method	Top-1	Top-2	Method	Top-1	Top-2
$\mathrm{SVM}^{\mathrm{OVA}}$	54.3	85.8	top-1 SVM_1	**65.7**	83.9
$\mathrm{LR}^{\mathrm{OVA}}$	54.7	81.7	top-2 $\mathrm{SVM}_{0/1}$	54.4 / 54.5	87.1 / 87.0
$\mathrm{SVM}^{\mathrm{Multi}}$	58.9	89.3	top-2 Ent	54.6	87.6
$\mathrm{LR}^{\mathrm{Multi}}$	54.7	81.7	top-2 $\mathrm{Ent}_{\mathrm{tr}}$	58.4	**96.1**

Table 2: Top-k accuracy (%) on synthetic data. **Left:** Baselines methods. **Right:** Top-k SVM (nonsmooth / smooth) and top-k softmax losses (convex and nonconvex).

Samples of different classes are plotted next to each other for better visibility as there is significant class overlap. We visualize top-1/2 predictions with two colored circles (outside the black circle). We sample 200/200/200K points for training/validation/testing and tune the $C = \frac{1}{\lambda n}$ parameter in the range 2^{-18} to 2^{18}. Results are in Table 2.

	ALOI				**Letter**				**News 20**				**Caltech 101 Silhouettes**			
State-of-the-art	93 ± 1.2 [44]				97.98 [23] (RBF kernel)				86.9 [42]				62.1	79.6	83.4	[51]
Method	Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10
$\mathrm{SVM}^{\mathrm{OVA}}$	82.4	89.5	91.5	93.7	63.0	82.0	88.1	94.6	84.3	95.4	97.9	**99.5**	61.8	76.5	80.8	86.6
$\mathrm{LR}^{\mathrm{OVA}}$	86.1	93.0	94.8	96.6	68.1	86.1	90.6	96.2	84.9	96.3	97.8	99.3	63.2	80.4	84.4	89.4
$\mathrm{SVM}^{\mathrm{Multi}}$	90.0	95.1	96.7	98.1	76.5	89.2	93.1	97.7	85.4	94.9	97.2	99.1	62.8	77.8	82.0	86.9
$\mathrm{LR}^{\mathrm{Multi}}$	89.8	95.7	97.1	98.4	75.3	90.3	94.3	98.0	84.5	96.4	98.1	**99.5**	63.2	**81.2**	85.1	89.7
top-3 SVM	89.2	95.5	97.2	98.4	74.0	91.0	94.4	97.8	85.1	96.6	98.2	99.3	63.4	79.7	83.6	88.3
top-5 SVM	87.3	95.6	97.4	98.6	70.8	**91.5**	95.1	98.4	84.3	96.7	98.4	99.3	63.3	80.0	84.3	88.7
top-10 SVM	85.0	95.5	97.3	**98.7**	61.6	88.9	96.0	99.6	82.7	96.5	98.4	99.3	63.0	80.5	84.6	89.1
top-1 SVM_1	**90.6**	95.5	96.7	98.2	**76.8**	89.9	93.6	97.6	**85.6**	96.3	98.0	99.3	**63.9**	80.3	84.0	89.0
top-3 SVM_1	89.6	95.7	97.3	98.4	74.1	90.9	94.5	97.9	85.1	96.6	98.4	99.4	63.3	80.1	84.0	89.2
top-5 SVM_1	87.6	95.7	**97.5**	98.6	70.8	**91.5**	95.2	98.6	84.5	96.7	98.4	99.4	63.3	80.5	84.5	89.1
top-10 SVM_1	85.2	95.6	97.4	**98.7**	61.7	89.1	95.9	**99.7**	82.9	96.5	98.4	**99.5**	63.1	80.5	84.8	89.1
top-3 Ent	89.0	95.8	97.2	98.4	73.0	90.8	94.9	98.5	84.7	96.6	98.3	99.4	63.3	81.1	85.0	89.9
top-5 Ent	87.9	95.8	97.2	98.4	69.7	90.9	95.1	98.8	84.3	**96.8**	**98.6**	99.4	63.2	80.9	85.2	89.9
top-10 Ent	86.0	95.6	97.3	98.5	65.0	89.7	**96.2**	99.6	82.7	96.4	98.5	99.4	62.5	80.8	**85.4**	90.1
top-3 $\mathrm{Ent}_{\mathrm{tr}}$	89.3	**95.9**	97.3	98.5	63.6	91.1	95.6	98.8	83.4	96.4	98.3	99.4	60.7	81.1	85.2	**90.2**
top-5 $\mathrm{Ent}_{\mathrm{tr}}$	87.9	95.7	97.3	98.6	50.3	87.7	96.1	99.4	83.2	96.0	98.2	99.4	58.3	79.8	85.2	**90.2**
top-10 $\mathrm{Ent}_{\mathrm{tr}}$	85.2	94.8	97.1	98.5	46.5	80.9	93.7	99.6	82.9	95.7	97.9	99.4	51.9	78.4	84.6	**90.2**

	Indoor 67				**CUB**				**Flowers**				**FMD**		
State-of-the-art	82.0 [58]				62.8 [13] / 76.37 [61]				86.8 [40]				77.4 [13] / 82.4 [13]		
Method	Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5
$\mathrm{SVM}^{\mathrm{OVA}}$	81.9	94.3	96.5	98.0	60.6	77.1	83.4	89.9	82.0	91.7	94.3	96.8	77.4	92.4	96.4
$\mathrm{LR}^{\mathrm{OVA}}$	82.0	94.9	97.2	98.7	62.3	80.5	87.4	93.5	82.6	92.2	94.8	97.6	79.6	94.2	**98.2**
$\mathrm{SVM}^{\mathrm{Multi}}$	82.5	**95.4**	97.3	99.1	61.0	79.2	85.7	92.3	82.5	92.2	94.8	96.4	77.6	93.8	97.2
$\mathrm{LR}^{\mathrm{Multi}}$	82.4	95.2	**98.0**	99.1	62.3	81.7	87.9	**93.9**	82.9	92.4	95.1	97.8	79.0	94.6	97.8
top-3 SVM	81.6	95.1	97.7	99.0	61.3	80.4	86.3	92.5	81.9	92.2	95.0	96.1	78.8	94.6	97.8
top-5 SVM	79.9	95.0	97.7	99.0	60.9	81.2	87.2	92.9	81.7	92.4	95.1	97.8	78.4	94.4	97.6
top-10 SVM	78.4	95.1	97.4	99.0	59.6	81.3	87.7	93.4	80.5	91.9	95.1	97.7			
top-1 SVM_1	**82.6**	95.2	97.6	99.0	61.9	80.2	86.9	93.1	**83.0**	92.4	95.1	97.6	78.6	93.8	98.0
top-3 SVM_1	81.6	95.1	97.8	99.0	61.9	81.1	86.6	93.2	82.5	92.3	95.2	97.7	79.0	94.4	98.0
top-5 SVM_1	80.4	95.1	97.8	99.1	61.3	81.3	87.4	92.9	82.0	**92.5**	95.1	97.8	79.4	94.4	97.6
top-10 SVM_1	78.3	95.1	97.5	99.0	59.8	81.4	87.8	93.4	80.6	91.9	95.1	97.7			
top-3 Ent	81.4	**95.4**	97.6	**99.2**	**62.5**	81.8	87.9	**93.9**	82.5	92.0	**95.3**	97.8	**79.8**	94.8	98.0
top-5 Ent	80.3	95.0	97.7	99.0	62.0	**81.9**	88.1	93.8	82.1	92.2	95.1	**97.9**	79.4	94.4	98.0
top-10 Ent	79.2	95.1	97.6	99.0	61.2	81.6	**88.2**	93.8	80.9	92.1	95.0	97.7			
top-3 $\mathrm{Ent}_{\mathrm{tr}}$	79.8	95.0	97.5	99.1	62.0	81.4	87.6	93.4	82.1	92.2	95.2	97.6	78.4	**95.4**	**98.2**
top-5 $\mathrm{Ent}_{\mathrm{tr}}$	76.4	94.3	97.3	99.0	61.4	81.2	87.7	93.7	81.4	92.0	95.0	97.7	77.2	94.0	97.8
top-10 $\mathrm{Ent}_{\mathrm{tr}}$	72.6	92.8	97.1	98.9	59.7	80.7	87.2	93.4	77.9	91.1	94.3	97.3			

	SUN 397 (10 splits)				**Places 205** (val)				**ILSVRC 2012** (val)			
State-of-the-art	66.9 [58]				60.6		88.5	[58]	76.3		93.2	[50]
Method	Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10
$\mathrm{SVM}^{\mathrm{Multi}}$	65.8 ± 0.1	85.1 ± 0.2	90.8 ± 0.1	95.3 ± 0.1	58.4	78.7	84.7	89.9	68.3	82.9	87.0	91.1
$\mathrm{LR}^{\mathrm{Multi}}$	**67.5 ± 0.1**	**87.7 ± 0.2**	**92.9 ± 0.1**	**96.8 ± 0.1**	59.0	**80.6**	**87.6**	**94.3**	67.2	83.2	87.7	92.2
top-3 SVM	66.5 ± 0.2	86.5 ± 0.1	91.8 ± 0.1	95.9 ± 0.1	58.6	80.3	87.3	93.3	68.2	84.0	88.1	92.1
top-5 SVM	66.3 ± 0.2	87.0 ± 0.2	92.2 ± 0.2	96.3 ± 0.1	58.4	80.5	87.4	94.0	67.8	**84.1**	88.2	92.4
top-10 SVM	64.8 ± 0.3	87.2 ± 0.2	92.6 ± 0.1	96.6 ± 0.1	58.0	80.4	87.4	**94.3**	67.0	83.8	88.3	**92.6**
top-1 SVM_1	67.4 ± 0.2	86.8 ± 0.1	92.0 ± 0.1	96.1 ± 0.1	**59.2**	80.5	87.3	93.8	**68.7**	83.9	88.0	92.1
top-3 SVM_1	67.0 ± 0.2	87.0 ± 0.1	92.2 ± 0.1	96.2 ± 0.0	58.9	80.5	**87.6**	93.9	68.2	**84.1**	88.2	92.3
top-5 SVM_1	66.5 ± 0.2	87.2 ± 0.1	92.4 ± 0.2	96.3 ± 0.0	58.5	80.5	87.5	94.1	67.9	**84.1**	**88.4**	92.5
top-10 SVM_1	64.9 ± 0.3	87.3 ± 0.2	92.6 ± 0.2	96.6 ± 0.1	58.0	80.4	87.5	**94.3**	67.1	83.8	88.3	**92.6**
top-3 Ent	67.2 ± 0.2	**87.7 ± 0.2**	**92.9 ± 0.1**	**96.8 ± 0.1**	58.7	**80.6**	**87.6**	94.2	66.8	83.1	87.8	92.2
top-5 Ent	66.6 ± 0.3	**87.7 ± 0.2**	**92.9 ± 0.1**	**96.8 ± 0.1**	58.1	80.4	87.4	94.2	66.5	83.0	87.7	92.2
top-10 Ent	65.2 ± 0.3	87.4 ± 0.1	92.8 ± 0.1	**96.8 ± 0.1**	57.0	80.0	87.2	94.1	65.8	82.8	87.6	92.1

Table 3: Top-k accuracy (%) on various datasets. The first line is a reference to the state-of-the-art on each dataset and reports top-1 accuracy except when the numbers are aligned with Top-k. We compare the one-vs-all and multiclass baselines with the top-k SVM^{α} [28] as well as the proposed smooth top-k $\mathrm{SVM}^{\alpha}_{\gamma}$, top-$k$ Ent, and the nonconvex top-k $\mathrm{Ent}_{\mathrm{tr}}$.

In each column we provide the results for the model that optimizes the corresponding top-k accuracy, which is in general different for top-1 and top-2. First, we note that all top-1 baselines perform similar in top-1 performance, except for $\text{SVM}^{\text{Multi}}$ and top-1 SVM_1 which show better results. Next, we see that our top-2 losses improve the top-2 accuracy and the improvement is most significant for the nonconvex top-2 Ent_{tr} loss, which is close to the optimal solution for this dataset. This is because top-2 Ent_{tr} is a tight bound on the top-2 error and ignores top-1 errors in the loss. Unfortunately, similar significant improvements were not observed on the real-world data sets that we tried.

5. Experimental Results

The goal of this section is to provide an extensive empirical evaluation of the top-k performance of different losses in multiclass classification. To this end, we evaluate the loss functions introduced in § 2 on 11 datasets (500 to 2.4M training examples, 10 to 1000 classes), from various problem domains (vision and non-vision; fine-grained, scene and general object classification). The detailed statistics of the datasets is given in Table 4.

Dataset	m	n	d	Dataset	m	n	d
ALOI [44]	1K	54K	128	Indoor 67 [38]	67	5354	4K
Caltech 101 Sil [51]	101	4100	784	Letter [23]	26	10.5K	16
CUB [57]	202	5994	4K	News 20 [27]	20	15.9K	16K
Flowers [35]	102	2040	4K	Places 205 [62]	205	2.4M	4K
FMD [49]	10	500	4K	SUN 397 [60]	397	19.9K	4K
ILSVRC 2012 [47]	1K	1.3M	4K				

Table 4: Statistics of the datasets used in the experiments (m – # classes, n – # training examples, d – # features).

Please refer to Table 1 for an overview of the methods and our naming convention. A broad selection of results is also reported at the end of the paper. As other ranking based losses did not perform well in [28], we do no further comparison here.

Solvers. We use `LibLinear` [17] for the one-vs-all baselines SVM^{OVA} and LR^{OVA}; and our code from [28] for top-k SVM. We extended the latter to support the smooth top-k SVM_γ and top-k Ent. The multiclass loss baselines $\text{SVM}^{\text{Multi}}$ and LR^{Multi} correspond respectively to top-1 SVM and top-1 Ent. For the nonconvex top-k Ent_{tr}, we use the LR^{Multi} solution as an initial point and perform gradient descent with line search. We cross-validate hyper-parameters in the range 10^{-5} to 10^{3}, extending it when the optimal value is at the boundary.

Features. For ALOI, Letter, and News20 datasets, we use the features provided by the `LibSVM` [12] datasets. For ALOI, we randomly split the data into equally sized training and test sets preserving class distributions. The Letter dataset comes with a separate validation set, which we used for model selection only. For News20, we use PCA to reduce dimensionality of sparse features from 62060 to 15478 preserving all non-singular PCA components[4].

For Caltech101 Silhouettes, we use the features and the train/val/test splits provided by [51].

For CUB, Flowers, FMD, and ILSVRC 2012, we use `MatConvNet` [56] to extract the outputs of the last fully connected layer of the `imagenet-vgg-verydeep-16` model which is pre-trained on ImageNet [16] and achieves state-of-the-art results in image classification [50].

For Indoor 67, SUN 397, and Places 205, we use the `Places205-VGGNet-16` model by [58] which is pre-trained on Places 205 [62] and outperforms the ImageNet pre-trained model on scene classification tasks [58]. Further results can be found at the end of the paper. In all cases we obtain a similar behavior in terms of the ranking of the considered losses as discussed below.

Discussion. The experimental results are given in Table 3. There are several interesting observations that one can make. While the OVA schemes perform quite similar to the multiclass approaches (logistic OVA vs. softmax, hinge OVA vs. multiclass SVM), which confirms earlier observations in [2, 43], the OVA schemes performed worse on ALOI and Letter. Therefore it seems safe to recommend to use multiclass losses instead of the OVA schemes.

Comparing the softmax vs. multiclass SVM losses, we see that there is no clear winner in top-1 performance, but softmax consistently outperforms multiclass SVM in top-k performance for $k > 1$. This might be due to the strong property of softmax being top-k calibrated for all k. Please note that this trend is uniform across all datasets, in particular, also for the ones where the features are not coming from a convnet. Both the smooth top-k hinge and the top-k entropy losses perform slightly better than softmax if one compares the corresponding top-k errors. However, the good performance of the truncated top-k loss on synthetic data does not transfer to the real world datasets. This might be due to a relatively high dimension of the feature spaces, but requires further investigation.

Fine-tuning experiments. We also performed a number of fine-tuning experiments where the original network was trained further for 1-3 epochs with the smooth top-k hinge and the truncated top-k entropy losses[5]. The motivation was to see if the full end-to-end training would be more beneficial compared to training just the classifier. Results are reported in Table 5. First, we note that the setting is now slightly different: there is no feature extraction step with the `MatConvNet` and there is a non-regularized bias term in Caffe [24]. Next, we see that the top-k specific losses are able to improve the performance compared to the reference model, and that the top-5 SVM_1 loss achieves the best

[4] The top-k SVM solvers that we used were designed for dense inputs.
[5] Code: https://github.com/mlapin/caffe/tree/topk

Places 205 (val)

Method	Top-1	Top-3	Top-5	Top-10
LR^{Multi}	59.97	81.39	88.17	94.59
top-3 SVM_1 (FT)	60.73	82.09	88.58	94.56
top-5 SVM_1 (FT)	**60.88**	**82.18**	**88.78**	94.75
top-3 Ent_{tr} (FT)	60.51	81.86	88.69	94.78
top-5 Ent_{tr} (FT)	60.48	81.66	88.66	94.80
LR^{Multi} (FT)	60.73	82.07	88.71	**94.82**

ILSVRC 2012 (val)

Method	Top-1	Top-3	Top-5	Top-10
LR^{Multi}	68.60	84.29	88.66	92.83
top-3 SVM_1 (FT)	71.66	86.63	90.55	94.17
top-5 SVM_1 (FT)	71.60	86.67	90.56	94.23
top-3 Ent_{tr} (FT)	71.41	86.80	90.77	94.35
top-5 Ent_{tr} (FT)	71.20	86.57	90.75	**94.38**
LR^{Multi} (FT)	**72.11**	**87.08**	**90.88**	**94.38**

Table 5: Top-k accuracy (%), as reported by Caffe [24], on large scale datasets after fine-tuning (FT) for approximately one epoch on Places and 3 epochs on ILSVRC. The first line (LR^{Multi}) is the reference performance w/o fine-tuning.

top-1..5 performance on Places 205. However, in this set of experiments, we also observed similar improvements when fine-tuning with the standard softmax loss, which achieves best performance on ILSVRC 2012.

We conclude that a safe choice for multiclass problems seems to be the softmax loss as it yields competitive results in all top-k errors. An interesting alternative is the smooth top-k hinge loss which is faster to train (see Section 3) and achieves competitive performance. If one wants to optimize directly for a top-k error (at the cost of a higher top-1 error), then further improvements are possible using either the smooth top-k SVM or the top-k entropy losses.

6. Conclusion

We have done an extensive experimental study of top-k performance optimization. We observed that the softmax loss and the smooth top-1 hinge loss are competitive across all top-k errors and should be considered the primary candidates in practice. Our new top-k loss functions can further improve these results slightly, especially if one is targeting a particular top-k error as the performance measure. Finally, we would like to highlight our new optimization scheme based on SDCA for the top-k entropy loss which also includes the softmax loss and is of an independent interest.

References

[1] S. Agarwal. The infinite push: A new support vector ranking algorithm that directly optimizes accuracy at the absolute top of the list. In *SDM*, pages 839–850, 2011. 1

[2] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Good practice in large-scale learning for image classification. *PAMI*, 36(3):507–520, 2014. 1, 4, 16

[3] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006. 3

[4] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. 14

[5] A. Beck and M. Teboulle. Smoothing and first order methods, a unified framework. *SIAM Journal on Optimization*, 22:557–580, 2012. 6

[6] Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009. 4

[7] A. Bordes, L. Bottou, P. Gallinari, and J. Weston. Solving multiclass support vector machines with LaRank. In *ICML*, pages 89–96, 2007. 1

[8] J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Cms Books in Mathematics Series. Springer Verlag, 2000. 11, 14

[9] S. Boyd, C. Cortes, M. Mohri, and A. Radovanovic. Accuracy at the top. In *NIPS*, pages 953–961, 2012. 1

[10] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. 8, 9, 10

[11] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *ICML*, pages 89–96, 2005. 1

[12] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:1–27, 2011. 16

[13] M. Cimpoi, S. Maji, and A. Vedaldi. Deep filter banks for texture recognition and segmentation. In *CVPR*, 2015. 15

[14] R. M. Corless, G. H. Gonnet, D. E. Hare, D. J. Jeffrey, and D. E. Knuth. On the lambert W function. *Advances in Computational Mathematics*, 5(1):329–359, 1996. 12, 13

[15] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR*, 2:265–292, 2001. 4, 7

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 16

[17] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 9:1871–1874, 2008. 16

[18] T. Fukushima. Precise and fast computation of Lambert W-functions without transcendental function evaluations. *Journal of Computational and Applied Mathematics*, 244:77 – 89, 2013. 12, 13

[19] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *ICCV*, pages 221–228, 2009. 23, 24

[20] M. R. Gupta, S. Bengio, and J. Weston. Training highly multiclass classifiers. *JMLR*, 15:1461–1492, 2014. 1

[21] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, Berlin, 2001. 7

[22] A. S. Householder. *The Numerical Treatment of a Single Nonlinear Equation*. McGraw-Hill, 1970. 13

[23] C.-W. Hsu and C.-J. Lin. A comparison of methods for multi-class support vector machines. *Neural Networks*, 13(2):415–425, 2002. 15, 16

[24] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 16, 17

[25] T. Joachims. A support vector method for multivariate performance measures. In *ICML*, pages 377–384, 2005. 1

[26] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012. 4

[27] K. Lang. Newsweeder: Learning to filter netnews. In *ICML*, pages 331–339, 1995. 16

[28] M. Lapin, M. Hein, and B. Schiele. Top-k multiclass SVM. In *NIPS*, 2015. 1, 2, 5, 7, 8, 11, 12, 14, 15, 16, 28, 32, 34, 36

[29] M. Lapin, B. Schiele, and M. Hein. Scalable multitask representation learning for scene classification. In *CVPR*, 2014. 31

[30] N. Li, R. Jin, and Z.-H. Zhou. Top rank optimization in linear time. In *NIPS*, pages 1502–1510, 2014. 1

[31] L. Liu, T. G. Dietterich, N. Li, and Z. Zhou. Transductive optimization of top k precision. *CoRR*, abs/1510.05976, 2015. 1

[32] J. Mairal. *Sparse Coding for Machine Learning, Image Processing and Computer Vision*. PhD thesis, Ecole Normale Superieure de Cachan, 2010. 7

[33] J. Mairal, R. Jenatton, F. R. Bach, and G. R. Obozinski. Network flow algorithms for structured sparsity. In *NIPS*, pages 1558–1566, 2010. 14

[34] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005. 6

[35] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pages 722–729, 2008. 16

[36] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Science+ Business Media, 2006. 13

[37] K. B. Petersen, M. S. Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 450:7–15, 2008. 8

[38] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009. 16

[39] A. Rakotomamonjy. Sparse support vector infinite push. In *ICML*, pages 1335–1342. ACM, 2012. 1

[40] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In *CVPRW, DeepVision workshop*, 2014. 15

[41] M. Reid and B. Williamson. Composite binary losses. *JMLR*, 11:2387–2422, 2010. 5

[42] J. D. Rennie. Improving multi-class text classification with naive bayes. Technical report, Massachusetts Institute of Technology, 2001. 15

[43] R. Rifkin and A. Klautau. In defense of one-vs-all classification. *JMLR*, 5:101–141, 2004. 16

[44] A. Rocha and S. Klein Goldenstein. Multiclass from binary: Expanding one-versus-all, one-versus-one and ecoc-based approaches. *Neural Networks and Learning Systems, IEEE Transactions on*, 25(2):289–302, 2014. 15, 16

[45] S. Ross, J. Zhou, Y. Yue, D. Dey, and D. Bagnell. Learning policies for contextual submodular prediction. In *ICML*, pages 1364–1372, 2013. 1

[46] C. Rudin. The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list. *JMLR*, 10:2233–2271, 2009. 1

[47] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, 2014. 1, 16

[48] S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, pages 1–41, 2014. 1, 2, 6, 7, 8, 11, 12

[49] L. Sharan, R. Rosenholtz, and E. Adelson. Material perception: What can you see in a brief glance? *Journal of Vision*, 9(8):784–784, 2009. 16

[50] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 4, 15, 16

[51] K. Swersky, B. J. Frey, D. Tarlow, R. S. Zemel, and R. P. Adams. Probabilistic n-choose-k models for classification and ranking. In *NIPS*, pages 3050–3058, 2012. 1, 15, 16

[52] A. Tewari and P. Bartlett. On the consistency of multiclass classification methods. *JMLR*, 8:1007–1025, 2007. 3, 4

[53] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, pages 1453–1484, 2005. 1

[54] N. Usunier, D. Buffoni, and P. Gallinari. Ranking with ordered weighted pairwise classification. In *ICML*, pages 1057–1064, 2009. 1, 5, 10

[55] D. Veberič. Lambert W function for applications in physics. *Computer Physics Communications*, 183(12):2622–2628, 2012. 12, 13

[56] A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab. In *Proceeding of the ACM Int. Conf. on Multimedia*, 2015. 16

[57] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical report, California Institute of Technology, 2011. 16

[58] L. Wang, S. Guo, W. Huang, and Y. Qiao. Places205-vggnet models for scene recognition. *CoRR*, abs/1508.01667, 2015. 15, 16

[59] J. Weston, S. Bengio, and N. Usunier. Wsabie: scaling up to large vocabulary image annotation. *IJCAI*, pages 2764–2770, 2011. 1

[60] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 1, 16

[61] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based rcnn for fine-grained detection. In *ECCV*, 2014. 15

[62] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014. 1, 16

ALOI

Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
$\mathrm{SVM}^{\mathrm{OVA}}$	82.4	87.4	89.5	90.7	91.5	92.1	92.6	93.1	93.4	93.7
$\mathrm{LR}^{\mathrm{OVA}}$	86.1	91.1	93.0	94.1	94.8	95.4	95.8	96.1	96.4	96.6
top-1 SVM^{α} / $\mathrm{SVM}^{\mathrm{Multi}}$	90.0	93.4	95.1	96.0	96.7	97.1	97.5	97.7	97.9	98.1
top-2 SVM^{α}	90.0	94.0	95.5	96.4	97.0	97.4	97.7	97.9	98.1	98.3
top-3 SVM^{α}	89.2	94.2	95.5	96.7	97.2	97.6	97.8	98.1	98.2	98.4
top-4 SVM^{α}	88.4	94.3	95.6	96.8	97.4	97.8	98.0	98.2	98.4	98.5
top-5 SVM^{α}	87.3	94.1	95.6	96.9	97.4	97.8	98.0	98.3	98.4	98.6
top-10 SVM^{α}	85.0	93.0	95.5	96.5	97.3	97.8	98.2	98.4	98.6	98.7
top-1 SVM_1^{α}	90.6	94.2	95.5	96.2	96.7	97.0	97.6	97.9	98.1	98.2
top-2 SVM_1^{α}	90.3	94.3	95.6	96.3	96.8	97.1	97.7	98.0	98.2	98.3
top-3 SVM_1^{α}	89.6	94.4	95.7	96.7	97.3	97.6	97.9	98.1	98.3	98.4
top-4 SVM_1^{α}	88.7	94.4	95.7	96.9	97.4	97.8	98.0	98.2	98.4	98.5
top-5 SVM_1^{α}	87.6	94.3	95.7	96.9	97.5	97.8	98.1	98.3	98.4	98.6
top-10 SVM_1^{α}	85.2	93.1	95.6	96.6	97.4	97.8	98.2	98.4	98.6	98.7
top-1 SVM^{β} / $\mathrm{SVM}^{\mathrm{Multi}}$	90.0	93.4	95.1	96.0	96.7	97.1	97.5	97.7	97.9	98.1
top-2 SVM^{β}	90.2	93.9	95.2	96.0	96.7	97.1	97.5	97.7	97.9	98.1
top-3 SVM^{β}	90.2	94.1	95.4	96.0	96.5	96.9	97.5	97.8	98.0	98.1
top-4 SVM^{β}	90.1	94.2	95.4	96.1	96.6	97.0	97.3	97.5	98.0	98.2
top-5 SVM^{β}	90.0	94.3	95.5	96.2	96.7	97.1	97.3	97.5	97.7	98.2
top-10 SVM^{β}	89.5	94.2	95.7	96.5	96.9	97.3	97.5	98.0	97.9	98.3
top-1 SVM_1^{β}	90.6	94.2	95.5	96.2	96.7	97.0	97.6	97.9	98.1	98.2
top-2 SVM_1^{β}	90.6	94.2	95.5	96.2	96.7	97.0	97.3	97.9	98.1	98.2
top-3 SVM_1^{β}	90.4	94.3	95.6	96.3	96.7	97.1	97.4	97.6	98.1	98.2
top-4 SVM_1^{β}	90.3	94.4	95.6	96.3	96.8	97.1	97.4	97.6	97.8	97.9
top-5 SVM_1^{β}	90.2	94.4	95.7	96.3	96.8	97.2	97.5	97.7	97.8	98.0
top-10 SVM_1^{β}	89.5	94.3	95.7	96.6	97.0	97.4	97.6	97.8	98.1	98.2
top-1 Ent / $\mathrm{LR}^{\mathrm{Multi}}$	89.8	94.2	95.7	96.5	97.1	97.5	97.8	98.0	98.2	98.4
top-2 Ent	89.4	94.2	95.8	96.6	97.1	97.5	97.8	98.0	98.2	98.4
top-3 Ent	89.0	94.3	95.8	96.6	97.2	97.5	97.8	98.0	98.2	98.4
top-4 Ent	88.5	94.2	95.8	96.7	97.2	97.6	97.8	98.1	98.3	98.4
top-5 Ent	87.9	94.2	95.8	96.7	97.2	97.6	97.9	98.1	98.3	98.4
top-10 Ent	86.0	93.2	95.6	96.7	97.3	97.7	98.0	98.2	98.4	98.5
top-2 $\mathrm{Ent}_{\mathrm{tr}}$	89.8	94.4	95.9	96.7	97.2	97.6	97.9	98.1	98.3	98.5
top-3 $\mathrm{Ent}_{\mathrm{tr}}$	89.3	94.3	95.9	96.7	97.3	97.7	98.0	98.2	98.3	98.5
top-4 $\mathrm{Ent}_{\mathrm{tr}}$	88.7	94.0	95.8	96.7	97.3	97.7	98.0	98.2	98.4	98.5
top-5 $\mathrm{Ent}_{\mathrm{tr}}$	87.9	93.7	95.7	96.7	97.3	97.7	98.0	98.2	98.4	98.6
top-10 $\mathrm{Ent}_{\mathrm{tr}}$	85.2	92.4	94.8	96.3	97.1	97.6	98.0	98.2	98.4	98.5

Table 6: Comparison of different methods in top-k accuracy (%).

Letter

Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
$\mathrm{SVM}^{\mathrm{OVA}}$	63.0	75.7	82.0	85.7	88.1	89.9	91.4	92.8	93.9	94.6
$\mathrm{LR}^{\mathrm{OVA}}$	68.1	81.1	86.1	88.4	90.6	92.2	93.4	94.6	95.3	96.2
top-1 SVM^{α} / $\mathrm{SVM}^{\mathrm{Multi}}$	76.5	85.5	89.2	91.5	93.1	94.3	95.5	96.5	97.0	97.7
top-2 SVM^{α}	76.1	86.9	90.1	92.2	93.3	94.8	96.0	96.5	97.2	97.7
top-3 SVM^{α}	74.0	87.0	91.0	93.0	94.4	95.4	96.2	96.7	97.3	97.8
top-4 SVM^{α}	71.4	86.4	91.4	93.5	94.8	95.7	96.6	97.2	97.6	98.2
top-5 SVM^{α}	70.8	85.9	91.5	93.9	95.1	96.2	96.9	97.5	98.1	98.4
top-10 SVM^{α}	61.6	82.5	88.9	93.6	96.0	97.6	98.3	98.9	99.2	99.6
top-1 SVM_1^{α}	76.8	86.0	89.9	92.1	93.6	94.9	95.8	96.3	97.0	97.6
top-2 SVM_1^{α}	76.2	87.0	90.3	92.5	94.0	94.9	96.0	96.6	97.3	97.7
top-3 SVM_1^{α}	74.1	87.1	90.9	93.2	94.5	95.6	96.4	96.9	97.3	97.9
top-4 SVM_1^{α}	72.1	86.7	91.4	93.2	94.7	95.7	96.7	97.3	97.8	98.0
top-5 SVM_1^{α}	70.8	86.2	91.5	93.8	95.2	96.3	97.0	97.6	98.2	98.6
top-10 SVM_1^{α}	61.7	82.9	89.1	93.6	95.9	97.5	98.3	98.9	99.3	99.7
top-1 SVM^{β} / $\mathrm{SVM}^{\mathrm{Multi}}$	76.5	85.5	89.2	91.5	93.1	94.3	95.5	96.5	97.0	97.7
top-2 SVM^{β}	76.5	86.4	90.2	92.1	93.8	94.8	95.9	96.7	97.3	97.8
top-3 SVM^{β}	75.6	86.9	90.6	92.7	94.2	95.3	95.8	96.7	97.4	97.9
top-4 SVM^{β}	74.9	86.9	90.9	93.1	94.6	95.5	96.3	96.9	97.5	98.1
top-5 SVM^{β}	74.5	86.9	91.0	93.4	94.9	95.7	96.5	97.2	97.8	98.2
top-10 SVM^{β}	72.9	85.9	90.8	93.5	95.3	96.2	97.2	97.8	98.3	98.8
top-1 SVM_1^{β}	76.8	86.0	89.9	92.1	93.6	94.9	95.8	96.3	97.0	97.6
top-2 SVM_1^{β}	76.5	86.6	90.2	92.4	93.8	95.0	96.0	96.8	97.4	97.8
top-3 SVM_1^{β}	75.6	86.9	90.6	92.8	94.2	95.3	96.1	96.6	97.3	97.9
top-4 SVM_1^{β}	75.1	87.0	90.8	93.1	94.6	95.5	96.3	97.0	97.8	98.1
top-5 SVM_1^{β}	74.6	87.0	91.1	93.4	94.9	95.8	96.6	97.2	97.8	98.2
top-10 SVM_1^{β}	72.9	86.1	90.8	93.6	95.2	96.3	97.1	97.9	98.3	98.8
top-1 Ent / $\mathrm{LR}^{\mathrm{Multi}}$	75.2	86.5	90.1	93.1	94.5	95.7	96.5	97.4	98.1	98.1
top-2 Ent	74.6	86.5	90.6	93.2	94.7	95.7	96.5	97.5	98.0	98.4
top-3 Ent	73.0	86.4	90.8	93.3	94.9	95.9	96.8	97.6	98.2	98.5
top-4 Ent	71.9	86.1	91.0	93.6	95.3	96.3	97.1	97.7	98.3	98.6
top-5 Ent	69.7	85.6	90.9	93.7	95.1	96.7	97.3	97.9	98.6	98.8
top-10 Ent	65.0	82.5	89.7	93.6	96.2	97.7	98.4	98.9	99.3	99.6
top-2 $\mathrm{Ent}_{\mathrm{tr}}$	70.1	86.3	90.7	93.7	95.1	96.1	97.0	97.7	98.1	98.5
top-3 $\mathrm{Ent}_{\mathrm{tr}}$	63.6	83.9	91.1	93.9	95.6	96.7	97.5	98.1	98.5	98.8
top-4 $\mathrm{Ent}_{\mathrm{tr}}$	58.7	80.3	90.0	93.9	96.1	97.2	98.0	98.5	98.9	99.3
top-5 $\mathrm{Ent}_{\mathrm{tr}}$	50.3	76.4	87.7	93.7	96.1	97.5	98.3	98.7	99.2	99.4
top-10 $\mathrm{Ent}_{\mathrm{tr}}$	46.5	67.9	80.9	88.9	93.7	96.3	97.7	98.7	99.3	99.6

Table 7: Comparison of different methods in top-k accuracy (%).

News 20

Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
$\mathrm{SVM}^{\mathrm{OVA}}$	84.3	93.0	95.4	97.0	97.9	98.5	98.8	99.0	99.3	99.5
$\mathrm{LR}^{\mathrm{OVA}}$	84.9	93.1	96.3	97.2	97.8	98.3	98.7	98.8	99.0	99.3
top-1 SVM^{α} / $\mathrm{SVM}^{\mathrm{Multi}}$	85.4	92.7	94.9	96.4	97.2	97.6	98.2	98.5	98.7	99.1
top-2 SVM^{α}	85.3	94.3	96.4	97.2	97.9	98.4	98.5	98.9	99.0	99.1
top-3 SVM^{α}	85.1	94.8	96.6	97.7	98.2	98.6	99.0	99.2	99.3	99.3
top-4 SVM^{α}	85.0	94.5	96.7	97.7	98.3	98.6	98.9	99.1	99.3	99.4
top-5 SVM^{α}	84.3	94.2	96.7	97.8	98.4	98.7	99.0	99.2	99.3	99.3
top-10 SVM^{α}	82.7	93.3	96.5	97.7	98.4	98.6	99.0	99.2	99.3	99.3
top-1 SVM^{α}_1	85.6	94.5	96.3	97.3	98.0	98.5	98.8	99.0	98.9	99.3
top-2 SVM^{α}_1	85.2	94.6	96.5	97.6	98.1	98.6	98.9	99.1	99.3	99.3
top-3 SVM^{α}_1	85.1	94.7	96.6	97.8	98.4	98.7	99.0	99.1	99.3	99.4
top-4 SVM^{α}_1	84.9	94.4	96.7	97.8	98.4	98.7	99.0	99.1	99.3	99.4
top-5 SVM^{α}_1	84.5	94.4	96.7	97.9	98.4	98.8	99.0	99.1	99.3	99.4
top-10 SVM^{α}_1	82.9	93.5	96.5	97.8	98.4	98.7	99.0	99.2	99.3	99.5
top-1 SVM^{β} / $\mathrm{SVM}^{\mathrm{Multi}}$	85.4	92.7	94.9	96.4	97.2	97.6	98.2	98.5	98.7	99.1
top-2 SVM^{β}	85.7	94.3	96.1	97.4	97.8	97.8	98.2	98.6	98.9	99.1
top-3 SVM^{β}	86.0	94.5	96.5	97.5	98.1	98.4	98.7	99.0	99.2	99.2
top-4 SVM^{β}	85.9	94.8	96.6	97.7	98.1	98.4	98.7	99.0	99.2	99.2
top-5 SVM^{β}	85.4	94.8	96.7	97.7	98.3	98.7	99.0	99.1	99.3	99.4
top-10 SVM^{β}	84.9	94.6	96.7	98.0	98.5	98.7	99.0	99.2	99.3	99.3
top-1 SVM^{β}_1	85.6	94.5	96.3	97.3	98.0	98.5	98.8	99.0	98.9	99.3
top-2 SVM^{β}_1	85.8	94.5	96.3	97.4	98.0	98.4	98.8	99.0	99.2	99.3
top-3 SVM^{β}_1	85.7	94.5	96.5	97.6	98.1	98.5	98.8	99.0	99.3	99.3
top-4 SVM^{β}_1	85.6	94.6	96.6	97.6	98.2	98.6	98.9	99.1	99.0	99.4
top-5 SVM^{β}_1	85.6	94.7	96.6	97.6	98.3	98.7	99.0	99.2	99.3	99.3
top-10 SVM^{β}_1	84.9	94.7	96.7	97.9	98.5	98.7	99.0	99.1	99.3	99.3
top-1 Ent / $\mathrm{LR}^{\mathrm{Multi}}$	84.5	94.2	96.4	97.6	98.1	98.5	99.0	99.1	99.3	99.5
top-2 Ent	84.7	94.4	96.5	97.7	98.3	98.6	98.9	99.2	99.3	99.3
top-3 Ent	84.7	94.6	96.6	97.8	98.3	98.7	98.9	99.2	99.3	99.4
top-4 Ent	84.5	94.5	96.7	97.9	98.5	98.7	99.0	99.2	99.3	99.4
top-5 Ent	84.3	94.3	96.8	97.8	98.6	98.8	99.0	99.1	99.3	99.4
top-10 Ent	82.7	93.3	96.4	97.8	98.5	98.7	98.9	99.1	99.3	99.4
top-2 $\mathrm{Ent}_{\mathrm{tr}}$	84.2	93.9	96.6	97.6	98.4	98.7	98.9	99.1	99.3	99.4
top-3 $\mathrm{Ent}_{\mathrm{tr}}$	83.4	93.6	96.4	97.7	98.3	98.6	98.9	99.2	99.2	99.4
top-4 $\mathrm{Ent}_{\mathrm{tr}}$	83.3	93.5	96.2	97.6	98.3	98.6	98.9	99.1	99.2	99.3
top-5 $\mathrm{Ent}_{\mathrm{tr}}$	83.2	93.3	96.0	97.7	98.2	98.5	98.9	99.1	99.2	99.4
top-10 $\mathrm{Ent}_{\mathrm{tr}}$	82.9	92.7	95.7	97.0	97.9	98.4	98.8	99.0	99.2	99.4

Table 8: Comparison of different methods in top-k accuracy (%).

Caltech 101 Silhouettes

Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
$\mathrm{SVM}^{\mathrm{OVA}}$	61.8	73.1	76.5	78.5	80.8	82.5	83.6	84.6	85.6	86.6
$\mathrm{LR}^{\mathrm{OVA}}$	63.2	77.1	80.4	82.6	84.4	85.7	87.0	87.9	88.6	89.4
top-1 SVM^{α} / $\mathrm{SVM}^{\mathrm{Multi}}$	62.8	74.6	77.8	80.0	82.0	83.3	84.4	85.0	85.9	86.9
top-2 SVM^{α}	63.1	76.2	79.0	81.0	82.7	84.4	85.5	86.2	87.0	87.6
top-3 SVM^{α}	63.4	76.7	79.7	81.5	83.6	85.1	85.8	86.6	87.6	88.3
top-4 SVM^{α}	63.2	76.6	79.8	82.4	84.0	85.3	86.0	86.9	87.8	88.6
top-5 SVM^{α}	63.3	76.8	80.0	82.7	84.3	85.6	86.5	87.5	88.1	88.7
top-10 SVM^{α}	63.0	77.3	80.5	82.7	84.6	85.9	86.8	88.0	88.8	89.1
top-1 SVM_1^{α}	63.9	76.9	80.3	82.4	84.0	85.4	86.4	87.2	88.4	89.0
top-2 SVM_1^{α}	63.8	76.9	80.5	82.3	84.1	85.3	86.5	87.4	88.4	89.0
top-3 SVM_1^{α}	63.3	77.2	80.1	82.3	84.0	85.7	86.5	87.3	88.0	89.2
top-4 SVM_1^{α}	63.1	77.0	80.3	82.4	84.3	85.5	86.6	87.3	88.2	89.2
top-5 SVM_1^{α}	63.3	77.1	80.5	82.7	84.5	85.9	86.7	87.6	88.5	89.1
top-10 SVM_1^{α}	63.1	77.2	80.5	82.8	84.8	86.0	87.0	88.1	88.5	89.1
top-1 SVM^{β} / $\mathrm{SVM}^{\mathrm{Multi}}$	62.8	74.6	77.8	80.0	82.0	83.3	84.4	85.0	85.9	86.9
top-2 SVM^{β}	63.5	76.2	79.3	81.1	82.6	84.3	85.3	86.3	87.0	87.9
top-3 SVM^{β}	63.9	76.6	79.7	81.4	83.4	84.9	85.7	86.6	87.4	88.0
top-4 SVM^{β}	63.9	76.9	80.1	82.0	83.5	85.0	85.9	87.0	87.8	88.7
top-5 SVM^{β}	63.6	77.0	80.4	82.6	84.2	85.3	86.3	87.4	88.1	89.0
top-10 SVM^{β}	64.0	77.1	80.5	83.0	84.9	86.2	87.3	87.9	88.8	89.4
top-1 SVM_1^{β}	63.9	76.9	80.3	82.4	84.0	85.4	86.4	87.2	88.4	89.0
top-2 SVM_1^{β}	63.9	77.1	80.4	82.3	84.1	85.5	86.5	87.3	88.4	89.0
top-3 SVM_1^{β}	64.0	77.0	80.4	82.4	84.2	85.4	86.6	87.3	88.4	89.1
top-4 SVM_1^{β}	63.7	77.0	80.2	82.6	84.4	85.7	86.7	87.4	88.4	89.0
top-5 SVM_1^{β}	63.7	77.2	80.5	82.4	84.4	85.9	86.9	87.6	88.4	89.1
top-10 SVM_1^{β}	64.2	77.3	80.6	83.2	85.0	86.2	87.1	88.0	88.8	89.5
top-1 Ent / $\mathrm{LR}^{\mathrm{Multi}}$	63.2	77.8	81.2	83.4	85.1	86.5	87.3	88.7	89.3	89.7
top-2 Ent	63.3	77.5	81.1	83.3	85.0	86.5	87.4	88.6	89.4	89.8
top-3 Ent	63.3	77.5	81.1	83.3	85.0	86.6	87.4	88.7	89.2	89.9
top-4 Ent	63.2	77.5	81.0	83.2	85.0	86.5	87.5	88.6	89.3	89.9
top-5 Ent	63.2	77.5	80.9	83.1	85.2	86.4	87.5	88.5	89.3	89.9
top-10 Ent	62.5	77.3	80.8	83.4	85.4	86.6	87.9	88.5	89.3	90.1
top-2 $\mathrm{Ent}_{\mathrm{tr}}$	62.7	77.5	81.3	83.5	85.5	86.9	87.8	88.7	89.3	90.0
top-3 $\mathrm{Ent}_{\mathrm{tr}}$	60.7	76.9	81.1	83.7	85.2	87.0	88.2	88.8	89.8	90.2
top-4 $\mathrm{Ent}_{\mathrm{tr}}$	60.0	75.9	80.4	83.1	85.1	86.9	88.2	88.9	89.7	90.5
top-5 $\mathrm{Ent}_{\mathrm{tr}}$	58.3	75.4	79.8	82.7	85.2	86.9	88.0	88.8	89.6	90.2
top-10 $\mathrm{Ent}_{\mathrm{tr}}$	51.9	72.1	78.4	82.0	84.6	86.2	87.5	88.5	89.4	90.2

Table 9: Comparison of different methods in top-k accuracy (%).

Caltech 101 (average of 39 kernels from [19], 5 splits)

Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
top-1 SVM^{α} / $\text{SVM}^{\text{Multi}}$	73.2 ± 1.0	81.6 ± 0.8	85.5 ± 0.8	87.8 ± 0.7	89.5 ± 0.6	90.8 ± 0.5	91.8 ± 0.4	92.7 ± 0.4	93.3 ± 0.4	93.8 ± 0.4
top-2 SVM^{α}	73.2 ± 0.9	81.7 ± 0.7	85.5 ± 0.9	87.8 ± 0.7	89.6 ± 0.6	90.9 ± 0.5	91.9 ± 0.4	92.7 ± 0.3	93.3 ± 0.4	93.9 ± 0.5
top-3 SVM^{α}	73.0 ± 1.0	81.6 ± 0.7	85.4 ± 0.9	87.8 ± 0.6	89.6 ± 0.6	90.9 ± 0.4	91.8 ± 0.3	92.7 ± 0.4	93.3 ± 0.4	93.8 ± 0.5
top-4 SVM^{α}	72.8 ± 1.0	81.4 ± 0.7	85.4 ± 0.9	87.9 ± 0.6	89.6 ± 0.5	90.9 ± 0.4	91.8 ± 0.3	92.7 ± 0.4	93.3 ± 0.4	93.8 ± 0.4
top-5 SVM^{α}	72.6 ± 1.0	81.2 ± 0.7	85.2 ± 0.8	87.7 ± 0.6	89.5 ± 0.6	90.8 ± 0.5	91.7 ± 0.4	92.6 ± 0.4	93.2 ± 0.5	93.7 ± 0.5
top-10 SVM^{α}	71.0 ± 0.8	80.2 ± 1.0	84.2 ± 0.7	87.0 ± 0.8	88.8 ± 0.7	90.1 ± 0.5	91.3 ± 0.4	92.2 ± 0.5	92.9 ± 0.4	93.5 ± 0.4
top-1 SVM^{α}_{1}	73.2 ± 1.0	81.6 ± 0.8	85.5 ± 0.8	87.8 ± 0.7	89.5 ± 0.6	90.8 ± 0.5	91.8 ± 0.4	92.7 ± 0.4	93.3 ± 0.4	93.8 ± 0.4
top-2 SVM^{α}_{1}	73.2 ± 0.9	81.7 ± 0.7	85.5 ± 0.9	87.8 ± 0.7	89.6 ± 0.6	90.9 ± 0.5	91.9 ± 0.4	92.7 ± 0.3	93.3 ± 0.4	93.9 ± 0.5
top-3 SVM^{α}_{1}	73.0 ± 1.0	81.6 ± 0.7	85.4 ± 0.8	87.8 ± 0.6	89.6 ± 0.6	90.9 ± 0.4	91.8 ± 0.3	92.7 ± 0.4	93.3 ± 0.4	93.8 ± 0.4
top-4 SVM^{α}_{1}	72.8 ± 1.0	81.4 ± 0.7	85.4 ± 0.9	87.9 ± 0.6	89.6 ± 0.5	90.9 ± 0.4	91.8 ± 0.3	92.7 ± 0.4	93.3 ± 0.5	93.8 ± 0.4
top-5 SVM^{α}_{1}	72.5 ± 1.0	81.2 ± 0.7	85.2 ± 0.8	87.7 ± 0.6	89.5 ± 0.6	90.8 ± 0.4	91.7 ± 0.4	92.6 ± 0.4	93.2 ± 0.5	93.8 ± 0.4
top-10 SVM^{α}_{1}	71.0 ± 0.8	80.2 ± 1.0	84.2 ± 0.7	87.0 ± 0.8	88.8 ± 0.7	90.1 ± 0.5	91.3 ± 0.4	92.2 ± 0.5	92.9 ± 0.4	93.5 ± 0.4
top-1 SVM^{β} / $\text{SVM}^{\text{Multi}}$	73.2 ± 1.0	81.6 ± 0.8	85.5 ± 0.8	87.8 ± 0.7	89.5 ± 0.6	90.8 ± 0.5	91.8 ± 0.4	92.7 ± 0.4	93.3 ± 0.4	93.8 ± 0.4
top-2 SVM^{β}	73.2 ± 1.0	81.6 ± 0.8	85.5 ± 0.8	87.8 ± 0.7	89.5 ± 0.6	90.8 ± 0.5	91.8 ± 0.4	92.7 ± 0.4	93.3 ± 0.4	93.8 ± 0.4
top-3 SVM^{β}	73.2 ± 1.0	81.6 ± 0.8	85.5 ± 0.8	87.8 ± 0.7	89.5 ± 0.6	90.8 ± 0.5	91.8 ± 0.4	92.7 ± 0.4	93.3 ± 0.4	93.8 ± 0.4
top-4 SVM^{β}	73.2 ± 1.0	81.6 ± 0.8	85.5 ± 0.8	87.8 ± 0.7	89.5 ± 0.6	90.8 ± 0.5	91.8 ± 0.4	92.7 ± 0.4	93.3 ± 0.4	93.8 ± 0.4
top-5 SVM^{β}	73.2 ± 1.0	81.6 ± 0.8	85.5 ± 0.8	87.8 ± 0.7	89.5 ± 0.6	90.8 ± 0.5	91.8 ± 0.4	92.7 ± 0.4	93.3 ± 0.4	93.8 ± 0.4
top-10 SVM^{β}	73.2 ± 1.0	81.6 ± 0.8	85.5 ± 0.8	87.8 ± 0.7	89.5 ± 0.6	90.9 ± 0.5	91.8 ± 0.4	92.7 ± 0.4	93.3 ± 0.4	93.8 ± 0.5
top-1 SVM^{β}_{1}	73.2 ± 1.0	81.6 ± 0.8	85.5 ± 0.8	87.8 ± 0.7	89.5 ± 0.6	90.8 ± 0.5	91.8 ± 0.4	92.7 ± 0.4	93.3 ± 0.4	93.8 ± 0.4
top-2 SVM^{β}_{1}	73.2 ± 1.0	81.6 ± 0.8	85.5 ± 0.8	87.8 ± 0.7	89.5 ± 0.6	90.8 ± 0.5	91.8 ± 0.4	92.7 ± 0.4	93.3 ± 0.4	93.8 ± 0.4
top-3 SVM^{β}_{1}	73.2 ± 1.0	81.6 ± 0.8	85.5 ± 0.8	87.8 ± 0.7	89.5 ± 0.6	90.8 ± 0.5	91.8 ± 0.4	92.7 ± 0.4	93.3 ± 0.4	93.8 ± 0.4
top-4 SVM^{β}_{1}	73.2 ± 1.0	81.6 ± 0.8	85.5 ± 0.8	87.8 ± 0.7	89.5 ± 0.6	90.8 ± 0.5	91.8 ± 0.4	92.7 ± 0.4	93.3 ± 0.4	93.8 ± 0.4
top-5 SVM^{β}_{1}	73.2 ± 1.0	81.6 ± 0.8	85.5 ± 0.8	87.8 ± 0.7	89.5 ± 0.6	90.8 ± 0.5	91.8 ± 0.4	92.7 ± 0.4	93.3 ± 0.4	93.8 ± 0.4
top-10 SVM^{β}_{1}	73.2 ± 1.0	81.6 ± 0.8	85.5 ± 0.8	87.8 ± 0.7	89.5 ± 0.6	90.8 ± 0.5	91.8 ± 0.4	92.7 ± 0.4	93.3 ± 0.4	93.8 ± 0.4
top-1 Ent / LR^{Multi}	72.7 ± 0.8	80.9 ± 0.9	84.9 ± 0.9	87.4 ± 0.7	89.1 ± 0.7	90.5 ± 0.6	91.6 ± 0.5	92.4 ± 0.3	93.0 ± 0.5	93.6 ± 0.5
top-2 Ent	72.6 ± 0.9	80.9 ± 0.8	85.0 ± 0.9	87.5 ± 0.8	89.2 ± 0.5	90.6 ± 0.5	91.6 ± 0.5	92.4 ± 0.4	93.1 ± 0.4	93.6 ± 0.4
top-3 Ent	72.5 ± 0.9	80.8 ± 0.8	85.0 ± 0.9	87.3 ± 0.7	89.2 ± 0.6	90.5 ± 0.5	91.6 ± 0.4	92.4 ± 0.4	93.1 ± 0.4	93.6 ± 0.5
top-4 Ent	72.2 ± 1.0	80.7 ± 0.8	84.8 ± 0.9	87.3 ± 0.7	89.0 ± 0.7	90.5 ± 0.5	91.5 ± 0.5	92.4 ± 0.3	93.1 ± 0.5	93.6 ± 0.5
top-5 Ent	72.0 ± 0.8	80.5 ± 0.9	84.7 ± 0.8	87.2 ± 0.7	89.0 ± 0.6	90.4 ± 0.5	91.5 ± 0.3	92.3 ± 0.2	93.0 ± 0.4	93.6 ± 0.4
top-10 Ent	70.2 ± 0.9	79.8 ± 1.1	83.5 ± 0.6	86.6 ± 0.5	88.4 ± 0.6	89.7 ± 0.8	90.8 ± 0.7	91.9 ± 0.4	92.7 ± 0.4	93.0 ± 0.5

Table 10: Comparison of different methods in top-k accuracy (%).

Caltech 256 (average of 39 kernels from [19])

Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
top-1 SVM^{α} / $\mathrm{SVM}^{\mathrm{Multi}}$	47.1	56.3	61.4	64.7	67.3	69.3	71.5	72.6	74.1	75.2
top-2 SVM^{α}	47.1	56.3	61.4	64.6	67.2	69.4	71.3	72.6	74.1	75.2
top-3 SVM^{α}	47.0	56.3	61.5	64.6	67.3	69.4	71.3	72.6	74.1	75.1
top-4 SVM^{α}	46.9	56.2	61.4	64.5	67.3	69.3	71.3	72.8	74.1	75.2
top-5 SVM^{α}	46.8	56.1	61.2	64.4	67.1	69.3	71.2	72.8	74.2	75.1
top-10 SVM^{α}	45.4	55.3	60.7	64.1	66.7	69.0	71.0	72.6	73.9	75.2
top-1 SVM_1^{α}	47.1	56.3	61.4	64.5	67.1	69.3	71.5	72.8	74.1	75.2
top-2 SVM_1^{α}	47.1	56.3	61.4	64.4	67.0	69.3	71.3	72.7	74.0	75.2
top-3 SVM_1^{α}	47.0	56.2	61.5	64.5	67.3	69.4	71.3	72.7	74.1	75.1
top-4 SVM_1^{α}	46.9	56.2	61.4	64.5	67.3	69.3	71.1	72.8	74.2	75.2
top-5 SVM_1^{α}	46.8	56.1	61.3	64.3	67.0	69.4	71.1	72.9	74.2	75.2
top-10 SVM_1^{α}	45.4	55.3	60.8	64.1	66.6	68.9	71.0	72.6	74.0	75.2
top-1 SVM^{β} / $\mathrm{SVM}^{\mathrm{Multi}}$	47.1	56.3	61.4	64.7	67.3	69.3	71.5	72.6	74.1	75.2
top-2 SVM^{β}	47.2	56.3	61.4	64.7	67.2	69.4	71.4	72.6	74.1	75.2
top-3 SVM^{β}	47.2	56.3	61.4	64.7	67.2	69.4	71.4	72.6	74.1	75.2
top-4 SVM^{β}	47.2	56.3	61.4	64.7	67.2	69.4	71.4	72.6	74.1	75.2
top-5 SVM^{β}	47.2	56.3	61.4	64.7	67.2	69.3	71.4	72.6	74.1	75.2
top-10 SVM^{β}	47.2	56.3	61.4	64.7	67.2	69.3	71.5	72.7	74.1	75.2
top-1 SVM_1^{β}	47.1	56.3	61.4	64.5	67.1	69.3	71.5	72.8	74.1	75.2
top-2 SVM_1^{β}	47.1	56.3	61.4	64.5	67.1	69.3	71.5	72.8	74.1	75.2
top-3 SVM_1^{β}	47.1	56.3	61.4	64.5	67.1	69.3	71.5	72.7	74.1	75.2
top-4 SVM_1^{β}	47.1	56.3	61.4	64.5	67.1	69.3	71.5	72.7	74.1	75.2
top-5 SVM_1^{β}	47.1	56.3	61.4	64.5	67.1	69.3	71.5	72.7	74.1	75.2
top-10 SVM_1^{β}	47.2	56.3	61.4	64.5	67.1	69.3	71.5	72.7	74.1	75.2
top-1 Ent / $\mathrm{LR}^{\mathrm{Multi}}$	46.2	55.6	60.9	64.1	66.7	69.1	71.0	72.5	73.7	74.9
top-2 Ent	46.2	55.7	60.9	64.1	66.7	69.0	71.0	72.0	73.3	74.5
top-3 Ent	46.1	55.8	61.0	64.2	66.8	69.0	70.9	72.5	73.7	74.8
top-4 Ent	46.1	55.5	60.9	64.2	66.7	68.9	70.9	71.5	73.9	74.8
top-5 Ent	46.0	55.4	60.7	64.1	66.7	68.8	70.7	72.7	73.2	74.2
top-10 Ent	45.0	54.6	59.9	63.7	66.2	68.4	70.4	72.3	73.5	74.7

Table 11: Comparison of different methods in top-k accuracy (%).

CUB

Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
$\mathrm{SVM}^{\mathrm{OVA}}$	60.6	71.4	77.1	80.7	83.4	85.2	86.6	87.9	89.0	89.9
$\mathrm{LR}^{\mathrm{OVA}}$	62.3	74.8	80.5	84.6	87.4	89.4	90.7	91.9	92.8	93.5
top-1 SVM^{α} / $\mathrm{SVM}^{\mathrm{Multi}}$	61.0	73.3	79.2	82.8	85.7	87.8	89.5	90.7	91.6	92.3
top-2 SVM^{α}	61.2	73.7	79.9	83.5	85.9	88.2	89.9	91.0	91.9	92.6
top-3 SVM^{α}	61.3	74.9	80.4	83.9	86.3	88.1	89.9	91.3	91.9	92.5
top-4 SVM^{α}	61.1	75.1	81.1	84.6	86.7	88.7	89.8	90.9	91.9	92.8
top-5 SVM^{α}	60.9	74.7	81.2	84.7	87.2	89.0	90.4	91.0	92.1	92.9
top-10 SVM^{α}	59.6	73.9	81.3	85.1	87.7	89.6	90.7	91.7	92.7	93.4
top-1 SVM_1^{α}	61.9	74.3	80.2	84.0	86.9	88.6	90.1	91.4	92.3	93.1
top-2 SVM_1^{α}	62.0	74.7	80.5	84.0	86.9	88.6	90.1	91.3	92.2	93.0
top-3 SVM_1^{α}	61.9	75.1	81.1	84.2	86.6	88.6	90.2	91.4	92.2	93.2
top-4 SVM_1^{α}	61.7	75.1	81.2	84.7	87.1	89.0	90.3	91.4	92.3	93.0
top-5 SVM_1^{α}	61.3	75.0	81.3	85.0	87.4	89.2	90.6	91.2	92.2	92.9
top-10 SVM_1^{α}	59.8	73.9	81.4	85.2	87.8	89.7	90.7	91.8	92.7	93.4
top-1 SVM^{β} / $\mathrm{SVM}^{\mathrm{Multi}}$	61.0	73.3	79.2	82.8	85.7	87.8	89.5	90.7	91.6	92.3
top-2 SVM^{β}	61.0	73.4	79.2	83.1	86.0	88.1	89.8	91.1	91.9	92.6
top-3 SVM^{β}	61.3	73.7	79.8	83.5	86.4	88.2	89.9	91.2	91.9	92.5
top-4 SVM^{β}	61.5	73.9	79.7	83.8	86.6	88.3	89.8	90.9	91.9	92.8
top-5 SVM^{β}	61.8	74.4	79.8	83.8	86.4	88.5	90.1	91.1	91.6	92.4
top-10 SVM^{β}	62.1	74.9	80.8	84.4	86.8	88.8	90.1	91.3	92.2	93.1
top-1 SVM_1^{β}	61.9	74.3	80.2	84.0	86.9	88.6	90.1	91.4	92.3	93.1
top-2 SVM_1^{β}	61.9	74.3	80.2	84.0	86.9	88.6	90.2	91.3	92.2	93.1
top-3 SVM_1^{β}	62.0	74.3	80.2	83.9	86.7	88.6	90.2	91.4	92.2	93.2
top-4 SVM_1^{β}	61.9	74.4	80.2	83.9	86.8	88.5	90.2	91.3	92.3	93.1
top-5 SVM_1^{β}	61.9	74.4	80.4	83.9	86.7	88.7	90.1	91.5	92.4	93.0
top-10 SVM_1^{β}	62.4	74.9	80.9	84.3	86.8	88.8	90.3	91.3	92.1	92.9
top-1 Ent / $\mathrm{LR}^{\mathrm{Multi}}$	62.3	75.2	81.7	85.2	87.9	89.6	91.3	92.5	93.2	93.9
top-2 Ent	62.4	75.4	81.6	85.2	87.9	89.7	91.3	92.5	93.2	93.9
top-3 Ent	62.5	75.6	81.8	85.4	87.9	89.6	91.2	92.4	93.2	93.9
top-4 Ent	62.3	75.6	81.4	85.4	87.8	89.6	91.1	92.4	93.2	93.9
top-5 Ent	62.0	75.4	81.9	85.4	88.1	90.0	91.2	92.4	93.2	93.8
top-10 Ent	61.2	74.7	81.6	85.3	88.2	90.0	91.3	92.4	93.3	93.8
top-2 $\mathrm{Ent}_{\mathrm{tr}}$	61.9	74.8	81.2	84.7	87.6	89.4	90.9	91.9	92.7	93.4
top-3 $\mathrm{Ent}_{\mathrm{tr}}$	62.0	75.0	81.4	84.9	87.6	89.4	90.8	91.9	93.0	93.4
top-4 $\mathrm{Ent}_{\mathrm{tr}}$	61.8	74.5	81.1	84.9	87.7	89.6	91.0	92.0	93.0	93.6
top-5 $\mathrm{Ent}_{\mathrm{tr}}$	61.4	74.8	81.2	85.0	87.7	89.7	91.1	92.0	92.9	93.7
top-10 $\mathrm{Ent}_{\mathrm{tr}}$	59.7	73.2	80.7	84.7	87.2	89.4	90.8	91.9	92.8	93.4

Table 12: Comparison of different methods in top-k accuracy (%).

Flowers

Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
$\mathrm{SVM}^{\mathrm{OVA}}$	82.0	89.2	91.7	93.2	94.3	95.2	95.9	95.8	96.2	96.8
$\mathrm{LR}^{\mathrm{OVA}}$	82.6	89.4	92.2	93.9	94.8	95.8	96.4	96.9	97.3	97.6
top-1 SVM^{α} / $\mathrm{SVM}^{\mathrm{Multi}}$	82.5	89.5	92.2	93.8	94.8	95.6	96.2	95.5	96.0	96.4
top-2 SVM^{α}	82.3	89.5	92.3	93.9	95.0	95.7	96.5	95.6	95.9	96.2
top-3 SVM^{α}	81.9	89.3	92.2	93.8	95.0	95.8	96.4	97.0	96.0	96.1
top-4 SVM^{α}	81.8	89.3	92.3	94.0	95.0	95.9	96.6	97.0	95.6	97.8
top-5 SVM^{α}	81.7	89.1	92.4	94.1	95.1	95.8	96.6	97.1	97.4	97.8
top-10 SVM^{α}	80.5	88.8	91.9	93.7	95.1	95.9	96.5	97.1	97.4	97.7
top-1 SVM_1^{α}	83.0	89.8	92.4	94.0	95.1	95.9	96.4	96.7	97.3	97.6
top-2 SVM_1^{α}	82.6	89.6	92.4	94.0	95.2	95.9	96.5	96.9	97.3	97.6
top-3 SVM_1^{α}	82.5	89.7	92.3	94.1	95.2	95.8	96.5	97.1	97.3	97.7
top-4 SVM_1^{α}	82.3	89.3	92.4	94.1	95.2	96.0	96.6	97.1	97.5	97.7
top-5 SVM_1^{α}	82.0	89.3	92.5	94.1	95.1	95.9	96.6	97.1	97.5	97.8
top-10 SVM_1^{α}	80.6	88.8	91.9	93.7	95.1	95.9	96.6	97.0	97.4	97.7
top-1 SVM^{β} / $\mathrm{SVM}^{\mathrm{Multi}}$	82.5	89.5	92.2	93.8	94.8	95.6	96.2	95.5	96.0	96.4
top-2 SVM^{β}	82.5	89.6	92.2	93.7	94.9	95.6	96.2	95.6	95.9	96.3
top-3 SVM^{β}	82.4	89.8	92.1	93.7	94.8	95.7	96.2	96.8	95.8	96.1
top-4 SVM^{β}	82.4	89.5	92.1	93.7	94.8	95.6	96.2	96.7	95.6	96.0
top-5 SVM^{β}	82.5	89.7	92.0	93.7	94.9	95.6	96.2	96.7	95.8	96.2
top-10 SVM^{β}	82.7	89.7	92.3	93.9	95.0	95.6	96.2	96.7	97.2	97.5
top-1 SVM_1^{β}	83.0	89.8	92.4	94.0	95.1	95.9	96.4	96.7	97.3	97.6
top-2 SVM_1^{β}	83.0	89.8	92.4	94.0	95.1	95.9	96.4	96.7	97.3	97.6
top-3 SVM_1^{β}	83.0	89.8	92.4	94.0	95.1	95.9	96.5	96.7	97.4	97.6
top-4 SVM_1^{β}	82.9	89.7	92.4	94.0	95.0	96.0	96.5	96.9	97.3	97.6
top-5 SVM_1^{β}	83.0	89.9	92.4	93.9	95.1	95.9	96.5	96.7	97.4	97.6
top-10 SVM_1^{β}	82.7	89.8	92.4	94.0	95.1	96.0	96.2	96.7	97.3	97.6
top-1 Ent / $\mathrm{LR}^{\mathrm{Multi}}$	82.9	89.7	92.4	94.0	95.1	96.0	96.6	97.1	97.4	97.8
top-2 Ent	82.6	89.7	92.4	94.0	95.3	96.0	96.6	97.1	97.5	97.8
top-3 Ent	82.5	89.5	92.0	94.1	95.3	96.1	96.6	97.1	97.4	97.8
top-4 Ent	82.2	89.5	92.4	94.1	95.3	96.0	96.7	97.1	97.5	97.8
top-5 Ent	82.1	89.4	92.2	94.1	95.1	95.9	96.6	97.1	97.5	97.9
top-10 Ent	80.9	88.9	92.1	93.9	95.0	95.9	96.5	97.0	97.4	97.7
top-2 $\mathrm{Ent}_{\mathrm{tr}}$	82.1	89.4	92.3	93.8	95.0	96.0	96.5	97.2	97.4	97.8
top-3 $\mathrm{Ent}_{\mathrm{tr}}$	82.1	89.2	92.2	94.2	95.2	96.0	96.7	97.0	97.4	97.6
top-4 $\mathrm{Ent}_{\mathrm{tr}}$	81.9	88.8	92.3	94.1	95.0	95.9	96.6	97.2	97.5	97.7
top-5 $\mathrm{Ent}_{\mathrm{tr}}$	81.4	89.0	92.0	93.8	95.0	95.7	96.2	97.0	97.3	97.7
top-10 $\mathrm{Ent}_{\mathrm{tr}}$	77.9	87.8	91.1	93.0	94.3	95.2	96.0	96.6	96.9	97.3

Table 13: Comparison of different methods in top-k accuracy (%).

FMD

Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
$\mathrm{SVM}^{\mathrm{OVA}}$	77.4	87.2	92.4	94.4	96.4	97.8	99.0	99.8	100.0	100.0
$\mathrm{LR}^{\mathrm{OVA}}$	79.6	90.2	94.2	96.6	98.2	98.8	99.2	100.0	99.8	100.0
top-1 SVM^{α} / $\mathrm{SVM}^{\mathrm{Multi}}$	77.6	88.8	93.8	95.4	97.2	98.4	100.0	100.0	100.0	100.0
top-2 SVM^{α}	78.2	89.2	94.0	95.8	97.4	98.6	99.8	100.0	100.0	100.0
top-3 SVM^{α}	78.8	89.2	94.6	96.4	97.8	98.8	99.4	99.8	100.0	100.0
top-4 SVM^{α}	78.2	89.4	94.6	96.8	98.0	98.8	99.4	99.6	100.0	100.0
top-5 SVM^{α}	78.4	89.2	94.4	96.8	97.6	98.6	99.2	99.6	99.8	100.0
top-1 SVM_1^{α}	78.6	89.4	93.8	96.0	98.0	99.0	99.4	99.6	100.0	100.0
top-2 SVM_1^{α}	78.4	90.2	93.8	96.2	97.6	99.0	99.2	99.6	99.8	100.0
top-3 SVM_1^{α}	79.0	89.6	94.4	96.2	98.0	99.0	99.2	99.6	99.8	100.0
top-4 SVM_1^{α}	79.2	89.4	94.6	96.6	97.8	98.8	99.2	99.6	99.8	100.0
top-5 SVM_1^{α}	79.4	89.2	94.4	96.8	97.6	98.8	99.2	99.2	99.8	100.0
top-1 SVM^{β} / $\mathrm{SVM}^{\mathrm{Multi}}$	77.6	88.8	93.8	95.4	97.2	98.4	100.0	100.0	100.0	100.0
top-2 SVM^{β}	79.0	89.6	93.6	95.4	97.4	98.6	99.8	100.0	100.0	100.0
top-3 SVM^{β}	79.8	89.6	93.6	95.4	97.8	98.8	99.2	99.8	100.0	100.0
top-4 SVM^{β}	80.4	90.0	93.8	95.6	97.8	98.8	99.2	99.6	100.0	100.0
top-5 SVM^{β}	80.2	90.2	94.8	95.8	97.4	98.8	99.6	100.0	100.0	100.0
top-1 SVM_1^{β}	78.6	89.4	93.8	96.0	98.0	99.0	99.4	99.6	100.0	100.0
top-2 SVM_1^{β}	78.6	89.4	93.8	96.0	98.0	99.0	99.4	99.6	100.0	100.0
top-3 SVM_1^{β}	79.6	89.8	94.0	96.2	98.2	99.0	99.2	99.6	100.0	100.0
top-4 SVM_1^{β}	79.4	90.2	94.2	96.2	97.8	98.8	99.4	99.6	100.0	100.0
top-5 SVM_1^{β}	80.0	90.2	94.6	96.0	97.6	98.8	99.4	100.0	100.0	100.0
top-1 Ent / $\mathrm{LR}^{\mathrm{Multi}}$	79.0	90.6	94.6	96.6	97.8	98.8	99.2	100.0	99.8	100.0
top-2 Ent	79.4	89.6	94.6	97.6	98.0	98.8	99.2	100.0	99.8	100.0
top-3 Ent	79.8	89.4	94.8	97.4	98.0	98.8	99.2	99.2	99.8	100.0
top-4 Ent	79.2	90.2	94.8	97.0	97.8	98.8	99.0	99.2	99.8	100.0
top-5 Ent	79.4	90.4	94.4	97.2	98.0	98.8	99.2	99.2	99.8	100.0
top-2 $\mathrm{Ent}_{\mathrm{tr}}$	79.0	89.4	94.4	96.6	98.2	99.0	99.2	100.0	100.0	100.0
top-3 $\mathrm{Ent}_{\mathrm{tr}}$	78.4	89.6	95.4	97.4	98.2	98.6	99.2	100.0	100.0	100.0
top-4 $\mathrm{Ent}_{\mathrm{tr}}$	77.8	89.4	94.8	96.6	98.0	98.8	99.4	99.8	100.0	100.0
top-5 $\mathrm{Ent}_{\mathrm{tr}}$	77.2	89.4	94.0	96.4	97.8	98.8	99.2	100.0	100.0	100.0

Table 14: Comparison of different methods in top-k accuracy (%).

Indoor 67 (AlexNet trained on Places 205, FC7 output, provided by [28])

Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
$\mathrm{SVM}^{\mathrm{OVA}}$	71.7	80.1	85.9	90.2	92.6	94.3	94.9	95.7	96.4	96.9
$\mathrm{LR}^{\mathrm{OVA}}$	73.1	84.5	89.6	91.8	93.3	94.3	95.3	96.1	96.6	97.0
top-1 SVM^{α} / $\mathrm{SVM}^{\mathrm{Multi}}$	74.0	85.2	89.3	91.9	93.4	94.9	95.6	95.8	96.4	96.9
top-2 SVM^{α}	73.1	85.7	90.4	92.2	94.5	95.1	96.2	96.6	97.0	97.3
top-3 SVM^{α}	71.6	86.3	91.1	93.2	94.7	95.7	96.4	96.6	97.1	97.2
top-4 SVM^{α}	71.4	85.7	90.7	93.3	94.8	95.7	96.2	96.6	97.2	97.8
top-5 SVM^{α}	70.7	85.7	90.4	93.2	94.7	95.5	96.1	96.9	97.5	97.9
top-10 SVM^{α}	70.0	85.4	90.0	93.1	94.6	95.6	96.2	97.1	97.5	97.5
top-1 SVM_1^{α}	74.0	86.0	90.7	92.8	94.5	95.9	96.1	96.8	97.1	97.4
top-2 SVM_1^{α}	72.7	85.9	90.5	93.0	94.3	95.8	96.3	96.6	97.1	97.4
top-3 SVM_1^{α}	72.2	86.1	90.8	93.1	94.6	95.4	96.3	96.6	97.3	97.7
top-4 SVM_1^{α}	71.3	86.2	90.7	93.1	94.6	95.5	96.4	96.7	97.4	97.7
top-5 SVM_1^{α}	71.0	86.2	90.2	92.8	94.7	95.4	96.3	97.0	97.5	97.8
top-10 SVM_1^{α}	70.3	85.2	89.9	93.1	94.6	95.7	96.0	97.0	97.4	97.7
top-1 SVM^{β} / $\mathrm{SVM}^{\mathrm{Multi}}$	74.0	85.2	89.3	91.9	93.4	94.9	95.6	95.8	96.4	96.9
top-2 SVM^{β}	74.0	85.9	89.8	92.2	94.1	95.1	95.7	96.4	97.0	97.3
top-3 SVM^{β}	73.0	86.3	90.6	92.8	94.4	95.9	96.1	96.3	96.9	97.2
top-4 SVM^{β}	73.1	86.2	90.8	92.7	94.5	95.8	96.2	96.6	97.1	97.7
top-5 SVM^{β}	72.6	85.6	90.7	93.0	94.5	95.7	96.2	96.8	97.4	97.6
top-10 SVM^{β}	71.9	85.3	90.4	93.4	94.4	95.6	96.2	97.2	97.5	97.8
top-1 SVM_1^{β}	74.0	86.0	90.7	92.8	94.5	95.9	96.1	96.8	97.1	97.4
top-2 SVM_1^{β}	73.9	85.9	90.8	92.8	94.5	95.7	96.2	96.6	97.1	97.5
top-3 SVM_1^{β}	73.7	86.3	90.7	92.7	94.5	95.6	96.3	96.6	97.3	97.5
top-4 SVM_1^{β}	73.0	86.2	90.7	92.5	94.3	95.7	96.3	96.7	97.4	97.5
top-5 SVM_1^{β}	72.7	85.9	90.8	92.7	94.4	95.6	96.2	97.0	97.5	97.8
top-10 SVM_1^{β}	72.2	85.5	90.7	93.4	94.6	95.6	96.3	97.1	97.5	97.8
top-1 Ent / $\mathrm{LR}^{\mathrm{Multi}}$	72.5	86.0	91.2	93.8	94.9	95.7	96.6	97.1	97.5	97.8
top-2 Ent	72.1	85.9	91.3	93.7	94.8	95.9	96.6	97.0	97.5	97.8
top-3 Ent	71.4	86.0	91.0	93.6	94.9	95.7	96.6	97.2	97.6	97.7
top-4 Ent	71.3	86.0	91.2	93.4	94.8	95.7	96.6	97.2	97.5	97.6
top-5 Ent	71.0	85.7	90.6	93.0	94.7	95.8	96.5	97.1	97.5	97.6
top-10 Ent	69.6	84.8	90.1	92.5	94.9	95.9	96.4	97.1	97.4	97.5
top-2 $\mathrm{Ent}_{\mathrm{tr}}$	68.2	85.3	90.4	92.8	94.6	95.6	96.5	97.1	97.6	97.8
top-3 $\mathrm{Ent}_{\mathrm{tr}}$	68.3	83.9	89.8	92.6	94.5	95.1	96.3	97.2	97.5	97.5
top-4 $\mathrm{Ent}_{\mathrm{tr}}$	68.2	82.4	89.3	92.2	94.2	95.4	96.2	97.0	97.4	97.5
top-5 $\mathrm{Ent}_{\mathrm{tr}}$	67.1	82.8	89.6	92.8	94.3	95.3	95.9	96.7	97.4	97.3
top-10 $\mathrm{Ent}_{\mathrm{tr}}$	66.4	82.6	87.9	91.9	93.8	95.3	95.9	96.8	97.0	97.3

Table 15: Comparison of different methods in top-k accuracy (%).

Indoor 67 (VGG-16 trained on Places 205, FC6 output)

Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
$\mathrm{SVM}^{\mathrm{OVA}}$	80.4	89.9	94.3	95.3	96.0	96.9	97.3	97.7	97.9	98.3
$\mathrm{LR}^{\mathrm{OVA}}$	82.5	92.6	95.1	96.3	97.5	97.5	98.3	98.6	98.4	98.7
top-1 SVM^{α} / $\mathrm{SVM}^{\mathrm{Multi}}$	82.6	91.4	95.0	96.1	96.0	97.9	97.5	97.8	98.3	98.9
top-2 SVM^{α}	83.1	92.3	95.4	96.8	97.7	98.0	98.3	98.7	98.7	98.5
top-3 SVM^{α}	83.0	92.5	95.4	97.5	97.8	98.3	98.5	98.7	98.7	98.9
top-4 SVM^{α}	82.9	92.5	95.6	97.0	97.8	98.2	98.7	98.7	98.8	99.0
top-5 SVM^{α}	82.5	92.3	95.7	97.0	97.7	98.1	98.5	98.7	98.8	99.0
top-10 SVM^{α}	80.4	92.2	95.8	96.9	97.7	98.4	98.8	99.0	99.1	99.2
top-1 SVM_1^{α}	83.1	92.5	95.1	97.1	97.8	98.1	98.5	98.3	98.4	99.0
top-2 SVM_1^{α}	83.0	93.1	95.2	97.1	97.8	98.1	98.4	98.8	98.4	99.0
top-3 SVM_1^{α}	82.9	93.0	95.4	97.2	97.8	98.2	98.5	98.8	98.8	98.9
top-4 SVM_1^{α}	82.5	92.5	95.7	97.2	97.8	98.2	98.7	98.7	98.8	99.0
top-5 SVM_1^{α}	82.1	92.5	95.7	97.0	97.7	98.3	98.7	98.7	99.0	99.1
top-10 SVM_1^{α}	80.1	92.1	95.9	96.9	97.5	98.1	98.7	99.0	99.2	99.3
top-1 SVM^{β} / $\mathrm{SVM}^{\mathrm{Multi}}$	82.6	91.4	95.0	96.1	96.0	97.9	97.5	97.8	98.3	98.9
top-2 SVM^{β}	83.2	92.1	95.2	96.6	97.5	98.0	98.4	98.7	98.7	98.5
top-3 SVM^{β}	82.9	92.5	95.1	96.9	97.9	98.3	98.4	98.7	98.4	98.7
top-4 SVM^{β}	82.5	92.6	95.2	97.1	97.8	98.4	98.6	98.8	98.8	98.8
top-5 SVM^{β}	82.2	92.8	95.4	97.1	97.8	98.3	98.6	98.8	98.8	98.9
top-10 SVM^{β}	81.9	92.8	95.7	97.1	97.7	98.4	98.7	99.0	99.0	99.2
top-1 SVM_1^{β}	83.1	92.5	95.1	97.1	97.8	98.1	98.5	98.3	98.4	99.0
top-2 SVM_1^{β}	82.9	92.8	95.1	97.0	97.8	98.2	98.5	98.7	98.4	99.0
top-3 SVM_1^{β}	82.8	92.9	95.1	97.0	97.8	98.4	98.5	98.8	98.8	99.0
top-4 SVM_1^{β}	82.2	92.7	95.4	97.1	97.8	98.4	98.6	98.8	98.8	99.0
top-5 SVM_1^{β}	82.3	92.8	95.4	97.0	97.8	98.4	98.7	98.8	98.8	99.0
top-10 SVM_1^{β}	81.6	92.9	95.5	97.0	97.7	98.4	98.7	98.9	99.0	99.3
top-1 Ent / $\mathrm{LR}^{\mathrm{Multi}}$	82.4	92.9	95.4	97.1	97.2	98.1	98.7	99.1	99.1	99.1
top-2 Ent	82.4	93.2	95.7	97.2	97.2	98.1	98.7	99.0	99.0	99.2
top-3 Ent	81.9	93.3	95.7	97.1	97.5	98.1	98.8	98.9	99.0	99.1
top-4 Ent	81.9	92.8	95.7	96.9	97.6	98.2	98.7	98.9	99.0	99.2
top-5 Ent	81.9	92.5	95.8	96.9	97.5	98.4	98.7	99.0	99.0	99.2
top-10 Ent	81.2	92.5	95.7	96.9	97.6	98.4	98.7	99.0	99.1	99.2
top-2 $\mathrm{Ent}_{\mathrm{tr}}$	81.0	92.0	95.5	96.9	97.6	98.1	98.4	98.7	98.8	99.0
top-3 $\mathrm{Ent}_{\mathrm{tr}}$	81.2	92.0	95.7	96.9	97.6	98.4	98.6	98.7	99.0	99.1
top-4 $\mathrm{Ent}_{\mathrm{tr}}$	81.1	91.9	95.2	97.0	97.7	98.4	98.6	98.7	98.7	99.1
top-5 $\mathrm{Ent}_{\mathrm{tr}}$	80.9	91.9	95.1	96.5	97.5	98.1	98.6	98.9	99.1	99.0
top-10 $\mathrm{Ent}_{\mathrm{tr}}$	78.7	90.9	94.6	96.2	97.5	98.3	98.6	98.8	99.0	99.2

Table 16: Comparison of different methods in top-k accuracy (%).

Indoor 67 (VGG-16 trained on Places 205, FC7 output)

Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
$\mathrm{SVM}^{\mathrm{OVA}}$	81.9	91.0	94.3	95.7	96.5	97.2	97.4	97.7	97.8	98.0
$\mathrm{LR}^{\mathrm{OVA}}$	82.0	91.6	94.9	96.3	97.2	97.8	98.1	98.3	98.6	98.7
top-1 SVM^{α} / $\mathrm{SVM}^{\mathrm{Multi}}$	82.5	91.7	95.4	96.9	97.3	97.8	97.6	98.5	98.8	99.1
top-2 SVM^{α}	82.0	91.3	95.1	96.6	97.6	97.8	98.0	98.6	98.7	99.0
top-3 SVM^{α}	81.6	91.4	95.1	96.8	97.7	98.2	98.6	98.7	98.8	99.0
top-4 SVM^{α}	80.8	91.3	95.2	96.9	97.7	98.6	98.6	98.7	98.9	99.1
top-5 SVM^{α}	79.9	91.3	95.0	96.5	97.7	98.4	98.6	98.8	98.8	99.0
top-10 SVM^{α}	78.4	91.0	95.1	96.6	97.4	98.3	98.5	98.8	99.0	99.0
top-1 SVM_1^{α}	82.6	91.6	95.2	96.9	97.6	98.1	98.4	98.6	98.7	99.0
top-2 SVM_1^{α}	82.5	91.6	95.0	96.7	97.8	98.0	97.9	98.1	98.9	98.7
top-3 SVM_1^{α}	81.6	91.5	95.1	96.6	97.8	98.3	98.7	98.7	98.7	99.0
top-4 SVM_1^{α}	81.0	91.3	95.1	96.8	97.8	98.4	98.6	98.7	99.0	99.3
top-5 SVM_1^{α}	80.4	91.3	95.1	96.6	97.8	98.4	98.6	98.8	98.8	99.1
top-10 SVM_1^{α}	78.3	91.0	95.1	96.6	97.5	98.4	98.5	98.8	99.0	99.0
top-1 SVM^{β} / $\mathrm{SVM}^{\mathrm{Multi}}$	82.5	91.7	95.4	96.9	97.3	97.8	97.6	98.5	98.8	99.1
top-2 SVM^{β}	82.5	91.3	95.1	96.9	97.7	97.8	98.3	98.5	98.7	98.7
top-3 SVM^{β}	82.1	91.4	95.0	97.1	97.7	98.1	98.2	98.6	98.7	99.0
top-4 SVM^{β}	82.2	91.6	95.2	96.9	97.5	98.1	98.4	98.7	98.9	98.6
top-5 SVM^{β}	82.2	91.3	94.9	96.9	97.6	98.2	98.4	98.7	98.9	99.0
top-10 SVM^{β}	81.7	91.8	95.2	96.7	97.3	98.2	98.7	98.9	99.0	99.1
top-1 SVM_1^{β}	82.6	91.6	95.2	96.9	97.6	98.1	98.4	98.6	98.7	99.0
top-2 SVM_1^{β}	82.7	91.5	95.1	96.9	97.7	98.2	97.9	98.1	98.8	98.7
top-3 SVM_1^{β}	82.5	91.4	95.1	96.8	97.6	98.3	98.6	98.1	98.4	98.7
top-4 SVM_1^{β}	82.6	91.3	95.0	97.0	97.6	98.3	98.5	98.1	98.9	98.7
top-5 SVM_1^{β}	82.5	91.6	95.1	96.7	97.5	98.3	98.5	98.7	98.9	98.8
top-10 SVM_1^{β}	81.6	91.7	95.3	96.8	97.2	98.2	98.6	98.8	99.0	99.2
top-1 Ent / $\mathrm{LR}^{\mathrm{Multi}}$	82.3	91.4	95.2	97.2	98.0	98.4	98.7	98.8	99.0	99.1
top-2 Ent	81.9	91.9	95.1	96.9	97.8	98.4	98.7	98.7	99.0	99.3
top-3 Ent	81.4	91.5	95.4	96.7	97.6	98.3	98.7	98.8	99.0	99.2
top-4 Ent	80.8	91.6	95.4	96.6	97.7	98.3	98.7	98.8	99.0	99.1
top-5 Ent	80.3	91.3	95.0	96.6	97.7	98.2	98.7	98.8	99.0	99.0
top-10 Ent	79.2	91.0	95.1	96.7	97.6	98.3	98.7	98.7	99.0	99.0
top-2 $\mathrm{Ent}_{\mathrm{tr}}$	79.6	90.8	94.4	96.3	97.0	97.9	98.4	98.7	98.9	99.0
top-3 $\mathrm{Ent}_{\mathrm{tr}}$	79.8	90.9	95.0	96.2	97.5	98.1	98.5	98.7	99.0	99.1
top-4 $\mathrm{Ent}_{\mathrm{tr}}$	78.7	90.1	94.9	96.3	97.4	98.1	98.5	98.9	99.0	99.1
top-5 $\mathrm{Ent}_{\mathrm{tr}}$	76.4	90.2	94.3	96.0	97.3	97.8	98.4	98.6	98.7	99.0
top-10 $\mathrm{Ent}_{\mathrm{tr}}$	72.6	88.7	92.8	95.7	97.1	97.6	98.3	98.6	98.7	98.9

Table 17: Comparison of different methods in top-k accuracy (%).

SUN 397 (Fisher Vector kernels provided by [29])

Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
top-1 SVM^{α} / $\text{SVM}^{\text{Multi}}$	48.9 ± 0.3	60.6 ± 0.3	66.8 ± 0.2	70.8 ± 0.3	73.8 ± 0.2	76.1 ± 0.2	78.0 ± 0.2	79.5 ± 0.2	80.9 ± 0.2	82.0 ± 0.2
top-2 SVM^{α}	48.9 ± 0.3	60.5 ± 0.3	66.8 ± 0.2	70.8 ± 0.3	73.8 ± 0.2	76.1 ± 0.2	78.0 ± 0.2	79.5 ± 0.2	80.9 ± 0.2	82.0 ± 0.2
top-3 SVM^{α}	48.9 ± 0.3	60.5 ± 0.3	66.8 ± 0.2	70.8 ± 0.3	73.8 ± 0.2	76.1 ± 0.2	78.0 ± 0.3	79.5 ± 0.2	80.9 ± 0.2	82.0 ± 0.2
top-4 SVM^{α}	48.8 ± 0.3	60.5 ± 0.3	66.8 ± 0.2	70.8 ± 0.3	73.8 ± 0.2	76.1 ± 0.2	78.0 ± 0.2	79.6 ± 0.2	80.9 ± 0.2	82.0 ± 0.2
top-5 SVM^{α}	48.7 ± 0.3	60.5 ± 0.3	66.8 ± 0.2	70.9 ± 0.3	73.9 ± 0.3	76.2 ± 0.2	78.1 ± 0.2	79.6 ± 0.2	80.9 ± 0.2	82.1 ± 0.2
top-10 SVM^{α}	48.3 ± 0.4	60.4 ± 0.3	66.9 ± 0.2	71.1 ± 0.3	74.1 ± 0.2	76.5 ± 0.2	78.4 ± 0.2	80.0 ± 0.2	81.3 ± 0.2	82.5 ± 0.2
top-1 SVM_1^{α}	48.9 ± 0.3	60.6 ± 0.3	66.9 ± 0.2	71.1 ± 0.3	74.2 ± 0.2	76.5 ± 0.3	78.5 ± 0.2	80.1 ± 0.2	81.4 ± 0.2	82.6 ± 0.2
top-2 SVM_1^{α}	48.9 ± 0.3	60.6 ± 0.3	66.9 ± 0.2	71.1 ± 0.3	74.2 ± 0.2	76.5 ± 0.3	78.5 ± 0.2	80.1 ± 0.2	81.4 ± 0.2	82.6 ± 0.2
top-3 SVM_1^{α}	48.9 ± 0.3	60.6 ± 0.2	66.9 ± 0.2	71.1 ± 0.3	74.2 ± 0.2	76.5 ± 0.3	78.5 ± 0.2	80.1 ± 0.2	81.4 ± 0.2	82.6 ± 0.2
top-4 SVM_1^{α}	48.8 ± 0.3	60.6 ± 0.2	66.9 ± 0.2	71.1 ± 0.3	74.2 ± 0.2	76.5 ± 0.2	78.4 ± 0.2	80.1 ± 0.2	81.4 ± 0.2	82.6 ± 0.2
top-5 SVM_1^{α}	48.7 ± 0.3	60.6 ± 0.3	66.9 ± 0.2	71.1 ± 0.3	74.2 ± 0.2	76.5 ± 0.3	78.4 ± 0.2	80.1 ± 0.2	81.4 ± 0.2	82.6 ± 0.2
top-10 SVM_1^{α}	48.3 ± 0.4	60.4 ± 0.3	66.9 ± 0.2	71.1 ± 0.3	74.1 ± 0.2	76.5 ± 0.2	78.5 ± 0.2	80.0 ± 0.2	81.4 ± 0.2	82.6 ± 0.2
top-1 SVM^{β} / $\text{SVM}^{\text{Multi}}$	48.9 ± 0.3	60.6 ± 0.3	66.8 ± 0.2	70.8 ± 0.3	73.8 ± 0.2	76.1 ± 0.2	78.0 ± 0.2	79.5 ± 0.2	80.9 ± 0.2	82.0 ± 0.2
top-2 SVM^{β}	48.9 ± 0.3	60.6 ± 0.3	66.8 ± 0.2	70.8 ± 0.3	73.8 ± 0.2	76.1 ± 0.2	78.0 ± 0.2	79.5 ± 0.2	80.9 ± 0.2	82.0 ± 0.2
top-3 SVM^{β}	48.9 ± 0.3	60.6 ± 0.3	66.8 ± 0.2	70.8 ± 0.3	73.8 ± 0.2	76.1 ± 0.2	78.0 ± 0.2	79.5 ± 0.2	80.9 ± 0.2	82.0 ± 0.2
top-4 SVM^{β}	48.9 ± 0.3	60.6 ± 0.3	66.8 ± 0.2	70.8 ± 0.3	73.8 ± 0.2	76.1 ± 0.2	78.0 ± 0.2	79.5 ± 0.2	80.9 ± 0.2	82.0 ± 0.2
top-5 SVM^{β}	48.9 ± 0.3	60.6 ± 0.3	66.8 ± 0.2	70.8 ± 0.3	73.8 ± 0.2	76.1 ± 0.2	78.0 ± 0.2	79.5 ± 0.2	80.9 ± 0.2	82.0 ± 0.2
top-10 SVM^{β}	48.9 ± 0.3	60.6 ± 0.3	66.8 ± 0.2	70.8 ± 0.3	73.8 ± 0.2	76.3 ± 0.2	78.2 ± 0.2	79.8 ± 0.2	81.1 ± 0.2	82.3 ± 0.2
top-1 SVM_1^{β}	48.9 ± 0.3	60.6 ± 0.3	66.9 ± 0.2	71.1 ± 0.3	74.2 ± 0.2	76.5 ± 0.3	78.5 ± 0.2	80.1 ± 0.2	81.4 ± 0.2	82.6 ± 0.2
top-2 SVM_1^{β}	48.9 ± 0.3	60.6 ± 0.3	66.9 ± 0.2	71.1 ± 0.3	74.2 ± 0.2	76.5 ± 0.3	78.5 ± 0.2	80.1 ± 0.2	81.4 ± 0.2	82.6 ± 0.2
top-3 SVM_1^{β}	48.9 ± 0.3	60.6 ± 0.3	66.9 ± 0.2	71.1 ± 0.3	74.2 ± 0.2	76.5 ± 0.3	78.5 ± 0.2	80.1 ± 0.2	81.4 ± 0.2	82.6 ± 0.2
top-4 SVM_1^{β}	48.9 ± 0.3	60.6 ± 0.3	66.9 ± 0.2	71.1 ± 0.3	74.2 ± 0.2	76.5 ± 0.3	78.4 ± 0.2	80.1 ± 0.2	81.4 ± 0.2	82.6 ± 0.2
top-5 SVM_1^{β}	48.9 ± 0.3	60.6 ± 0.3	66.9 ± 0.2	71.1 ± 0.3	74.2 ± 0.2	76.5 ± 0.3	78.5 ± 0.2	80.1 ± 0.2	81.4 ± 0.2	82.6 ± 0.2
top-10 SVM_1^{β}	48.9 ± 0.3	60.6 ± 0.3	66.9 ± 0.2	71.0 ± 0.3	74.2 ± 0.2	76.5 ± 0.2	78.5 ± 0.2	80.1 ± 0.2	81.4 ± 0.2	82.6 ± 0.2
top-1 Ent / LR^{Multi}	48.5 ± 0.3	60.5 ± 0.2	66.9 ± 0.3	71.2 ± 0.3	74.3 ± 0.3	76.7 ± 0.3	78.6 ± 0.2	80.2 ± 0.2	81.6 ± 0.2	82.7 ± 0.2
top-2 Ent	48.5 ± 0.3	60.5 ± 0.2	66.9 ± 0.3	71.2 ± 0.3	74.4 ± 0.3	76.7 ± 0.3	78.6 ± 0.2	80.2 ± 0.2	81.6 ± 0.2	82.7 ± 0.2
top-3 Ent	48.5 ± 0.3	60.5 ± 0.2	66.9 ± 0.3	71.2 ± 0.3	74.4 ± 0.3	76.7 ± 0.3	78.6 ± 0.2	80.2 ± 0.2	81.6 ± 0.2	82.8 ± 0.2
top-4 Ent	48.5 ± 0.3	60.5 ± 0.2	66.9 ± 0.3	71.2 ± 0.3	74.3 ± 0.3	76.7 ± 0.3	78.6 ± 0.2	80.2 ± 0.2	81.6 ± 0.2	82.8 ± 0.2
top-5 Ent	48.4 ± 0.3	60.5 ± 0.2	66.9 ± 0.3	71.2 ± 0.3	74.3 ± 0.3	76.7 ± 0.3	78.6 ± 0.2	80.2 ± 0.3	81.6 ± 0.2	82.8 ± 0.2
top-10 Ent	48.0 ± 0.4	60.3 ± 0.3	66.8 ± 0.3	71.1 ± 0.3	74.3 ± 0.3	76.7 ± 0.2	78.7 ± 0.2	80.2 ± 0.2	81.6 ± 0.2	82.8 ± 0.2

Table 18: Comparison of different methods in top-k accuracy (%).

SUN 397 (AlexNet trained on Places 205, FC7 output, provided by [28])

Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
$\mathrm{SVM}^{\mathrm{OVA}}$	44.1 ± 0.4	60.8 ± 0.3	69.9 ± 0.2	75.8 ± 0.1	79.8 ± 0.1	82.8 ± 0.1	85.1 ± 0.2	86.9 ± 0.2	88.3 ± 0.2	89.6 ± 0.2
$\mathrm{LR}^{\mathrm{OVA}}$	53.9 ± 0.2	69.2 ± 0.2	76.5 ± 0.2	80.9 ± 0.2	84.0 ± 0.3	86.2 ± 0.2	87.9 ± 0.2	89.2 ± 0.2	90.3 ± 0.2	91.2 ± 0.2
top-1 SVM^{α} / $\mathrm{SVM}^{\mathrm{Multi}}$	58.2 ± 0.2	71.7 ± 0.2	78.2 ± 0.1	82.3 ± 0.2	85.0 ± 0.2	87.1 ± 0.2	88.8 ± 0.2	90.0 ± 0.2	91.0 ± 0.2	91.9 ± 0.2
top-2 SVM^{α}	58.8 ± 0.2	72.7 ± 0.2	79.3 ± 0.2	83.3 ± 0.2	85.9 ± 0.2	87.8 ± 0.2	89.2 ± 0.2	90.3 ± 0.2	91.3 ± 0.2	92.2 ± 0.2
top-3 SVM^{α}	59.0 ± 0.1	73.2 ± 0.2	79.9 ± 0.2	83.8 ± 0.2	86.5 ± 0.2	88.3 ± 0.2	89.7 ± 0.2	90.9 ± 0.2	91.8 ± 0.2	92.6 ± 0.2
top-4 SVM^{α}	58.9 ± 0.1	73.5 ± 0.2	80.3 ± 0.2	84.2 ± 0.2	86.8 ± 0.2	88.6 ± 0.2	90.0 ± 0.2	91.1 ± 0.2	92.0 ± 0.2	92.8 ± 0.2
top-5 SVM^{α}	58.9 ± 0.1	73.7 ± 0.2	80.5 ± 0.2	84.4 ± 0.3	87.0 ± 0.2	88.8 ± 0.2	90.2 ± 0.2	91.3 ± 0.2	92.2 ± 0.2	93.0 ± 0.2
top-10 SVM^{α}	58.0 ± 0.2	73.6 ± 0.1	80.8 ± 0.1	84.8 ± 0.2	87.4 ± 0.2	89.3 ± 0.2	90.7 ± 0.2	91.8 ± 0.2	92.7 ± 0.2	93.4 ± 0.2
top-1 SVM_1^{α}	59.7 ± 0.1	73.8 ± 0.1	80.3 ± 0.2	84.2 ± 0.2	86.8 ± 0.2	88.6 ± 0.2	90.0 ± 0.2	91.1 ± 0.2	92.0 ± 0.2	92.8 ± 0.2
top-2 SVM_1^{α}	59.6 ± 0.1	73.8 ± 0.1	80.4 ± 0.2	84.3 ± 0.2	86.8 ± 0.2	88.6 ± 0.3	90.1 ± 0.2	91.2 ± 0.2	92.1 ± 0.2	92.9 ± 0.2
top-3 SVM_1^{α}	59.4 ± 0.1	73.8 ± 0.1	80.4 ± 0.2	84.4 ± 0.2	86.9 ± 0.3	88.7 ± 0.3	90.1 ± 0.2	91.3 ± 0.2	92.2 ± 0.2	92.9 ± 0.2
top-4 SVM_1^{α}	59.2 ± 0.1	73.9 ± 0.1	80.6 ± 0.2	84.5 ± 0.2	87.1 ± 0.2	88.8 ± 0.2	90.2 ± 0.2	91.4 ± 0.2	92.3 ± 0.2	93.0 ± 0.2
top-5 SVM_1^{α}	59.1 ± 0.1	74.0 ± 0.2	80.7 ± 0.2	84.6 ± 0.3	87.2 ± 0.2	89.0 ± 0.2	90.4 ± 0.2	91.5 ± 0.2	92.4 ± 0.2	93.1 ± 0.2
top-10 SVM_1^{α}	58.1 ± 0.2	73.7 ± 0.2	80.9 ± 0.2	84.9 ± 0.2	87.5 ± 0.2	89.4 ± 0.2	90.7 ± 0.2	91.8 ± 0.2	92.7 ± 0.1	93.4 ± 0.1
top-1 SVM^{β} / $\mathrm{SVM}^{\mathrm{Multi}}$	58.2 ± 0.2	71.7 ± 0.2	78.2 ± 0.1	82.3 ± 0.2	85.0 ± 0.2	87.1 ± 0.2	88.8 ± 0.2	90.0 ± 0.2	91.0 ± 0.2	91.9 ± 0.2
top-2 SVM^{β}	58.8 ± 0.2	72.7 ± 0.2	79.3 ± 0.2	83.2 ± 0.2	85.9 ± 0.2	87.7 ± 0.2	89.0 ± 0.2	90.3 ± 0.2	91.3 ± 0.2	92.2 ± 0.2
top-3 SVM^{β}	59.1 ± 0.2	73.2 ± 0.2	79.8 ± 0.2	83.8 ± 0.2	86.4 ± 0.2	88.2 ± 0.2	89.6 ± 0.2	90.8 ± 0.2	91.7 ± 0.2	92.5 ± 0.2
top-4 SVM^{β}	59.2 ± 0.1	73.6 ± 0.2	80.2 ± 0.2	84.2 ± 0.2	86.7 ± 0.2	88.5 ± 0.2	89.9 ± 0.2	91.1 ± 0.2	92.0 ± 0.2	92.7 ± 0.2
top-5 SVM^{β}	59.3 ± 0.2	73.8 ± 0.2	80.4 ± 0.3	84.4 ± 0.3	87.0 ± 0.3	88.7 ± 0.3	90.1 ± 0.2	91.2 ± 0.2	92.1 ± 0.2	92.9 ± 0.2
top-10 SVM^{β}	59.3 ± 0.1	74.1 ± 0.2	80.9 ± 0.2	84.9 ± 0.2	87.5 ± 0.2	89.3 ± 0.2	90.7 ± 0.2	91.7 ± 0.2	92.6 ± 0.2	93.4 ± 0.2
top-1 SVM_1^{β}	59.7 ± 0.1	73.8 ± 0.1	80.3 ± 0.2	84.2 ± 0.2	86.8 ± 0.2	88.6 ± 0.2	90.0 ± 0.2	91.1 ± 0.2	92.0 ± 0.2	92.8 ± 0.2
top-2 SVM_1^{β}	59.7 ± 0.1	73.8 ± 0.2	80.3 ± 0.2	84.3 ± 0.2	86.8 ± 0.2	88.6 ± 0.3	90.0 ± 0.2	91.2 ± 0.2	92.1 ± 0.2	92.9 ± 0.2
top-3 SVM_1^{β}	59.6 ± 0.1	73.8 ± 0.1	80.4 ± 0.3	84.3 ± 0.2	86.9 ± 0.3	88.7 ± 0.3	90.1 ± 0.2	91.2 ± 0.2	92.1 ± 0.2	92.9 ± 0.2
top-4 SVM_1^{β}	59.6 ± 0.2	73.9 ± 0.2	80.5 ± 0.2	84.4 ± 0.2	87.0 ± 0.2	88.8 ± 0.3	90.2 ± 0.2	91.3 ± 0.2	92.2 ± 0.2	93.0 ± 0.2
top-5 SVM_1^{β}	59.5 ± 0.1	74.1 ± 0.1	80.7 ± 0.2	84.6 ± 0.2	87.1 ± 0.2	88.9 ± 0.3	90.3 ± 0.2	91.4 ± 0.2	92.3 ± 0.2	93.1 ± 0.2
top-10 SVM_1^{β}	59.3 ± 0.1	74.2 ± 0.2	81.0 ± 0.2	85.0 ± 0.2	87.5 ± 0.2	89.3 ± 0.2	90.7 ± 0.2	91.8 ± 0.2	92.7 ± 0.2	93.4 ± 0.2
top-1 Ent / $\mathrm{LR}^{\mathrm{Multi}}$	59.5 ± 0.2	74.2 ± 0.2	81.1 ± 0.2	85.1 ± 0.2	87.7 ± 0.2	89.6 ± 0.2	91.0 ± 0.1	92.1 ± 0.2	93.0 ± 0.2	93.7 ± 0.2
top-2 Ent	59.5 ± 0.2	74.3 ± 0.2	81.1 ± 0.2	85.1 ± 0.2	87.7 ± 0.2	89.6 ± 0.2	91.0 ± 0.2	92.1 ± 0.2	93.0 ± 0.2	93.7 ± 0.2
top-3 Ent	59.4 ± 0.1	74.3 ± 0.2	81.2 ± 0.2	85.2 ± 0.2	87.8 ± 0.2	89.6 ± 0.2	91.0 ± 0.2	92.1 ± 0.2	93.0 ± 0.2	93.7 ± 0.2
top-4 Ent	59.2 ± 0.2	74.3 ± 0.2	81.2 ± 0.2	85.2 ± 0.2	87.8 ± 0.2	89.7 ± 0.2	91.1 ± 0.2	92.2 ± 0.2	93.0 ± 0.2	93.7 ± 0.2
top-5 Ent	58.9 ± 0.1	74.3 ± 0.2	81.2 ± 0.2	85.2 ± 0.2	87.8 ± 0.2	89.7 ± 0.2	91.0 ± 0.2	92.1 ± 0.1	93.0 ± 0.2	93.7 ± 0.2
top-10 Ent	58.0 ± 0.2	73.7 ± 0.2	81.0 ± 0.2	85.1 ± 0.2	87.8 ± 0.2	89.7 ± 0.2	91.0 ± 0.2	92.2 ± 0.1	93.1 ± 0.2	93.8 ± 0.1

Table 19: Comparison of different methods in top-k accuracy (%).

SUN 397 (VGG-16 trained on Places 205, FC7 output)

Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
$\mathrm{SVM}^{\mathrm{OVA}}$	65.4 ± 0.2	77.6 ± 0.1	83.8 ± 0.2	87.2 ± 0.1	89.6 ± 0.1	91.3 ± 0.1	92.6 ± 0.2	93.6 ± 0.2	94.3 ± 0.2	94.9 ± 0.1
$\mathrm{LR}^{\mathrm{OVA}}$	67.6 ± 0.1	81.5 ± 0.2	87.2 ± 0.2	90.4 ± 0.2	92.4 ± 0.1	93.7 ± 0.1	94.7 ± 0.1	95.4 ± 0.1	96.0 ± 0.1	96.4 ± 0.1
top-1 SVM^{α} / $\mathrm{SVM}^{\mathrm{Multi}}$	65.8 ± 0.1	79.0 ± 0.2	85.1 ± 0.2	88.4 ± 0.2	90.8 ± 0.1	92.3 ± 0.1	93.3 ± 0.1	94.2 ± 0.1	94.8 ± 0.1	95.3 ± 0.1
top-2 SVM^{α}	66.4 ± 0.2	80.2 ± 0.2	86.1 ± 0.1	89.4 ± 0.1	91.5 ± 0.1	92.9 ± 0.2	93.8 ± 0.2	94.6 ± 0.2	95.3 ± 0.1	95.7 ± 0.1
top-3 SVM^{α}	66.5 ± 0.2	80.6 ± 0.2	86.5 ± 0.1	89.7 ± 0.2	91.8 ± 0.1	93.2 ± 0.1	94.2 ± 0.1	95.0 ± 0.1	95.3 ± 0.1	95.9 ± 0.1
top-4 SVM^{α}	66.4 ± 0.2	80.8 ± 0.2	86.8 ± 0.1	90.0 ± 0.2	92.1 ± 0.2	93.4 ± 0.1	94.4 ± 0.1	95.1 ± 0.1	95.7 ± 0.1	96.2 ± 0.1
top-5 SVM^{α}	66.3 ± 0.2	81.0 ± 0.2	87.0 ± 0.2	90.2 ± 0.1	92.2 ± 0.2	93.6 ± 0.1	94.5 ± 0.1	95.2 ± 0.1	95.8 ± 0.1	96.3 ± 0.1
top-10 SVM^{α}	64.8 ± 0.3	80.9 ± 0.1	87.2 ± 0.2	90.5 ± 0.2	92.6 ± 0.1	93.9 ± 0.1	94.9 ± 0.1	95.6 ± 0.1	96.2 ± 0.1	96.6 ± 0.1
top-1 SVM_1^{α}	67.4 ± 0.2	81.1 ± 0.2	86.8 ± 0.1	90.0 ± 0.1	92.0 ± 0.1	93.4 ± 0.1	94.3 ± 0.1	95.1 ± 0.1	95.7 ± 0.1	96.1 ± 0.1
top-2 SVM_1^{α}	67.2 ± 0.2	81.1 ± 0.2	86.9 ± 0.2	90.1 ± 0.2	92.1 ± 0.1	93.4 ± 0.1	94.4 ± 0.1	95.1 ± 0.1	95.7 ± 0.1	96.2 ± 0.1
top-3 SVM_1^{α}	67.0 ± 0.2	81.2 ± 0.2	87.0 ± 0.1	90.2 ± 0.2	92.2 ± 0.1	93.5 ± 0.1	94.5 ± 0.1	95.2 ± 0.1	95.7 ± 0.1	96.2 ± 0.0
top-4 SVM_1^{α}	66.8 ± 0.2	81.2 ± 0.2	87.1 ± 0.1	90.3 ± 0.2	92.3 ± 0.2	93.6 ± 0.1	94.5 ± 0.1	95.2 ± 0.1	95.8 ± 0.1	96.3 ± 0.0
top-5 SVM_1^{α}	66.5 ± 0.2	81.3 ± 0.2	87.2 ± 0.1	90.4 ± 0.2	92.4 ± 0.2	93.7 ± 0.1	94.6 ± 0.1	95.3 ± 0.1	95.9 ± 0.1	96.3 ± 0.0
top-10 SVM_1^{α}	64.9 ± 0.3	80.9 ± 0.1	87.3 ± 0.2	90.6 ± 0.2	92.6 ± 0.2	94.0 ± 0.1	94.9 ± 0.1	95.6 ± 0.1	96.2 ± 0.1	96.6 ± 0.1
top-1 SVM^{β} / $\mathrm{SVM}^{\mathrm{Multi}}$	65.8 ± 0.1	79.0 ± 0.2	85.1 ± 0.2	88.4 ± 0.2	90.8 ± 0.1	92.3 ± 0.1	93.3 ± 0.1	94.2 ± 0.1	94.8 ± 0.1	95.3 ± 0.1
top-2 SVM^{β}	66.4 ± 0.2	80.1 ± 0.1	86.0 ± 0.2	89.3 ± 0.2	91.4 ± 0.1	92.7 ± 0.2	93.8 ± 0.1	94.6 ± 0.2	95.3 ± 0.1	95.8 ± 0.1
top-3 SVM^{β}	66.8 ± 0.2	80.7 ± 0.2	86.5 ± 0.1	89.7 ± 0.2	91.7 ± 0.1	93.1 ± 0.2	94.2 ± 0.1	94.7 ± 0.1	95.3 ± 0.1	95.9 ± 0.1
top-4 SVM^{β}	66.9 ± 0.2	80.9 ± 0.2	86.8 ± 0.1	90.0 ± 0.2	92.0 ± 0.2	93.4 ± 0.1	94.4 ± 0.1	95.1 ± 0.1	95.7 ± 0.1	96.1 ± 0.1
top-5 SVM^{β}	67.0 ± 0.2	81.2 ± 0.2	87.0 ± 0.1	90.2 ± 0.2	92.2 ± 0.1	93.6 ± 0.1	94.5 ± 0.1	95.2 ± 0.1	95.8 ± 0.1	96.2 ± 0.1
top-10 SVM^{β}	66.9 ± 0.2	81.5 ± 0.2	87.4 ± 0.2	90.7 ± 0.1	92.6 ± 0.1	93.9 ± 0.1	94.8 ± 0.1	95.5 ± 0.1	96.1 ± 0.1	96.5 ± 0.0
top-1 SVM_1^{β}	67.4 ± 0.2	81.1 ± 0.2	86.8 ± 0.1	90.0 ± 0.1	92.0 ± 0.1	93.4 ± 0.1	94.3 ± 0.1	95.1 ± 0.1	95.7 ± 0.1	96.1 ± 0.1
top-2 SVM_1^{β}	67.3 ± 0.2	81.1 ± 0.2	86.8 ± 0.1	90.0 ± 0.2	92.0 ± 0.1	93.4 ± 0.1	94.4 ± 0.1	95.1 ± 0.1	95.7 ± 0.1	96.2 ± 0.1
top-3 SVM_1^{β}	67.3 ± 0.2	81.2 ± 0.2	86.9 ± 0.1	90.1 ± 0.1	92.1 ± 0.1	93.4 ± 0.1	94.4 ± 0.1	95.1 ± 0.1	95.7 ± 0.1	96.2 ± 0.1
top-4 SVM_1^{β}	67.2 ± 0.2	81.3 ± 0.2	87.0 ± 0.1	90.2 ± 0.2	92.2 ± 0.1	93.5 ± 0.1	94.5 ± 0.1	95.2 ± 0.1	95.8 ± 0.1	96.2 ± 0.1
top-5 SVM_1^{β}	67.2 ± 0.2	81.4 ± 0.2	87.2 ± 0.2	90.3 ± 0.1	92.3 ± 0.1	93.6 ± 0.1	94.6 ± 0.1	95.3 ± 0.1	95.9 ± 0.1	96.3 ± 0.1
top-10 SVM_1^{β}	66.9 ± 0.2	81.5 ± 0.2	87.5 ± 0.2	90.7 ± 0.1	92.6 ± 0.1	93.9 ± 0.1	94.9 ± 0.1	95.5 ± 0.1	96.1 ± 0.1	96.5 ± 0.1
top-1 Ent / $\mathrm{LR}^{\mathrm{Multi}}$	67.5 ± 0.1	81.7 ± 0.2	87.7 ± 0.2	90.9 ± 0.2	92.9 ± 0.1	94.2 ± 0.1	95.1 ± 0.1	95.8 ± 0.1	96.4 ± 0.1	96.8 ± 0.1
top-2 Ent	67.4 ± 0.2	81.8 ± 0.2	87.7 ± 0.2	90.9 ± 0.1	92.9 ± 0.1	94.2 ± 0.1	95.1 ± 0.1	95.8 ± 0.1	96.4 ± 0.1	96.8 ± 0.1
top-3 Ent	67.2 ± 0.2	81.8 ± 0.2	87.7 ± 0.2	90.9 ± 0.1	92.9 ± 0.1	94.2 ± 0.1	95.1 ± 0.1	95.8 ± 0.1	96.4 ± 0.1	96.8 ± 0.1
top-4 Ent	66.9 ± 0.2	81.7 ± 0.2	87.7 ± 0.2	91.0 ± 0.1	92.9 ± 0.1	94.2 ± 0.1	95.2 ± 0.1	95.9 ± 0.1	96.4 ± 0.1	96.8 ± 0.1
top-5 Ent	66.6 ± 0.3	81.6 ± 0.2	87.7 ± 0.2	91.0 ± 0.1	92.9 ± 0.1	94.2 ± 0.1	95.2 ± 0.1	95.9 ± 0.1	96.4 ± 0.1	96.8 ± 0.1
top-10 Ent	65.2 ± 0.3	81.0 ± 0.2	87.4 ± 0.1	90.8 ± 0.2	92.8 ± 0.1	94.2 ± 0.1	95.2 ± 0.1	95.9 ± 0.1	96.4 ± 0.1	96.8 ± 0.1

Table 20: Comparison of different methods in top-k accuracy (%).

Places 205 (AlexNet trained on Places 205, FC7 output, provided by [28])

Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
top-1 SVM^{α} / $\mathrm{SVM}^{\mathrm{Multi}}$	50.6	64.5	71.4	75.5	78.5	80.7	82.5	84.0	85.1	86.2
top-2 SVM^{α}	51.1	65.7	73.1	77.5	80.7	83.1	84.9	86.3	87.5	88.4
top-3 SVM^{α}	51.3	66.2	73.2	77.9	81.3	83.6	85.6	87.1	88.3	89.4
top-4 SVM^{α}	51.2	66.3	73.5	78.1	81.4	83.7	85.7	87.3	88.7	89.7
top-5 SVM^{α}	50.8	66.2	73.7	78.2	81.4	83.9	85.8	87.5	88.9	90.0
top-10 SVM^{α}	50.1	65.8	73.4	78.3	81.6	84.0	86.0	87.6	89.0	90.1
top-1 SVM_1^{α}	51.8	66.4	73.5	78.1	81.4	83.9	85.7	87.4	88.7	89.8
top-2 SVM_1^{α}	51.5	66.4	73.5	78.1	81.4	83.8	85.7	87.3	88.6	89.7
top-3 SVM_1^{α}	51.5	66.4	73.5	78.1	81.4	83.8	85.7	87.4	88.7	89.8
top-4 SVM_1^{α}	51.3	66.4	73.7	78.1	81.5	83.8	85.9	87.5	88.8	89.9
top-5 SVM_1^{α}	50.9	66.2	73.6	78.2	81.5	83.9	85.9	87.5	88.9	90.0
top-10 SVM_1^{α}	50.2	65.8	73.4	78.3	81.7	84.0	86.0	87.6	89.0	90.2
top-1 SVM^{β} / $\mathrm{SVM}^{\mathrm{Multi}}$	50.6	64.5	71.4	75.5	78.5	80.7	82.5	84.0	85.1	86.2
top-2 SVM^{β}	51.0	65.6	72.7	77.4	80.6	82.9	84.9	86.1	87.4	88.4
top-3 SVM^{β}	51.3	66.0	73.4	77.9	81.3	83.6	85.6	87.1	88.3	89.3
top-4 SVM^{β}	51.4	66.2	73.6	78.0	81.4	83.8	85.7	87.3	88.7	89.8
top-5 SVM^{β}	51.3	66.3	73.7	78.3	81.4	83.8	85.8	87.5	88.8	89.9
top-10 SVM^{β}	50.9	66.1	73.5	78.4	81.7	84.0	86.0	87.6	89.0	90.2
top-1 SVM_1^{β}	51.8	66.4	73.5	78.1	81.4	83.9	85.7	87.4	88.7	89.8
top-2 SVM_1^{β}	51.7	66.4	73.5	78.0	81.4	83.9	85.7	87.4	88.7	89.8
top-3 SVM_1^{β}	51.5	66.3	73.7	78.2	81.4	83.8	85.8	87.4	88.8	89.8
top-4 SVM_1^{β}	51.5	66.4	73.7	78.3	81.5	83.8	85.8	87.4	88.8	89.9
top-5 SVM_1^{β}	51.3	66.4	73.7	78.3	81.4	83.8	85.8	87.5	88.8	90.0
top-10 SVM_1^{β}	51.0	66.1	73.5	78.3	81.7	84.0	86.0	87.6	89.0	90.2
top-1 Ent / $\mathrm{LR}^{\mathrm{Multi}}$	51.1	66.1	73.5	78.1	81.5	84.1	85.9	87.6	88.9	90.0
top-2 Ent	51.0	66.1	73.4	78.1	81.5	84.0	85.8	87.6	88.9	90.0
top-3 Ent	50.9	66.1	73.4	78.1	81.5	83.9	85.8	87.5	88.9	89.9
top-4 Ent	50.7	66.0	73.3	78.0	81.5	83.9	85.7	87.5	88.9	89.9
top-5 Ent	50.3	65.8	73.3	77.8	81.3	83.9	85.7	87.3	88.8	89.9
top-10 Ent	48.9	64.9	72.7	77.5	81.0	83.7	85.6	87.2	88.7	89.8

Table 21: Comparison of different methods in top-k accuracy (%).

Places 205 (VGG-16 trained on Places 205, FC7 output)

Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
top-1 SVM^{α} / $\text{SVM}^{\text{Multi}}$	58.4	72.5	78.7	82.3	84.7	86.4	87.5	88.4	89.2	89.9
top-2 SVM^{α}	58.6	73.4	80.1	84.1	86.6	88.5	89.9	90.8	91.6	92.2
top-3 SVM^{α}	58.6	73.7	80.3	84.5	87.3	89.3	90.8	91.8	92.6	93.3
top-4 SVM^{α}	58.6	73.8	80.5	84.6	87.4	89.5	91.0	92.1	93.0	93.8
top-5 SVM^{α}	58.4	73.8	80.5	84.5	87.4	89.5	91.1	92.3	93.2	94.0
top-10 SVM^{α}	58.0	73.2	80.4	84.6	87.4	89.6	91.2	92.5	93.5	94.3
top-1 SVM_1^{α}	59.2	74.2	80.5	84.6	87.3	89.6	91.1	92.2	93.2	93.8
top-2 SVM_1^{α}	59.0	73.9	80.4	84.6	87.5	89.6	91.1	92.2	93.1	93.7
top-3 SVM_1^{α}	58.9	74.0	80.5	84.6	87.6	89.6	91.1	92.3	93.2	93.9
top-4 SVM_1^{α}	58.7	73.8	80.5	84.6	87.4	89.6	91.1	92.3	93.2	94.1
top-5 SVM_1^{α}	58.5	73.8	80.5	84.5	87.5	89.5	91.2	92.3	93.2	94.1
top-10 SVM_1^{α}	58.0	73.2	80.4	84.5	87.5	89.6	91.3	92.5	93.5	94.3
top-1 SVM^{β} / $\text{SVM}^{\text{Multi}}$	58.4	72.5	78.7	82.3	84.7	86.4	87.5	88.4	89.2	89.9
top-2 SVM^{β}	58.6	73.6	80.0	83.9	86.4	88.3	89.6	90.6	91.4	92.0
top-3 SVM^{β}	58.8	73.9	80.4	84.5	87.2	89.2	90.7	91.7	92.6	93.2
top-4 SVM^{β}	58.8	73.9	80.6	84.6	87.4	89.6	91.0	92.1	93.0	93.7
top-5 SVM^{β}	58.9	74.0	80.6	84.7	87.5	89.6	91.0	92.2	93.2	94.0
top-10 SVM^{β}	58.7	74.0	80.7	84.8	87.6	89.7	91.3	92.4	93.5	94.2
top-1 SVM_1^{β}	59.2	74.2	80.5	84.6	87.3	89.6	91.1	92.2	93.2	93.8
top-2 SVM_1^{β}	59.0	74.2	80.5	84.6	87.4	89.6	91.0	92.2	93.1	93.7
top-3 SVM_1^{β}	59.0	74.1	80.6	84.7	87.5	89.7	91.1	92.2	93.1	93.8
top-4 SVM_1^{β}	58.9	74.0	80.7	84.7	87.5	89.7	91.1	92.3	93.2	94.0
top-5 SVM_1^{β}	58.9	74.0	80.7	84.7	87.5	89.7	91.1	92.3	93.3	94.2
top-10 SVM_1^{β}	58.7	74.0	80.7	84.8	87.6	89.7	91.3	92.4	93.5	94.3
top-1 Ent / LR^{Multi}	59.0	73.9	80.6	84.8	87.6	89.7	91.3	92.5	93.5	94.3
top-2 Ent	58.9	73.8	80.6	84.7	87.6	89.6	91.2	92.4	93.5	94.3
top-3 Ent	58.7	73.8	80.6	84.8	87.6	89.6	91.2	92.5	93.5	94.2
top-4 Ent	58.5	73.6	80.5	84.6	87.5	89.6	91.2	92.4	93.4	94.2
top-5 Ent	58.1	73.5	80.4	84.5	87.4	89.6	91.2	92.4	93.4	94.2
top-10 Ent	57.0	72.8	80.0	84.2	87.2	89.3	91.0	92.3	93.4	94.1
top-2 Ent_{tr}	57.7	73.4	80.1	84.2	87.2	89.4	90.8	92.0	93.2	94.0
top-3 Ent_{tr}	56.8	72.8	80.0	84.1	87.1	89.3	90.9	92.2	93.3	94.2
top-4 Ent_{tr}	55.2	71.6	79.2	83.6	86.9	89.2	90.8	92.0	93.0	94.1
top-5 Ent_{tr}	54.2	71.2	79.0	83.5	86.9	89.1	90.8	92.0	93.1	94.0
top-10 Ent_{tr}	51.1	68.4	76.9	82.2	85.8	88.3	90.2	91.7	92.9	93.8

Table 22: Comparison of different methods in top-k accuracy (%).

ILSVRC 2012 (AlexNet trained on ImageNet, FC7 output, provided by [28])

Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
top-1 SVM^α / $\text{SVM}^{\text{Multi}}$	56.6	67.3	72.4	75.4	77.7	79.3	80.8	82.0	82.9	83.7
top-2 SVM^α	56.6	68.1	73.2	76.4	78.6	80.3	81.7	82.8	83.7	84.6
top-3 SVM^α	56.6	68.3	73.6	76.8	79.0	80.7	82.1	83.2	84.2	85.0
top-4 SVM^α	56.5	68.4	73.8	77.1	79.3	81.1	82.4	83.5	84.5	85.3
top-5 SVM^α	56.5	68.4	73.8	77.2	79.4	81.1	82.5	83.7	84.6	85.4
top-10 SVM^α	55.9	68.2	73.8	77.3	79.8	81.4	82.8	84.0	85.0	85.8
top-1 SVM_1^α	57.1	68.3	73.5	76.7	78.9	80.6	82.0	83.1	84.1	84.9
top-2 SVM_1^α	56.7	68.4	73.6	76.8	79.0	80.8	82.1	83.2	84.2	85.0
top-3 SVM_1^α	56.6	68.4	73.8	77.0	79.2	80.9	82.4	83.5	84.4	85.2
top-4 SVM_1^α	56.6	68.5	73.9	77.1	79.4	81.1	82.5	83.7	84.6	85.3
top-5 SVM_1^α	56.5	68.5	73.9	77.3	79.5	81.2	82.6	83.7	84.6	85.5
top-10 SVM_1^α	55.9	68.2	73.8	77.4	79.7	81.4	82.8	84.0	85.0	85.8
top-1 SVM^β / $\text{SVM}^{\text{Multi}}$	56.6	67.3	72.4	75.4	77.7	79.3	80.8	82.0	82.9	83.7
top-2 SVM^β	56.9	68.0	73.2	76.2	78.5	80.2	81.6	82.7	83.6	84.4
top-3 SVM^β	57.0	68.3	73.5	76.7	78.9	80.6	82.0	83.1	84.0	84.8
top-4 SVM^β	57.0	68.4	73.6	76.9	79.1	80.8	82.2	83.4	84.3	85.1
top-5 SVM^β	57.1	68.4	73.7	76.9	79.3	81.0	82.4	83.5	84.5	85.2
top-10 SVM^β	56.9	68.4	73.9	77.3	79.5	81.2	82.7	83.8	84.8	85.6
top-1 SVM_1^β	57.1	68.3	73.5	76.7	78.9	80.6	82.0	83.1	84.1	84.9
top-2 SVM_1^β	57.1	68.3	73.6	76.7	78.9	80.6	82.0	83.2	84.2	84.9
top-3 SVM_1^β	54.6	66.1	71.5	74.7	77.1	78.7	80.2	81.3	82.3	83.1
top-4 SVM_1^β	57.1	68.5	73.8	77.0	79.2	80.9	82.3	83.5	84.5	85.2
top-5 SVM_1^β	57.1	68.5	73.8	77.0	79.3	81.0	82.5	83.6	84.6	85.3
top-10 SVM_1^β	56.9	68.5	73.9	77.3	79.6	81.2	82.7	83.8	84.8	85.7
top-1 Ent / LR^{Multi}	55.8	67.4	73.1	76.6	79.0	80.8	82.2	83.4	84.4	85.3
top-2 Ent	55.6	67.4	73.0	76.5	79.0	80.8	82.2	83.4	84.4	85.2
top-3 Ent	55.5	67.4	73.0	76.5	78.9	80.7	82.2	83.4	84.4	85.2
top-4 Ent	55.4	67.3	73.0	76.5	78.9	80.7	82.1	83.4	84.4	85.2
top-5 Ent	55.2	67.2	72.9	76.5	78.9	80.7	82.1	83.4	84.3	85.2
top-10 Ent	54.8	66.9	72.7	76.4	78.8	80.6	82.1	83.3	84.3	85.2

Table 23: Comparison of different methods in top-k accuracy (%).

ILSVRC 2012 (VGG-16 trained on ImageNet, FC7 output)

Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
top-1 SVM^{α} / $\mathrm{SVM}^{\mathrm{Multi}}$	68.3	78.6	82.9	85.4	87.0	88.2	89.2	89.9	90.6	91.1
top-2 SVM^{α}	68.3	79.3	83.7	86.3	87.8	89.0	89.9	90.6	91.3	91.8
top-3 SVM^{α}	68.2	79.5	84.0	86.5	88.1	89.3	90.2	91.0	91.6	92.1
top-4 SVM^{α}	68.0	79.6	84.1	86.6	88.3	89.5	90.4	91.2	91.8	92.3
top-5 SVM^{α}	67.8	79.5	84.1	86.6	88.2	89.5	90.5	91.2	91.9	92.4
top-10 SVM^{α}	67.0	79.0	83.8	86.5	88.3	89.6	90.6	91.4	92.1	92.6
top-1 SVM_1^{α}	68.7	79.5	83.9	86.4	88.0	89.3	90.2	90.9	91.6	92.1
top-2 SVM_1^{α}	68.5	79.6	84.0	86.5	88.1	89.3	90.2	91.0	91.7	92.2
top-3 SVM_1^{α}	68.2	79.6	84.1	86.6	88.2	89.4	90.3	91.1	91.8	92.3
top-4 SVM_1^{α}	68.0	79.7	84.2	86.7	88.4	89.6	90.5	91.2	91.8	92.3
top-5 SVM_1^{α}	67.9	79.6	84.1	86.6	88.4	89.6	90.5	91.3	92.0	92.5
top-10 SVM_1^{α}	67.1	79.1	83.8	86.5	88.3	89.6	90.6	91.4	92.1	92.6
top-1 SVM^{β} / $\mathrm{SVM}^{\mathrm{Multi}}$	68.3	78.6	82.9	85.4	87.0	88.2	89.2	89.9	90.6	91.1
top-2 SVM^{β}	68.6	79.2	83.6	86.1	87.6	88.9	89.8	90.6	91.2	91.7
top-3 SVM^{β}	68.5	79.5	83.9	86.4	88.0	89.2	90.1	90.8	91.5	91.9
top-4 SVM^{β}	68.4	79.5	84.0	86.6	88.2	89.4	90.3	91.0	91.6	92.1
top-5 SVM^{β}	68.4	79.6	84.1	86.6	88.2	89.5	90.4	91.1	91.7	92.2
top-10 SVM^{β}	68.0	79.5	84.0	86.6	88.3	89.6	90.6	91.3	92.0	92.5
top-1 SVM_1^{β}	68.7	79.5	83.9	86.4	88.0	89.3	90.2	90.9	91.6	92.1
top-2 SVM_1^{β}	68.7	79.5	84.0	86.5	88.1	89.3	90.2	91.0	91.6	92.1
top-3 SVM_1^{β}	68.6	79.5	84.1	86.6	88.1	89.4	90.3	91.0	91.6	92.2
top-4 SVM_1^{β}	68.5	79.6	84.1	86.6	88.3	89.5	90.4	91.1	91.7	92.2
top-5 SVM_1^{β}	68.4	79.6	84.1	86.6	88.3	89.5	90.4	91.2	91.8	92.3
top-10 SVM_1^{β}	68.0	79.5	84.1	86.6	88.3	89.6	90.6	91.4	92.0	92.5
top-1 Ent / $\mathrm{LR}^{\mathrm{Multi}}$	67.2	78.5	83.2	85.9	87.7	89.1	90.1	91.0	91.7	92.2
top-2 Ent	67.1	78.4	83.2	85.9	87.8	89.1	90.1	91.0	91.7	92.2
top-3 Ent	66.8	78.4	83.1	85.9	87.8	89.1	90.1	91.0	91.7	92.2
top-4 Ent	66.7	78.3	83.1	85.8	87.8	89.1	90.1	91.0	91.7	92.2
top-5 Ent	66.5	78.2	83.0	85.8	87.7	89.1	90.1	91.0	91.6	92.2
top-10 Ent	65.8	77.8	82.8	85.7	87.6	89.0	90.0	90.9	91.6	92.1
top-2 $\mathrm{Ent}_{\mathrm{tr}}$	66.6	78.1	83.0	85.7	87.6	89.0	90.0	90.8	91.6	92.1
top-3 $\mathrm{Ent}_{\mathrm{tr}}$	65.9	77.8	82.8	85.7	87.5	89.0	89.9	90.8	91.5	92.1
top-4 $\mathrm{Ent}_{\mathrm{tr}}$	66.0	77.7	82.7	85.6	87.5	88.9	89.9	90.8	91.5	92.1
top-5 $\mathrm{Ent}_{\mathrm{tr}}$	65.0	77.1	82.3	85.2	87.3	88.7	89.8	90.6	91.3	91.9
top-10 $\mathrm{Ent}_{\mathrm{tr}}$	64.6	76.7	82.0	85.0	87.0	88.5	89.6	90.5	91.2	91.8

Table 24: Comparison of different methods in top-k accuracy (%).