
An end-to-end approach for the verification problem: learning the right distance

João Monteiro^{1,2} Isabela Albuquerque¹ Jahangir Alam^{1,2} R Devon Hjelm^{3,4} Tiago Falk¹

Abstract

In this contribution, we augment the metric learning setting by introducing a parametric pseudo-distance, trained jointly with the encoder. Several interpretations are thus drawn for the learned distance-like model’s output. We first show it approximates a likelihood ratio which can be used for hypothesis tests, and that it further induces a large divergence across the joint distributions of pairs of examples from the same and from different classes. Evaluation is performed under the verification setting consisting of determining whether sets of examples belong to the same class, even if such classes are novel and were never presented to the model during training. Empirical evaluation shows such method defines an end-to-end approach for the verification problem, able to attain better performance than simple scorers such as those based on cosine similarity and further outperforming widely used downstream classifiers. We further observe training is much simplified under the proposed approach compared to metric learning with actual distances, requiring no complex scheme to harvest pairs of examples.

1. Introduction

Learning useful representations from high-dimensional data is one of the main goals of modern machine learning. However, doing so is generally a side effect of the solution of a pre-defined task, e.g., while learning the decision surface in a classification problem, inner layers of artificial neural networks are shown to make salient cues of input data which are discriminable. Moreover, in unsupervised

settings, bottleneck layers of autoencoders as well as approximate posteriors from variational autoencoders have all been shown to embed relevant properties of input data which can be leveraged in downstream tasks. Rather than employing a neural network to solve some task and hope learned features are useful, approaches such as *siamese networks* (Bromley et al., 1994), which can be included in a set of approaches commonly referred to as *Metric Learning*, have been introduced with the goal of explicitly inducing features holding desirable properties such as class separability. In this setting, an encoder is trained so as to minimize or maximize a *distance* measured across pairs of encoded examples, depending on whether the examples within each pair belong to the same class or not, provided that class labels are available. Follow-up work leveraged this idea for several applications (Hadsell et al., 2006; Hoffer & Ailon, 2015), which include, for instance, the verification problem in biometrics, as is the case of FaceNet (Schroff et al., 2015) and Deep-Speaker (Li et al., 2017), which are used for face and speaker recognition, respectively. However, as pointed out in recent work (Schroff et al., 2015; Shi et al., 2016; Wu et al., 2017; Li et al., 2017; Zhang et al., 2018), careful selection of training pairs is crucial to ensure a reasonable sample complexity during training given that most triplets of examples quickly reach the condition such that distances measured between pairs from the same class are smaller than those of the pairs from different classes. As such, developing efficient strategies for harvesting negative pairs with small distances throughout training becomes primordial.

In this contribution, we are concerned with the metric learning setting briefly described above, and more specifically, we turn our attention to its application to the verification problem, i.e., that of comparing data pairs and determining whether they belong to the same class. The verification problem arises in applications where comparisons of two small samples is required such as face/finger-print/voice verification (Reynolds, 2002), image retrieval (Zhu et al., 2016; Wu et al., 2017), and so on. At test time, inference is often performed to answer two types of questions: (i) Do two given examples belong to the same class? and (ii) Does a test example belong to a specific claimed class? And in both cases test examples might belong to classes never presented to the model during training. Current verification

¹INRS-EMT, Université du Québec, Montreal, Canada. ²Centre de Recherche Informatique de Montréal, Montreal, Canada. ³Microsoft Research. ⁴Quebec Artificial Intelligence Institute, Université de Montréal, Montreal, Canada.. Correspondence to: João Monteiro <joao.monteiro@emt.inrs.ca>.

approaches are usually comprised of several components trained in a greedy manner (Kenny et al., 2013; Snyder et al., 2018b), and an end-to-end approach is still lacking.

Euclidean spaces will not, in general, be suitable for representing any desired type of structure expressed in the data (e.g. asymmetry (Pitis et al., 2020) or hierarchy (Nickel & Kiela, 2017)). To avoid the need to select an adequate distance given every new problem we are faced with, as well as to deal with the training difficulties mentioned previously, we propose to augment the metric learning framework and jointly train an encoder (which embeds raw data into a lower dimensional space) and a (pseudo) distance model tailored to the problem of interest. An end-to-end approach for verification is then defined by employing such pseudo-distance to compute similarity scores. Both models together, parametrized by neural networks, define a (pseudo) metric space in which inference can be performed efficiently since now semantic properties of the data (e.g., discrepancies across classes) are encoded by scores. While doing so, we found several interpretations appear from such learned pseudo-distance, and it can be further interpreted as a likelihood ratio in a Neyman-Pearson hypothesis test, as well as an approximate divergence measure between the joint distributions of positive (same classes) and negative (different classes) pairs of examples. Moreover, even though we do not enforce models to satisfy properties of an actual metric¹, we empirically observe such properties to appear.

Our contributions can be summarized as follows:

1. We propose an augmented metric learning framework where an encoder and a (pseudo) distance are trained jointly and define a (pseudo) metric space where inference can be done efficiently for verification.
2. We show that the optimal distance model for any fixed encoder yields the likelihood-ratio for a Neyman-Pearson hypothesis test, and it further induces a high Jensen-Shannon divergence between the joint distributions of positive and negative pairs.
3. The introduced setting is trained in an end-to-end fashion, and inference can be performed with a single forward pass, greatly simplifying current verification pipelines which involve several sub-components.
4. Evaluation on large scale verification tasks provides empirical evidence of the effectiveness in directly using outputs of the learned pseudo-distance for inference, outperforming commonly used downstream classifiers.

The remainder of this paper is organized as follows: metric learning and the verification problem are discussed in Section 2. The proposed method is presented in Section 3 along with our main guarantees, while empirical evaluation is presented in Section 4. Discussion and final remarks as well as future directions are presented in Section 5.

¹Symmetry, identity of indiscernibles, and triangle inequality.

2. Background and related work

2.1. Distance Metric Learning

Being able to efficiently assess similarity across samples from data under analysis is a long standing problem within machine learning. Algorithms such as K-means, nearest-neighbors classifiers, and kernel methods generally rely on the selection of some similarity or distance measure able to encode semantic relationships present in high-dimensional data into real scores. Under this view, approaches commonly referred to as *Distance Metric Learning*, introduced originally by Xing et al. (2003), try to learn a so-called Mahalanobis distance, which, given $x, y \in \mathbb{R}^n$, will have the form: $\sqrt{(x-y)^\top A(x-y)}$, where $A \in \mathbb{R}^{n \times n}$ is positive semidefinite. Several extensions of that setting were then introduced (Globerson & Roweis, 2006; Weinberger & Saul, 2009; Ying & Li, 2012).

Shalev-Shwartz et al. (2004), for instance, proposed an on-line version of the algorithm in (Xing et al., 2003), while an approach based on support vector machines was introduced in (Schultz & Joachims, 2004) for learning A . Davis et al. (2007) provided an information-theoretic approach to solve for A by minimizing the divergence between Gaussian distributions associated to the learned and the Euclidean distances, further showing such an approach to be equivalent to low-rank kernel learning (Kulis et al., 2006). Similar distances have also been used in other settings, such as similarity scoring for contrastive learning (Oord et al., 2018; Tian et al., 2019). Besides the Mahalanobis distance, other forms of distance/similarity have been considered in recent work. In (Lanckriet et al., 2004), for example, a kernel matrix is directly learned, implicitly defining a similarity function. In (Pitis et al., 2020), classes of neural networks are proposed to define pseudo-distances which satisfy the triangle inequality while not being necessarily symmetric.

For the particular case of Mahalanobis distance metric learning, one can show that $\exists W : \sqrt{(x-y)^\top A(x-y)} = \|Wx - Wy\|_2$ (Shalev-Shwartz et al., 2004), which means that there exists a linear projection of the data after which the Euclidean distance will correspond to the Mahalanobis distance on the original space. Chopra et al. (2005) substituted the linear projection by a learned non-linear encoder $\mathcal{E} : R^D \rightarrow R^d$ so that $\|\mathcal{E}(x) - \mathcal{E}(y)\|_2$ yields a (non-Mahalanobis) distance measure between raw data points yielding useful properties. Follow-up work has extended such idea to several applications (Schroff et al., 2015; Shi et al., 2016; Li et al., 2017; Zhang et al., 2018). One extra variation of $\|Wx - Wy\|_2$, besides the introduction of \mathcal{E} , is to switch the Euclidean distance $\|\cdot\|_2$ with an alternative better suited for the task of interest. That is the case in (Norouzi et al., 2012), where the Hamming distance is used over data encoded to a binary space. In (Courty et al., 2018), in turn, the encoder is trained so that Euclidean distances

in the encoded space approximate Wasserstein divergences, while Nickel & Kiela (2018) employ a hyperbolic distance which is argued to be suitable for their particular use case.

Based on the covered literature, one can conclude that there are two different directions aimed at achieving a similar goal: *learn to represent the data in a metric space where distances yield efficient inference mechanisms for various tasks*. While one corresponds to learning a meaningful distance or similarity from raw data, the other corresponds to, given a fixed distance metric, finding an encoding process yielding the desired metric space. Here, we propose an alternative to perform both these tasks *simultaneously*, i.e., jointly learn both the encoder and distance. Close to such an approach is the method discussed by Garcia & Vogiatzis (2019) where, similarly to our setting, both encoder and distance are trained, with the main differences lying in the facts that our method is fully end-to-end² while in their case training happens separately. Moreover, training of the distance model in that case is done by imitation learning of cosine similarities.

2.2. The Verification Problem

Given data instances $x \in \mathcal{X}$ such that each x can be associated to a class label $y \in \mathcal{Y}$ through a labeling function $f : \mathcal{X} \rightarrow \mathcal{Y}$, we define a *trial* as a pair of sets of examples $\{X_i, X_j\}$, provided that $f(x_i^k) = f(x_i^l) \forall k, l \in \{1, 2, \dots, |X_i|\}^2$ and $f(x_j^k) = f(x_j^l) \forall k, l \in \{1, 2, \dots, |X_j|\}^2$, so that we can assign class labels to such sets X_m defining $f(X_m) = f(x_m) \forall x_m \in X_m$. The verification problem can be thus viewed as, given a trial $T_{i,j} = \{X_i, X_j\}$, deciding whether $f(X_i) = f(X_j)$, in which case we refer to T as *target trial*, or $f(X_i) \neq f(X_j)$ and the trial will be called *non-target*.

The verification problem is illustrated in Figure 1. We categorize trials into two types in accordance to practical instances of the verification problem: type I trials are those such that X_i is referred to as enrollment sample, i.e., a set of data points representing a given class such as a gallery of face pictures from a given user in an access control application, while X_j will correspond to a single example x_{test} to be verified against the enrollment gallery. For the type II case, X_i is simply a *claim* corresponding to the class against which x_{test} will be verified. Classes corresponding to examples within test trials might have never been presented to the model, and sets X_i and X_j are typically small (< 10).

Under the Neyman-Pearson approach (Neyman & Pearson, 1933), verification is seen as a hypothesis test, where H_0 and H_1 correspond to the hypothesis such that T is target or otherwise, respectively (Jiang & Deng, 2001). The test is

²What authors refer to as end-to-end requires pretraining an encoder in the metric learning setting with a standard distance.

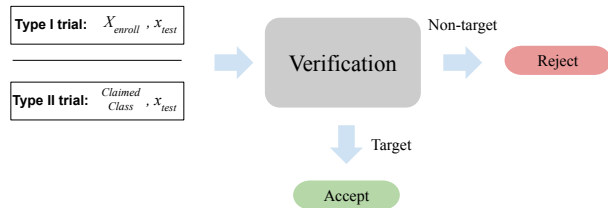


Figure 1. The verification problem.

thus performed through the following likelihood ratio (LR):

$$LR = \frac{p(T|H_0)}{p(T|H_1)}, \quad (1)$$

where $p(T|H_0)$ and $p(T|H_1)$ correspond to models of target, and non-target (or impostor) trials. The decision is made by comparing LR with a threshold τ .

One can then explicitly approximate LR through generative approaches (Deng & O’Shaughnessy, 2018), which is commonly done using Gaussian mixture models. In that case, the denominator is usually defined as a universal background model (GMM-UBM, Reynolds et al. (2000)), meaning that it is trained on data from all available classes, while the numerator is a fine-tuned model on enrollment data so that, for trial $\{X_1, X_2\}$, LR will be:

$$LR = \frac{p_{X_1}(X_2)}{p_{UBM}(X_2)} = \frac{p_{X_{Enroll}}(x_{test})}{p_{UBM}(x_{test})}. \quad (2)$$

Alternatively, Cumani et al. (2013) showed that discriminative settings, i.e., binary classifiers trained on top of data pairs to determine whether they belong to the same class, yielded likelihood ratios useful for verification. In their case, a binary SVM was trained on pairs of i-vectors (Dehak et al., 2010) for automatic speaker verification. We build upon such discriminative setting, but with the difference that we learn an encoding process along with the discriminator (here represented as a distance model), and show it to yield likelihood ratios required for verification through contrastive estimation results. This is more general than the result in (Cumani et al., 2013), which shows that there exists a generative classifier associated to each discriminator whose likelihood ratio matches the discriminator’s output, requiring such classifier’s assumptions to hold.

We remark that current verification approaches are composed of complex pipelines containing several components (Dehak et al., 2010; Kenny et al., 2013; Snyder et al., 2018b), including a pretrained data encoder, followed by a downstream classifier, such as probabilistic linear discriminant analysis (PLDA) (Ioffe, 2006; Prince & Elder, 2007), and score normalization (Auckenthaler et al., 2000), each contributing practical issues (e.g., cohort selection) to the over-

all system. This renders both training and testing of such systems difficult. The approach proposed herein is a step towards end-to-end verification, i.e., from data to scores via a single forward pass, thus simplifying inference.

3. Learning pseudo metric spaces

We consider the setting where both an encoding mechanism, as well as some type of similarity or distance across data points are to be learned. Assume $\mathcal{E} : \mathbb{R}^D \rightarrow \mathbb{R}^d$ and $\mathcal{D} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow (0, 1)$ are deterministic mappings which will be referred to as encoder and distance model, respectively, and will be both parametrized by neural networks. Such entities resemble a metric-space, thus we will refer to it as *pseudo metric space*. We empirically observed that introducing distance properties in \mathcal{D} , i.e., by constraining it to be symmetric and enforcing it to satisfy the triangle inequality, did not result in improved performance, yet rendered training unstable. However, since trained models are found to approximately behave as an actual distance, we make use of the analogy, but further provide alternative interpretations of \mathcal{D} 's outputs.

Data samples are such that $x \in \mathcal{X} \subset \mathbb{R}^D$, and $z = \mathcal{E}(x)$ represents embedded data in \mathbb{R}^d . It will be usually the case that $D \gg d$. Once more, each data example can be further assigned to one of L class labels through a labeling function $f : \mathcal{X} \rightarrow \{1, \dots, L\}$. Moreover, we define positive and negative pairs of examples denoted by $+$ or $-$ superscripts such that $x^+ = \{x_i, x_j\} \implies f(x_i) = f(x_j)$, as well as $x^- = \{x_i, x_j\} \implies f(x_i) \neq f(x_j)$. The same notation is employed in the embedding space so that $z^+ = \mathcal{E}(x^+) = \{\mathcal{E}(x_i), \mathcal{E}(x_j)\} \implies f(x_i) = f(x_j)$, and $z^- = \mathcal{E}(x^-) = \{\mathcal{E}(x_i), \mathcal{E}(x_j)\} \implies f(x_i) \neq f(x_j)$. We will denote the sets of all possible positive and negative pairs by \mathcal{X}^+ and \mathcal{X}^- , respectively, and further define a probability distribution p over \mathcal{X} which, along with f , will yield p^+ and p^- over \mathcal{X}^+ and \mathcal{X}^- . Similarly to the setting in (Hjelm et al., 2018), which introduces a discriminator over pairs of samples, we are interested in \mathcal{E}^* and \mathcal{D}^* such that:

$$\begin{aligned} \mathcal{E}^*, \mathcal{D}^* \in \underset{\mathcal{E}, \mathcal{D}}{\operatorname{argmin}} & -\mathbb{E}_{x^+ \sim p^+} \log(\mathcal{D} \circ \mathcal{E}(x^+)) \\ & -\mathbb{E}_{x^- \sim p^-} \log(1 - \mathcal{D} \circ \mathcal{E}(x^-)), \end{aligned} \quad (3)$$

and \circ indicates composition so that $\mathcal{D} \circ \mathcal{E}(x^+) = \mathcal{D}(\mathcal{E}(x^+))$. Such problem is separable in the parameters of \mathcal{E} and \mathcal{D} and iterative solution strategies might include either alternate or simultaneous updates. We found the latter to converge faster in terms of wall-clock time and both approaches reach similar performance. We thus perform simultaneous updates while training.

The problem stated in (3) corresponds to finding \mathcal{E} and \mathcal{D} which will ensure that semantically close or distant samples, as defined through f , will preserve such properties in terms

of distance in the new space, while doing so in lower dimension. We stress the fact that class labels define which samples should be close together or far apart, which means that the same underlying data can yield different pseudo-metric spaces if different semantic properties are used to define class labels. For example, if one considers that, for a given set of speech recordings, class labels are equivalent to speaker identities, recordings from the same speaker are expected to be clustered together in the embedding space, while different results can be achieved if class labels are assigned corresponding to spoken language, acoustic conditions, and so on.

3.1. Different interpretations for \mathcal{D}

Besides the view of \mathcal{D} as a distance-like object defining a metric-like space $\{\mathcal{E}(\mathcal{X}), \mathcal{D}\}$, here we discuss some other possible interpretations of its outputs. We start by justifying the choice of the training objective defined in (3) by showing it to yield the likelihood ratio of particular trials of type I corresponding to a single enrollment example against a single test example, i.e. $T = \{x_{enroll}, x_{test}\}$. In both of the next two propositions, proofs directly reuse results from the contrastive estimation and generative adversarial networks literature (Gutmann & Hyvärinen, 2010; Goodfellow et al., 2014) to show \mathcal{D} can be used for verification.

Proposition 1. *The optimal \mathcal{D} for any fixed \mathcal{E} yields a simple transformation of the likelihood ratio stated in Eq. 1 for trials of the type $T = \{x_{enroll}, x_{test}\}$.*

Proof. We first define p_z^+ and p_z^- , which correspond to the counterparts of p^+ and p^- induced by \mathcal{E} in the embedding space. Now consider the loss \mathcal{L} defined in Eq. 3:

$$\begin{aligned} \mathcal{L} &= -\mathbb{E}_{z^+ \sim p_z^+} \log(\mathcal{D}(z^+)) - \mathbb{E}_{z^- \sim p_z^-} \log(1 - \mathcal{D}(z^-)) \\ &= -\int_{z^+} p_z^+(z^+) \log(\mathcal{D}(z^+)) - \int_{z^-} p_z^-(z^-) \log(1 - \mathcal{D}(z^-)) \\ &= -\int_{z'} p_z^+(z') \log(\mathcal{D}(z')) + p_z^-(z') \log(1 - \mathcal{D}(z')), \end{aligned} \quad (4)$$

where z' corresponds to $z^+ \cup z^-$ or equivalently $\mathcal{E}(\mathcal{X}^+) \cup \mathcal{E}(\mathcal{X}^-)$. Since $\mathcal{D}(z') \in (0, 1) \forall z' \in z'$, above integrand $p_z^+(z') \log(\mathcal{D}(z')) + p_z^-(z') \log(1 - \mathcal{D}(z'))$, provided that the set from which we pick candidate solutions is rich enough, has its maximum at:

$$\begin{aligned} \mathcal{D}^*(z') &= \frac{p_z^+(z')}{p_z^+(z') + p_z^-(z')}, \\ &= \frac{1}{1 + \left(\frac{p_z^-(z')}{p_z^+(z')}\right)^{-1}}. \end{aligned} \quad (5)$$

The last step above is of course only valid for $z' \in \operatorname{supp}(p_z^+)$. Nevertheless, $\mathcal{D}^*(z')$ is in any case meaningful for verification. In fact, as will be discussed in

Proposition 2, the optimal encoder is the one that induces $\text{supp}(p_z^+) \cap \text{supp}(p_z^-) = \emptyset$. Considering trial $T = \{x_{\text{enroll}}, x_{\text{test}}\}$, we can write the ratio $\frac{p_z^+(z')}{p_z^-(z')}$ as:

$$\frac{p_z^+(z')}{p_z^-(z')} = \frac{p_z^+(\mathcal{E}(x_{\text{enroll}}), \mathcal{E}(x_{\text{test}}))}{p_z^-(\mathcal{E}(x_{\text{enroll}}), \mathcal{E}(x_{\text{test}}))} := \frac{p(T|H_0)}{p(T|H_1)}. \quad \square \quad (6)$$

Proposition 1 indicates that the discussed setting can be used in an end-to-end fashion to yield verification decision rules against a threshold τ for trials of a specific type.

The following lemma will be necessary for the next result:

Lemma 1. *If $\text{supp}(p_z^+) \cap \text{supp}(p_z^-) = \emptyset$, any positive threshold $0 < \tau < \infty$ yields optimal decision rules for trials $T = \{x_{\text{enroll}}, x_{\text{test}}\}$.*

Proof. We prove the lemma by inspecting the decision rule under the considered assumptions in the two possible test cases: if T is non-target $\implies \frac{p_z^+(\mathcal{E}(x_{\text{enroll}}), \mathcal{E}(x_{\text{test}}))}{p_z^-(\mathcal{E}(x_{\text{enroll}}), \mathcal{E}(x_{\text{test}}))} = 0 < \tau$. If T is target $\implies \frac{p_z^+(\mathcal{E}(x_{\text{enroll}}), \mathcal{E}(x_{\text{test}}))}{p_z^-(\mathcal{E}(x_{\text{enroll}}), \mathcal{E}(x_{\text{test}}))} \rightarrow \infty > \tau$, completing the proof. \square

We now proceed and use the optimal discriminator into \mathcal{L} , which yields the following result for the optimal encoder:

Proposition 2. *Minimizing \mathcal{L} yields optimal decision rules for any positive threshold.*

Proof. We plug \mathcal{D}^* into \mathcal{L} so that for any z' we obtain:

$$\begin{aligned} \mathcal{L} &= -\mathbb{E}_{z' \sim p_z^+} \log \left(\frac{p_z^+(z')}{p_z^+(z') + p_z^-(z')} \right) \\ &\quad - \mathbb{E}_{z' \sim p_z^-} \log \left(\frac{p_z^-(z')}{p_z^+(z') + p_z^-(z')} \right) \\ &= -KL(p_z^+ \| p_z^+ + p_z^-) - KL(p_z^- \| p_z^+ + p_z^-) \\ &= \log 4 - 2JSD(p_z^+ \| p_z^-). \end{aligned} \quad (7)$$

\mathcal{L} is therefore minimized ($\mathcal{L}^* = 0$) iff \mathcal{E} yields $\text{supp}(p_z^+) \cap \text{supp}(p_z^-) = \emptyset$, which results in optimal decision rules for any positive threshold, invoking lemma 1, and assuming such encoders are available in the set one searches over. \square

We thus showed the proposed training scheme to be convenient for 2-sample tests under small sample regimes, such as in the case of verification, given that: (i) the distance model is also a discriminator which approximates the likelihood ratio of the joint distributions over positive and negative pairs³, and the encoder will be such that it induces a high divergence across such distributions, rendering their ratio amenable to decision making even in cases where verified samples are as small as single enrollment and test examples.

³The joint distribution over negative pairs is simply the product of marginals: $p^-(x_i, x_j) = p(x_i)p(x_j)$.

On a speculative note, we provide yet another view of \mathcal{D} by defining the kernel function $\mathcal{K} = \mathcal{D}$. If we assume \mathcal{K} to satisfy Mercer’s condition (which won’t likely be the case within our setting since \mathcal{K} will not be symmetric nor positive semidefinite), we can invoke Mercer’s theorem and state that there is a feature map to a Hilbert space where verification can be performed through inner products. Training in the described setting could be viewed such that minimizing \mathcal{L} becomes equivalent to building such a Hilbert space where classes can be distinguished by directly scoring data points one against the other. We hypothesize that constraining \mathcal{K} to sets where Mercer’s condition does hold might yield an effective approach for the problems we consider herein, which we intend to investigate in future work.

3.2. Training

We now describe the procedure we adopt to minimize \mathcal{L} as well as some practical design decisions made based on empirical results. Both \mathcal{E} and \mathcal{D} are implemented as neural networks. In our experiments, \mathcal{E} will be convolutional (2-d for images and 1-d for audio) while \mathcal{D} is a stack of fully-connected layers which take as input concatenated embeddings of pairs of examples. Training is carried out with standard minibatch stochastic gradient descent with momentum. We perform simultaneous update steps for \mathcal{E} and \mathcal{D} since we observed that to be faster than alternate updates, while yielding the same performance. Standard regularization strategies such as weight decay and label smoothing (Szegedy et al., 2016) are also employed. We empirically found that employing an auxiliary multi-class classification loss significantly accelerates training. Since our approach requires labels to determine which pairs of examples are positive or negative, we make further use of the labels to compute such auxiliary loss, which will be indicated by \mathcal{L}_{CE} . To allow for computation of \mathcal{L}_{CE} , we project $z = \mathcal{E}(x)$ onto the simplex Δ^{L-1} using a fully-connected layer. Minimization is then performed on the sum of the two losses, i.e., we solve $\mathcal{E}, \mathcal{D} \in \text{argmin } \mathcal{L}' = \mathcal{L} + \mathcal{L}_{CE}$, where the CE subscript in \mathcal{L}_{CE} indicates the multi-class cross-entropy loss.

All hyperparameters are selected with a random search over a pre-defined grid. For the particular case of the auxiliary loss \mathcal{L}_{CE} , besides the standard cross-entropy, we also ran experiments considering one of its so-called *large margin* variations. We particularly evaluated models trained with the additive margin softmax approach (Wang et al., 2018). The choice between the two types of auxiliary losses (standard or large margin) was a further hyperparameter and the decision was based on the random search over the two options. The grid used for hyperparameters selection along with the values chosen for each evaluation are presented in the appendix. A pseudocode describing our training procedure is presented in Algorithm 1.

Algorithm 1 Training procedure.

```

 $\mathcal{E}, \mathcal{D} = \text{InitializeModels}()$ 
repeat
   $x, y = \text{SampleMinibatch}()$ 
   $z = \mathcal{E}(x)$ 
   $z^+ = \text{GetAllPositivePairs}(z, y)$ 
   $z^- = \text{GetAllNegativePairs}(z, y)$ 
   $y' = \text{ProjectOntoSimplex}(z)$ 
   $\mathcal{L}' = \mathcal{L}(z^+, z^-, \mathcal{D}) + \mathcal{L}_{CE}(y', y)$ 
   $\mathcal{E}, \mathcal{D} = \text{UpdateRule}(\mathcal{E}, \mathcal{D}, \mathcal{L}')$ 
until Maximum number of iterations reached
return  $\mathcal{E}, \mathcal{D}$ 

```

Table 1. Evaluation of models trained under the proposed approach on image data.

		Scoring	EER	1-AUC
Cifar-10	Triplet	Cosine	3.80%	0.98%
		E2E	3.43%	0.60%
	Proposed	Cosine	3.56%	1.03%
		Cosine + E2E	3.42%	0.80%
Mini-ImageNet (Validation)	Triplet	Cosine	28.91%	21.58%
		E2E	28.64%	21.01%
	Proposed	Cosine	30.66%	23.70%
		Cosine + E2E	28.49%	20.90%
Mini-ImageNet (Test)	Triplet	Cosine	29.68%	22.56%
		E2E	29.26%	22.04%
	Proposed	Cosine	32.97%	27.34%
		Cosine + E2E	29.32%	22.24%

4. Evaluation

We proceed to evaluation of the described framework and do so with four sets of experiments. In the first part of our evaluation, we run proof-of-concept experiments and make use of standard image datasets to simulate verification settings. We report results on all trials created for the test sets of Cifar-10 and Mini-ImageNet. In the former, the same 10 classes of examples appear for both train and test partitions, in what we refer to as closed-set verification. For the case of Mini-ImageNet, since that dataset was designed for few-shot learning applications, we have an open-set evaluation for verification since there are 64, 16, and 20 disjoint classes of training, validation, and test examples.

We then move on to a *large scale realistic evaluation*. To this end, we make use of the recently introduced VoxCeleb corpus (Nagrani et al., 2017; Chung et al., 2018), corresponding to audio recordings of interviews taken from youtube videos, which means there’s no control over the acoustic conditions present in the data. Moreover, while most of the corpus corresponds to speech in English, other languages are also present, so that test recordings are from different speakers relative to the train data, and potentially also from different languages and acoustic environments. We specifically employ the second release of the corpus so that training data is

composed of recordings from 5994 speakers while three test sets are available: (i) **VoxCeleb1 Test set**, which is made up of utterances from 40 speakers, (ii) **VoxCeleb1-E**, i.e., the complete first release of the data containing 1251 speakers, and (iii) **VoxCeleb1-H**, corresponding to a sub-set of the trials in **VoxCeleb1-E** so that non-target trials are designed to be hard to discriminate by using the meta-data to match factors such as nationality and gender of the speakers. We then report experiments performed to observe whether \mathcal{D} ’s outputs present properties of actual distances, and finally check the influence of \mathcal{D} ’s architecture on final performance.

Our main baselines for proof-of-concept experiments correspond to the same encoders as in the evaluation of our proposed approach, while \mathcal{D} is dropped and replaced by the Euclidean distance. In those cases however, in order to get the most challenging baselines, we perform online selection of hard negatives. Our baselines closely follow the setting described in (Monteiro et al., 2019). All such baselines are referred to as *triplet* in the tables with results as a reference to the training loss in those cases. Unless specified, all models, baseline or otherwise, are trained from scratch, and the same computation budget is used for training and hyperparameter search for all models we trained.

Performance is assessed in terms of the difference to 1 of the area under the operating curve, indicated by 1-AUC in the tables, and also in terms of equal error rate (EER). EER indicates the operating point (i.e. threshold selection) at which the miss and false alarm rates are equal. Both metrics are better when closer to 0. We consider different strategies to score test trials. Both cosine similarity and PLDA are considered in some cases, and when the output of \mathcal{D} is directly used as a score we then indicate it by E2E in a reference to *end-to-end*⁴. We further remark that cosine similarity can also be used to score trials in our proposed setting, and we observed some performance gains when applying simple sum fusion of the two available scores. Additional implementation details are included in the appendix.

4.1. Cifar-10 and Mini-ImageNet

The encoder for evaluation on both Cifar-10 and Mini-ImageNet was implemented as a ResNet-18 (He et al., 2016). Results are reported in Table 1.

Results indicate the proposed scheme indeed yields effective inference strategies under the verification setting compared to traditional metric learning approaches, while using a more simplified training scheme since: (i) no sort of approach for harvesting hard negative pairs (e.g., (Schroff et al., 2015; Wu et al., 2017)) is needed in our case, and those are usually expensive, (ii) the method does not require large batch sizes,

⁴Scoring trials with cosine similarity can be also seen as end-to-end.

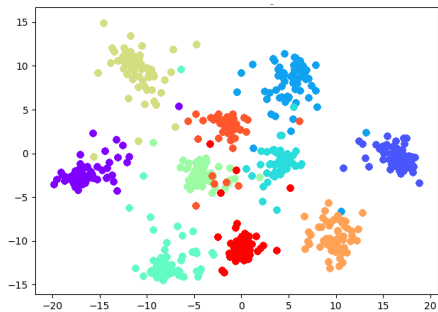


Figure 2. MNIST embeddings on a 2-dimensional space. Each color represents test examples corresponding to a digit from 0 to 9.

and (iii) we employ a simple loss with no hyperparameters that have to be tuned, as opposed to margin-based triplet or contrastive losses. We further highlight that the encoders trained with the proposed approach have the possibility for trials to be further scored with cosine similarities, which yields a performance improvement in some cases when combined with \mathcal{D} 's output

4.2. Large-scale verification with VoxCeleb

We now proceed and evaluate the proposed scheme in a more challenging scenario corresponding to realistic audio data for speaker verification. To do so, we implement \mathcal{E} as the well-known time-delay architecture (Waibel et al., 1989) employed within the x-vector setting, showed to be effective in summarizing speech into speaker- and spoken language-dependent representations (Snyder et al., 2018b;a). The model consists of a sequence of dilated 1-dimensional convolutions across the temporal dimension, followed by a time pooling layer, which simply concatenates element-wise first- and second-order statistics over time. Statistics are finally projected into an output vector through fully-connected layers. Speech is represented as 30 mel-frequency cepstral coefficients obtained with a short-time Fourier transform using a 25ms Hamming window with 60% overlap. All the data is downsampled to 16kHz beforehand. An energy-based voice activity detector is employed to filter out non-speech frames. We augment the data by creating noisy versions of training recordings using exactly the same approach as in (Snyder et al., 2018b). Model architecture and feature extraction details are included in the appendix.

We compared our models with a set of published results as well as the results provided by the popular Kaldi recipe⁵, considering scoring using cosine similarity or PLDA. For the Kaldi baseline, we found the same model as ours to yield relatively weak performance. As such, we decided to search over possible architectures in order to make it a

⁵Kaldi recipe: <https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb>

stronger baseline. We thus report the best model we could find which has the same structure as ours, i.e., it is made up of convolutions over time followed by temporal pooling and fully-connected layers, while the convolutional stack is deeper, which makes the comparison unfair in their favor.

We further evaluated our models using PLDA by running just the part of the same Kaldi recipe corresponding to the training of that downstream classifier on top of representations obtained from our encoder. Results are reported in Table 2 and support our claim that the proposed framework can be directly used in an end-to-end fashion. It is further observed that it outperformed standard downstream classifiers, such as PLDA, by a significant difference while not requiring any complex training procedure, as common metric learning approaches usually do. We employ simple random selection of training pairs. Ablation results are also reported, in which case we dropped the auxiliary loss \mathcal{L}_{CE} and trained the same \mathcal{E} and \mathcal{D} using the same budget in terms of number of iterations, showing that having the auxiliary loss improves performance in the considered evaluation.

Table 2. Evaluation of models trained under the proposed approach on VoxCeleb.

	Scoring	Training set	EER
VoxCeleb1 Test set			
Nagrani et al. (2017)	PLDA	VoxCeleb1	8.80%
Cai et al. (2018)	Cosine	VoxCeleb1	4.40%
Okabe et al. (2018)	Cosine	VoxCeleb1	3.85%
Hajibabaei & Dai (2018)	Cosine	VoxCeleb1	4.30%
Ravanelli & Bengio (2019)	Cosine	VoxCeleb1	5.80%
Chung et al. (2018)	Cosine	VoxCeleb2	3.95%
Xie et al. (2019)	Cosine	VoxCeleb2	3.22%
Hajavi & Etemad (2019)	Cosine	VoxCeleb2	4.26%
Xiang et al. (2019)	Cosine	VoxCeleb2	2.69%
Kaldi recipe ⁵	PLDA	VoxCeleb2	2.51%
Proposed	Cosine	VoxCeleb2	4.97%
Proposed	E2E	VoxCeleb2	2.51%
Proposed	Cosine + E2E	VoxCeleb2	2.51%
Proposed	PLDA	VoxCeleb2	3.75%
Ablation ($-\mathcal{L}_{CE}$)	E2E	VoxCeleb2	3.44%
VoxCeleb1-E			
Chung et al. (2018)	Cosine	VoxCeleb2	4.42%
Xie et al. (2019)	Cosine	VoxCeleb2	3.13%
Xiang et al. (2019)	Cosine	VoxCeleb2	2.76%
Kaldi recipe ⁵	PLDA	VoxCeleb2	2.60%
Proposed	Cosine	VoxCeleb2	4.77%
Proposed	E2E	VoxCeleb2	2.57%
Proposed	Cosine + E2E	VoxCeleb2	2.53%
Proposed	PLDA	VoxCeleb2	3.61%
Ablation ($-\mathcal{L}_{CE}$)	E2E	VoxCeleb2	3.70%
VoxCeleb1-H			
Chung et al. (2018)	Cosine	VoxCeleb2	7.33%
Xie et al. (2019)	Cosine	VoxCeleb2	5.06%
Xiang et al. (2019)	Cosine	VoxCeleb2	4.73%
Kaldi recipe ⁵	PLDA	VoxCeleb2	4.62%
Proposed	Cosine	VoxCeleb2	8.61%
Proposed	E2E	VoxCeleb2	4.73%
Proposed	Cosine + E2E	VoxCeleb2	4.69%
Proposed	PLDA	VoxCeleb2	5.98%
Ablation ($-\mathcal{L}_{CE}$)	E2E	VoxCeleb2	7.76%

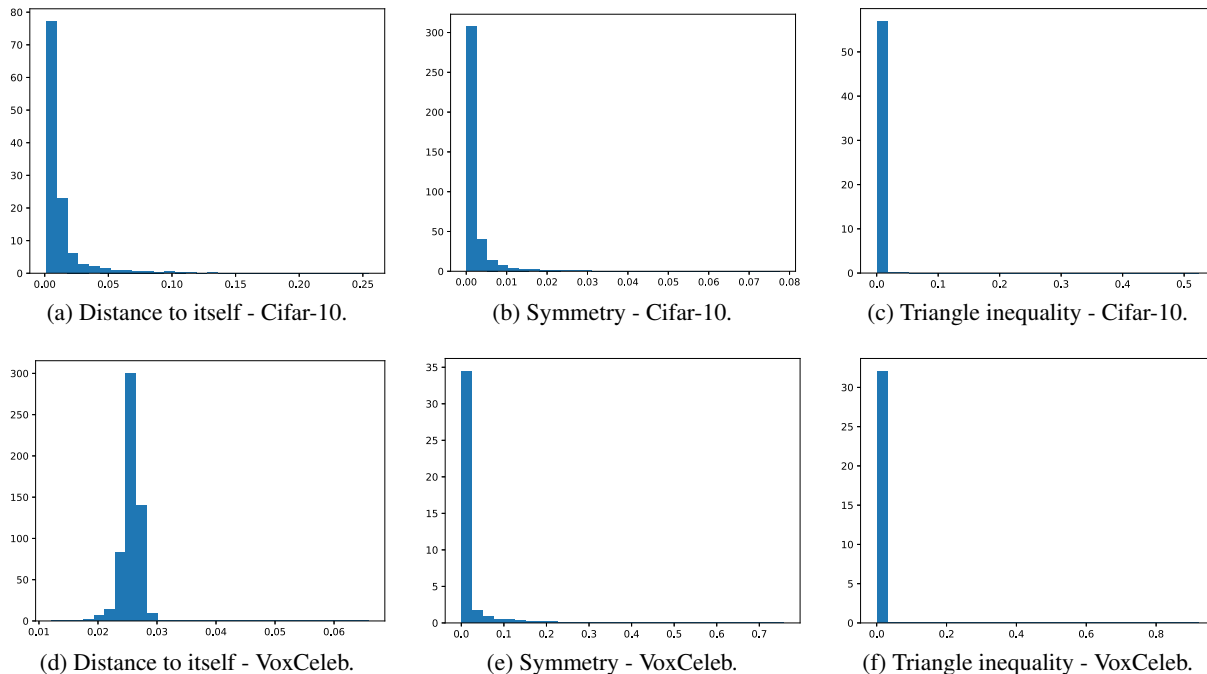


Figure 3. Evaluation of properties given by outputs of $\mathcal{D}' = 1 - \mathcal{D}$.

4.3. Checking for distance properties in \mathcal{D}

We now empirically evaluate how \mathcal{D} behaves in terms of properties of distances or metrics such as symmetry, for instance. We start by plotting embeddings from \mathcal{E} and do so by training an encoder on MNIST under the proposed setting (without the auxiliary loss \mathcal{L}_{CE} in this case) so that its outputs are given by $z \in \mathbb{R}^2$. We then plot the embeddings of the complete MNIST’s test set in Fig. 2, where the raw embeddings in \mathbb{R}^2 are directly displayed in the plot. Interestingly, classes are reasonably clustered in the Euclidean space even if such behavior was never enforced during training. We proceed and directly check for distance properties in $\mathcal{D}' = 1 - \mathcal{D}$. For the test set of Cifar-10 as well as for **VoxCeleb1 Test set**, we plot histograms of (i) the distance to itself for all test examples, (ii) a symmetry measure given by the absolute difference of the outputs of \mathcal{D}' measured in the two directions for all possible test pairs, and (iii) a measure of how much \mathcal{D}' satisfies the triangle inequality, which we do by measuring $\max[\mathcal{D}'(b, c) - (\mathcal{D}'(a, b) + \mathcal{D}'(a, c)), 0]$ for a random sample taken from all possible triplets of examples $\{a, b, c\}$. Proper metrics should have all such quantities equal 0. In Figures 3-a to 3-f, it can be seen that once more, even if any particular behavior is enforced over \mathcal{D} at its training phase, resulting models approximately behave as proper metrics. We thus hypothesize the relatively easier training observed in our setting, in the sense that it works without complicated schemes for selection of negative pairs, is due to the not so constrained distances induced by \mathcal{D} .

4.4. Varying the depth of \mathcal{D} for verification on ImageNet

We performed closed-set verification on the full ImageNet with distance models of increasing depths (1 to 5) to verify whether our setting is stable with respect to some of the introduced hyperparameters. With this experiment, we specifically intend to assess how difficult it would be in practice to find a good architecture for the distance model. Our models are compared against encoders with the same architecture, but trained using a standard metric learning approach, i.e the same training scheme as that employed for baselines reported in Table 1.

For this case, the encoder \mathcal{E} is implemented as the convolutional stack of a ResNet-50 followed by a fully-connected layer used to project the output representations to the desired dimensionality, and we employ an embedding dimension of 128 across all reported models. \mathcal{D} is once more implemented as a stack of fully-connected layers in which case we set the sizes of all hidden layers to 256. Training is performed such that the parameters of the convolutional portion of \mathcal{E} are initialized from a pretrained model for multi-class classification on ImageNet, and this approach is used for both our models as well as the baseline. We then perform stochastic gradient descent on the combined loss discussed in Section 3 using the standard multi-class cross entropy as auxiliary loss. Moreover, given the large number of classes in ImageNet compared to commonly used batch sizes, in order to be able to always find positive pairs throughout

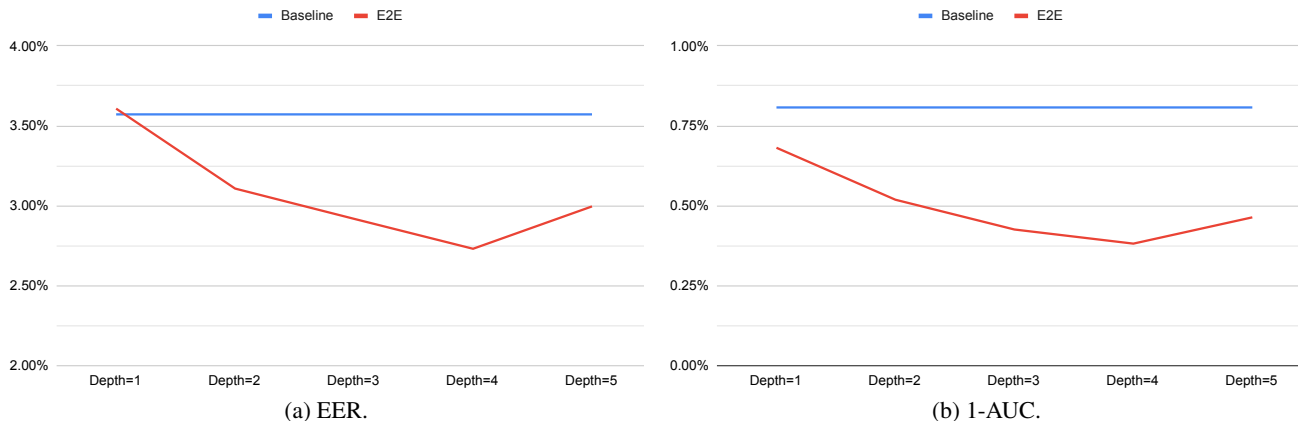


Figure 4. Verification on all trials created by pairing all test examples of ImageNet. Results indicate that defining the architecture of the distance model is not difficult in practice given that models of varying depths yield a relatively small performance range.

training, minibatches are constructed using the same strategy as that employed for experiments with *VoxCeleb*, i.e. we ensure at least 5 examples per class appear in each minibatch. The learning rate is set to 0.001 and is reduced by a factor of 0.1 every 10 epochs. Training is carried out for 50 epochs. Evaluation is performed over trials obtained from building all possible pairs of examples from the test partition of ImageNet. Results are reported in Figures 4-a and 4-b in terms of EER and the area over the operating curve (1-AUC), respectively. Scoring for the case of baseline encoders is performed with cosine similarity between encoded examples from test trials. While standard metric learning encoders make for strong baselines, all evaluated distance models are able to perform on pair (depth=1) or better than (depth>1) such models.

The results discussed herein provide empirical evidence for the claim that tuning the hyperparameters we introduced in comparison to previous settings, i.e. the architecture of the distance model, is not so challenging in that we achieve reasonably stable performance for verification on ImageNet when varying the depth of the distance model. Yet another empirical finding supporting that claim consists of the fact that similar architectures of the distance model were found to work well across all the datasets/domains we evaluated on. We specifically found that distance models with 3 or 4 hidden layers with 256 units each work well across datasets, which we believe might be a reasonable starting point for extending the approach we discussed to other datasets.

5. Conclusion

We introduced an end-to-end setting particularly tailored to perform small sample 2-sample tests and compare data pairs to determine whether they belong to the same class. Several interpretations of such framework are provided, including

joint encoder and distance metric learning, as well as contrastive estimation over data pairs. We used contrastive estimation results to show the solutions of the posed problem yield optimal decision rules under verification settings, resulting in correct decisions for any choice of threshold. In terms of practical contributions, the proposed method simplifies both the training under the metric learning framework, as it does not require any scheme to select negative pairs of examples, and also simplifies verification pipelines, which are usually made up of several individual components, each one contributing specific challenges at training and testing phases. Our models can be used in an end-to-end fashion by using \mathcal{D} 's outputs to score test trials yielding strong performance even in large scale and realistic open-set conditions where test classes are different from those seen at train time⁶. The proposed approach can be extended to any setting relying on distances to do inference such as image retrieval, prototypical networks (Snell et al., 2017), and clustering. Similarly to extensions of GANs (Nowozin et al., 2016; Arjovsky et al., 2017), variations of our approach where \mathcal{E} maximizes other types of divergences instead of Jensen-Shannon's might also be a relevant future research direction, requiring corresponding decision rules to be defined.

Acknowledgements

Part of the authors wish to acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) through grant RGPIN-2019-05381. The first author was partially funded by the *Bourse du CRIM pour Études Supérieures*.

⁶Code for reproducing our experiments can be found at: https://github.com/joaomonteirof/e2e_verification

References

- Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.
- Auckenthaler, R., Carey, M., and Lloyd-Thomas, H. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1-3):42–54, 2000.
- Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. Signature verification using a siamese time delay neural network. In *Advances in neural information processing systems*, pp. 737–744, 1994.
- Cai, W., Chen, J., and Li, M. Exploring the encoding layer and loss function in end-to-end speaker and language recognition system. In *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, pp. 74–81, 2018. doi: 10.21437/Odyssey.2018-11. URL <http://dx.doi.org/10.21437/Odyssey.2018-11>.
- Chopra, S., Hadsell, R., and LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pp. 539–546. IEEE, 2005.
- Chung, J. S., Nagrani, A., and Zisserman, A. Voxceleb2: Deep speaker recognition. *Proc. Interspeech 2018*, pp. 1086–1090, 2018.
- Courty, N., Flamary, R., and Ducoffe, M. Learning wasserstein embeddings. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=SJyEH91A->.
- Cumani, S., Brümmer, N., Burget, L., Laface, P., Plhot, O., and Vasilakakis, V. Pairwise discriminative speaker verification in the i-vector space. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(6):1217–1227, 2013.
- Davis, J. V., Kulis, B., Jain, P., Sra, S., and Dhillon, I. S. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pp. 209–216. ACM, 2007.
- Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2010.
- Deng, L. and O’Shaughnessy, D. *Speech processing: a dynamic and optimization-oriented approach*. CRC Press, 2018.
- Garcia, N. and Vogiatzis, G. Learning non-metric visual similarity for image retrieval. *Image and Vision Computing*, 82:18–25, 2019.
- Garcia-Romero, D., McCree, A., Shum, S., Brummer, N., and Vaquero, C. Unsupervised domain adaptation for i-vector speaker recognition. In *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2014.
- Globerson, A. and Roweis, S. T. Metric learning by collapsing classes. In *Advances in neural information processing systems*, pp. 451–458, 2006.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.
- Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 297–304, 2010.
- Hadsell, R., Chopra, S., and LeCun, Y. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1735–1742. IEEE, 2006.
- Hajavi, A. and Etemad, A. A deep neural network for short-segment speaker recognition. *Proc. Interspeech 2019*, pp. 2878–2882, 2019.
- Hajibabaei, M. and Dai, D. Unified hypersphere embedding for speaker recognition. *arXiv preprint arXiv:1807.08312*, 2018.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.
- Hoffer, E. and Ailon, N. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pp. 84–92. Springer, 2015.
- Ioffe, S. Probabilistic linear discriminant analysis. In *European Conference on Computer Vision*, pp. 531–542. Springer, 2006.
- Jiang, H. and Deng, L. A bayesian approach to the verification problem: Applications to speaker verification. *IEEE Transactions on Speech and Audio Processing*, 9(8):874–884, 2001.

- Kenny, P., Stafylakis, T., Ouellet, P., Alam, M. J., and Dumouchel, P. PLDA for speaker verification with utterances of arbitrary duration. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7649–7653. IEEE, 2013.
- Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., and Khudanpur, S. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5220–5224. IEEE, 2017.
- Kulis, B., Sustik, M., and Dhillon, I. Learning low-rank kernel matrices. In *Proceedings of the 23rd international conference on Machine learning*, pp. 505–512. ACM, 2006.
- Lanckriet, G. R., Cristianini, N., Bartlett, P., Ghaoui, L. E., and Jordan, M. I. Learning the kernel matrix with semidefinite programming. *Journal of Machine learning research*, 5(Jan):27–72, 2004.
- Li, C., Ma, X., Jiang, B., Li, X., Zhang, X., Liu, X., Cao, Y., Kannan, A., and Zhu, Z. Deep speaker: an end-to-end neural speaker embedding system. *arXiv preprint arXiv:1705.02304*, 2017.
- Monteiro, J., Alam, J., and Falk, T. H. Combining speaker recognition and metric learning for speaker-dependent representation learning. *Proc. Interspeech 2019*, pp. 4015–4019, 2019.
- Nagrani, A., Chung, J. S., and Zisserman, A. Voxceleb: A large-scale speaker identification dataset. *Proc. Interspeech 2017*, pp. 2616–2620, 2017.
- Neyman, J. and Pearson, E. S. IX. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.
- Nickel, M. and Kiela, D. Poincaré embeddings for learning hierarchical representations. In *Advances in neural information processing systems*, pp. 6338–6347, 2017.
- Nickel, M. and Kiela, D. Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In *International Conference on Machine Learning*, pp. 3776–3785, 2018.
- Norouzi, M., Fleet, D. J., and Salakhutdinov, R. R. Hamming distance metric learning. In *Advances in neural information processing systems*, pp. 1061–1069, 2012.
- Nowozin, S., Cseke, B., and Tomioka, R. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pp. 271–279, 2016.
- Oh Song, H., Xiang, Y., Jegelka, S., and Savarese, S. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4004–4012, 2016.
- Okabe, K., Koshinaka, T., and Shinoda, K. Attentive statistics pooling for deep speaker embedding. *Proc. Interspeech 2018*, pp. 2252–2256, 2018.
- Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- Pitis, S., Chan, H., Jamali, K., and Ba, J. An inductive bias for distances: Neural nets that respect the triangle inequality. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=HJeIDpVFPr>.
- Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society, 2011.
- Prince, S. J. and Elder, J. H. Probabilistic linear discriminant analysis for inferences about identity. In *IEEE 11th International Conference on Computer Vision (ICCV)*, pp. 1–8, 2007.
- Ravanelli, M. and Bengio, Y. Learning speaker representations with mutual information. *Proc. Interspeech 2019*, pp. 1153–1157, 2019.
- Reynolds, D. A. An overview of automatic speaker recognition technology. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pp. IV–4072. IEEE, 2002.
- Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1-3):19–41, 2000.
- Schroff, F., Kalenichenko, D., and Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.
- Schultz, M. and Joachims, T. Learning a distance metric from relative comparisons. In *Advances in neural information processing systems*, pp. 41–48, 2004.
- Shalev-Shwartz, S., Singer, Y., and Ng, A. Y. Online and batch learning of pseudo-metrics. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 94. ACM, 2004.

- Shi, H., Yang, Y., Zhu, X., Liao, S., Lei, Z., Zheng, W., and Li, S. Z. Embedding deep metric for person re-identification: A study against large variations. In *European conference on computer vision*, pp. 732–748. Springer, 2016.
- Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pp. 4077–4087, 2017.
- Snyder, D., Chen, G., and Povey, D. MUSAN: A Music, Speech, and Noise Corpus, 2015. arXiv:1510.08484v1.
- Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D., and Khudanpur, S. Spoken language recognition using x-vectors. In *Odyssey*, pp. 105–111, 2018a.
- Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. X-vectors: Robust DNN embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333. IEEE, 2018b.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.
- Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.
- Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. J. Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing*, 37(3):328–339, 1989.
- Wang, F., Cheng, J., Liu, W., and Liu, H. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018.
- Weinberger, K. Q. and Saul, L. K. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009.
- Wu, C.-Y., Manmatha, R., Smola, A. J., and Krahenbuhl, P. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2840–2848, 2017.
- Xiang, X., Wang, S., Huang, H., Qian, Y., and Yu, K. Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition. *arXiv preprint arXiv:1906.07317*, 2019.
- Xie, W., Nagrani, A., Chung, J. S., and Zisserman, A. Utterance-level aggregation for speaker recognition in the wild. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5791–5795. IEEE, 2019.
- Xing, E. P., Jordan, M. I., Russell, S. J., and Ng, A. Y. Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, pp. 521–528, 2003.
- Ying, Y. and Li, P. Distance metric learning with eigenvalue optimization. *Journal of machine Learning research*, 13 (Jan):1–26, 2012.
- Zhang, C., Koishida, K., and Hansen, J. H. Text-independent speaker verification based on triplet convolutional neural network embeddings. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 26(9):1633–1644, 2018.
- Zhu, H., Long, M., Wang, J., and Cao, Y. Deep hashing network for efficient similarity retrieval. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

A. Implementation details

A.1. \mathcal{D} architecture

\mathcal{D} is implemented as a stack of fully-connected layers with LeakyReLU activations. Dropout is further used in between the last hidden and the output layer. The number and size of hidden layers as well as the dropout probability were tuned for each experiment.

A.2. Cifar-10 and Mini-ImageNet

A.2.1. HYPERPARAMETERS

The grid used on the hyperparameter search for each hyperparameter is presented next. A budget of 100 runs was considered and each model was trained for 600 epochs. Hyperparameters yielding the best EER on the validation data for our proposed approach and the triplet baseline are represented by * and †, respectively. In all experiments, the minibatch size was set to 64 and 128 for Cifar-10 and Mini-ImageNet, respectively. A reduce-on-plateau schedule for the learning rate was employed, while its patience was a further hyperparameter included in the search.

Cifar-10:

- Learning rate: $\{0.5, 0.1, 0.01^{*\dagger}, 0.001\}$
- Weight decay: $\{0.01, 0.001^*, 0.0001^\dagger, 0.00001\}$
- Momentum: $\{0.1, 0.5, 0.9^{*\dagger}\}$
- Label smoothing: $\{0.01, 0.1, 0.2^{*\dagger}\}$
- Patience: $\{1, 5, 10^{*\dagger}, 20\}$
- Number of \mathcal{D} hidden layers: $\{2, 3^*, 4, 5\}$
- Size of \mathcal{D} hidden layers: $\{128, 256, 350^*, 512\}$
- \mathcal{D} dropout probability: $\{0.01, 0.1, 0.2^*\}$
- Type of auxiliary loss: $\{\text{Standard cross-entropy, Additive margin}^{*\dagger}\}$

Mini-Imagenet:

- Learning rate: $\{0.5, 0.1^\dagger, 0.01^*, 0.001\}$
- Weight decay: $\{0.01, 0.001^*, 0.0001^\dagger, 0.00001\}$
- Momentum: $\{0.1, 0.5, 0.9^{*\dagger}\}$
- Label smoothing: $\{0.01, 0.1^*, 0.2^\dagger\}$
- Patience: $\{1, 5, 10^{*\dagger}, 20\}$
- Number of \mathcal{D} hidden layers: $\{2, 3^*, 4, 5\}$
- Size of \mathcal{D} hidden layers: $\{128, 256, 350^*, 512\}$
- \mathcal{D} dropout probability: $\{0.01, 0.1^*, 0.2\}$
- Type of auxiliary loss: $\{\text{Standard cross-entropy, Additive margin}^{*\dagger}\}$

A.3. Voxceleb

A.3.1. ENCODER ARCHITECTURE

We implement \mathcal{E} as the well-known TDNN architecture employed within the x-vector setting (Snyder et al., 2018b), which consists of a sequence of dilated 1-dimensional convolutions across the temporal dimension, followed by a time pooling layer, which simply concatenates element-wise first- and second-order statistics over time. Concatenated statistics are finally projected into an output vector through two fully-connected layers. Pre-activation batch normalization is performed after each convolution and fully-connected layer. A summary of the employed architecture is shown in Table 3.

Table 3. Employed TDNN encoder. T indicates the duration of features in number of frames.

<i>Layer</i>	<i>Input Dimension</i>	<i>Output dimension</i>
<i>Conv1d+ReLU</i>	$30 \times T$	$512 \times T$
<i>Conv1d+ReLU</i>	$512 \times T$	$512 \times T$
<i>Conv1d+ReLU</i>	$512 \times T$	$512 \times T$
<i>Conv1d+ReLU</i>	$512 \times T$	$512 \times T$
<i>Conv1d+ReLU</i>	$512 \times T$	$1500 \times T$
<i>Statistical Pooling</i>	$1500 \times T$	3000
<i>Linear+ReLU</i>	$3000 \times T$	512
<i>Linear+ReLU</i>	512	d

A.3.2. DATA AUGMENTATION AND FEATURE EXTRACTION

We augment the training data by simulating diverse acoustic conditions using supplementary noisy speech, as done in (Snyder et al., 2018b). More specifically, we corrupt the original samples by adding reverberation (reverberation time varies from 0.25s - 0.75s) and background noise such as music (signal-to-noise ratio, SNR, within 5-15dB), and babble (SNR varies from 10dB to 20dB). Noise signals were selected from the MUSAN corpus (Snyder et al., 2015) and the room impulse responses samples from (Ko et al., 2017) were used to simulate reverberation. All the audio pre-processing steps including feature extraction, degradation with noise as well as silence frames removal was performed with the Kaldi toolkit (Povey et al., 2011) and are openly available as the first step of the recipe in <https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb>. The corpora used for augmentation are also openly available at <https://www.openslr.org/>.

In order to deal with recordings of varying duration within a minibatch, we pad all recordings to a maximum duration set in advance. We do so by repeating the signal up until it reaches the maximum duration or taking a random continuous chunk with the maximum duration for the case of long utterances.

A.3.3. MINIBATCH CONSTRUCTION

Given the large number of classes in the VoxCeleb case (corresponding to the number of speakers, i.e., 5994), we need to ensure several examples belonging to the same speaker exist in a minibatch to allow for positive pairs to exist. We thus create a list of sets of five recordings belonging to the same speaker, and such sets are randomly selected at training time. Minibatches are constructed through sequentially picking examples from the list, and the list is recreated once all elements are sampled. Such approach provides minibatches of size $N_e = S \cdot R$, where R and S correspond to the number of speakers per minibatch and recordings per speaker, respectively. While R is set to 5, S is set to 24, which gives an effective minibatch size of $N_e = 120$.

A.3.4. HYPERPARAMETERS

Training was carried out with a linear learning rate warm-up, employed during the first iterations, and the same exponential decay as in (Vaswani et al., 2017) is employed after that. A budget of 40 runs was considered and each model was trained for a budget of 600k iterations. The best set of hyperparameters, as assessed in terms of EER measured over a random set of trials created from **VoxCeleb1-E**, was then used to train a model from scratch for a total of 2M iterations. We report the results obtained by the best model within the 2M iterations in terms of the same metric used during the hyperparameter search. Selected values are indicated by *.

An end-to-end approach for the verification problem: learning the right distance

The grid used for the hyperparameter search is presented next. In all experiments, the minibatch size was set to 24, which, given the sampling strategy employed in this case, yields an effective batch size of 120. We further employed gradient clipping and searched over possible clipping thresholds.

- Base learning rate: {2.0, 1.5*, 1.0, 0.5, 0.1}
- Weight decay: {0.001*, 0.0001, 0.00001}
- Momentum: {0.7, 0.85, 0.95*}
- Label smoothing: {0.0, 0.1*, 0.2}
- Embedding size d : {128, 256*, 512}
- Maximum duration (in number of frames): {300, 500, 800*}
- Gradient clipping threshold: {10*, 20, 50}
- Number of \mathcal{D} hidden layers: {1, 2, 3, 4*}
- Size of \mathcal{D} hidden layers: {128, 256*, 350, 512}
- \mathcal{D} dropout probability: {0.01, 0.1*, 0.2}
- Type of auxiliary loss: {Standard cross-entropy, Additive margin*}

B. Large scale speaker verification under domain shift

In this experiment, we evaluate the performance of the proposed setting when test data significantly differs from training examples. To do so, we employ the data introduced for one of the tasks of the 2018 edition of the NIST Speaker Recognition Evaluation (SRE)⁷. We specifically consider the CTS task so that test data corresponds to spontaneous conversational telephone speech spoken in Tunisian Arabic, while the bulk of the train data is spoken in English. Besides the language mismatch, variations due to different codecs are further observed (PSTN vs. PSTN and VOIP).

The main training dataset (English) is built by combining the data from *NIST SREs* from 2004 to 2010, *Mixer 6*, as well as *Switchboard-2*, phases 1, 2, and 3, and the first release of *VoxCeleb*, yielding a total of approximately 14000 speakers. Audio representations correspond to 23 MFCCs obtained using a short-time Fourier transform with a 25ms Hamming window and 60% overlap. The audio data is downsampled to 8kHz. Further pre-processing steps are the same as those performed for experiments with *VoxCeleb* as reported in Section 4, i.e. an energy-based voice activity detector is followed by data augmentation performed via distorting original samples adding reverberation and background noise.

Baseline: For performance reference, we trained the well-known x-vector setting (Snyder et al., 2018b) using its Kaldi recipe⁸. In that case, PLDA is employed for scoring test trials. The same training data used to train our systems is employed in this case as well. The recipe performs the following steps: **i**-training of a TDNN (same architecture as in our case) as a multi-class classifier over the set of training speakers using the same training data utilized to train our proposed model; **ii**-preparation of PLDA’s training data, in which case the *SRE* partition of the training set is encoded using the second to last layer of the TDNN, embeddings are length-normalized and mean-centered using the average of an unlabelled sample from the target domain and finally have their dimensionality reduced using Linear Discriminant Analysis; **iii**-training of PLDA; **iv**-scoring of test trials. In addition to that, in order to cope with the described domain shift, the model adaptation scheme introduced in (Garcia-Romero et al., 2014) is also utilized for PLDA so that a second PLDA model is trained on top of target data. The final downstream classifier is then obtained by averaging the parameters of the original and target domain models. Both results obtained with and without the described scheme are reported in Table 4.

Table 4. Evaluation of models under domain shift. Target data corresponds to speech spoken in Arabic. Fine-tuning on datasets including target data yields an improvement in verification performance.

	<i>Training domain</i>	<i>Scoring</i>	<i>EER</i>
Snyder et al. (2018b)	English	PLDA	11.30%
	English+Arabic	Adapted PLDA	9.44%
Proposed	English	E2E	13.61%
	Multi-language	E2E	8.43%

For the case of the proposed approach, training is carried out using the training data described above corresponding to speech spoken in English. We reuse the setting found to work well on the experiments reported in Section 4 with the *VoxCeleb* corpus including all hyperparameters, architecture, data sampling and minibatch construction strategies, and computational budget. We additionally build a multi-language training set including data corresponding to the target domain so that we can fine-tune our model. The complementary training data corresponds to the data introduced for the 2012 (English) and 2016 (Cantonese+Tagalog) editions of NIST SRE as well as the development partition of NIST SRE 2018 which corresponds to the target domain of evaluation data (Arabic). This is done so as to increase the amount of data within the complementary partition and avoid overfitting to the small amount of target data. The combination of such data sources yields approximately 800 speakers. We thus train our models on the large out-of-domain dataset and fine-tune the resulting model in the multi-language complementary data.

Results in terms of equal error rate are presented in Table 4. While our model appears to be more domain dependent when compared to PLDA as indicated by results where only out-of-domain data is employed, it significantly improves once a relatively small amount of target domain data is provided. We stress the fact that the proposed setting dramatically simplifies verification pipelines and completely removes practical issues such as those related to processing steps prior to training of the downstream classifier.

⁷https://www.nist.gov/system/files/documents/2018/08/17/sre18_eval_plan_2018-05-31_v6.pdf

⁸<https://github.com/kaldi-asr/kaldi/tree/master/egs/sre16/v1/local/nnet3/xvector>

C. Image retrieval and clustering

Even though the setting we introduced is tailored to verification and we only have guarantees for that case, we further verified its performance on other tasks and once again concluded it reaches competitive performance while using a much simpler and general training/testing workflow. Such extra evaluation is performed using well-known image retrieval benchmarks: Caltech birds⁹ (CUB), CARS196¹⁰, and Stanford online products¹¹ (SOP).

We follow the experimental setting in past work and fine-tune pretrained models on ImageNet in each of the three datasets. The pretrained models thus correspond to the 5-layered case reported in Fig. 4. Fine-tuning is performed in each dataset with the same strategy to sample examples to form minibatches as reported in ImageNet experiments, while the learning rate schedule matches that of (Vaswani et al., 2017). Results in terms of Recall@k (Oh Song et al., 2016) for increasing k are presented in Figures 5-a, 5-b, and 5-c for the cases of CUB, CARS, and SOP, respectively, while clustering performance is reported in Table 5. We compare our models against results reported by Wu et al. (2017) corresponding to several metric learning schemes employed for retrieval. We thus indicate by *REF.* - and *REF.* + the worst and best performances they report for each metric/dataset. We further report the performance of the models trained only on ImageNet as well as an ablation case in which the auxiliary loss is dropped. In most cases our models performance lies in between *REF.* - and *REF.* +, i.e. a competitive performance with respect to settings heavily engineered for each case is obtained with our models where a much simpler training/inference workflow is used.

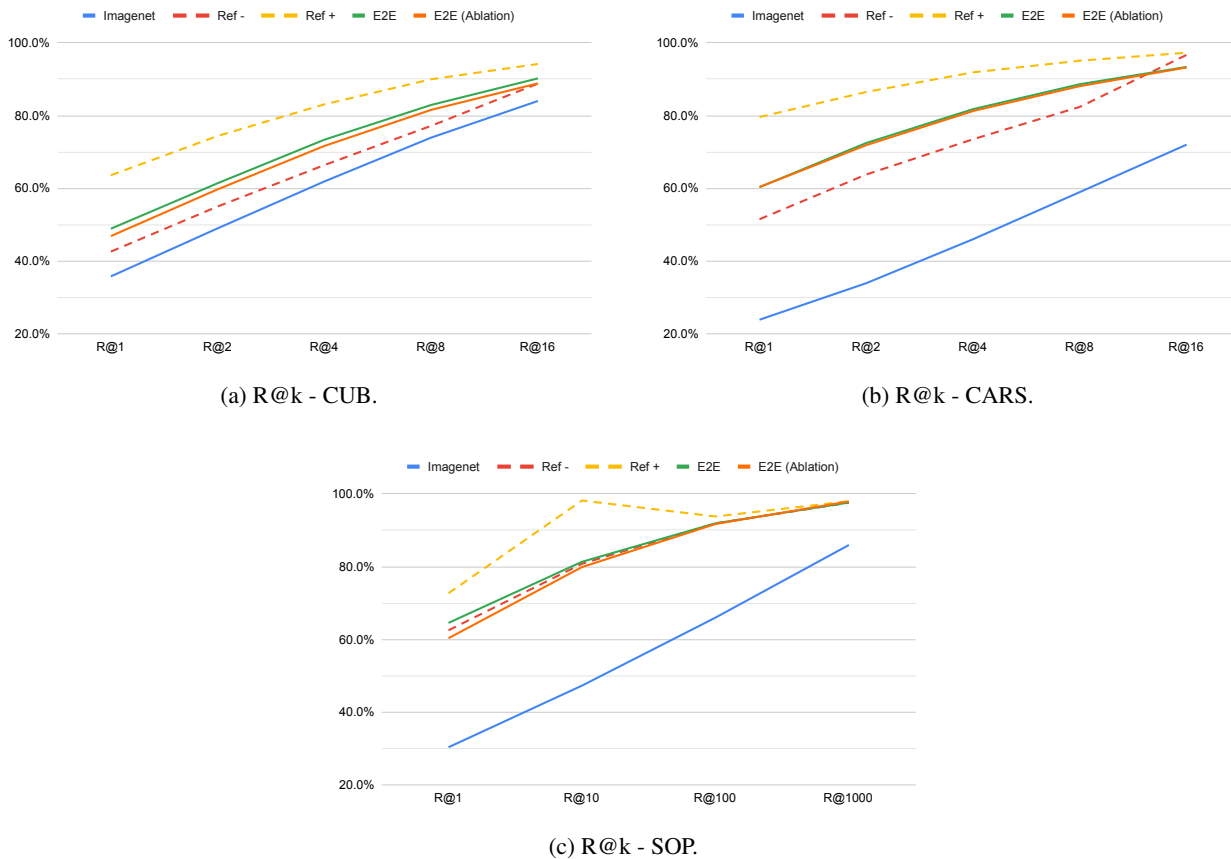


Figure 5. Evaluation on image retrieval.

Clustering results reported in Table 5 correspond to the normalized mutual information (NMI) measured between class labels and cluster assignments obtained by the k-means algorithm executed over the representations given by \mathcal{E} . For our

⁹<http://www.vision.caltech.edu/visipedia/CUB-200-2011.html>

¹⁰https://ai.stanford.edu/~jkrause/cars/car_dataset.html

¹¹https://cvgl.stanford.edu/projects/lifted_struct/

systems, we report a further result in which case a heuristic approach is used to enable the use of \mathcal{D} to assign clusters to each data point. Using said approach, we assign each data point to the cluster corresponding to the Euclidean centroid corresponding to the smallest distance given by \mathcal{D} .

Table 5. Clustering performance in terms of NMI. Results in parenthesis indicate the clustering performance obtained by using \mathcal{D} to assign clusters to test examples.

	CUB	CARS	SOP
<i>ImageNet</i>	52.5% (53.9%)	52.5% (35.5%)	81.6% (79.9%)
<i>Ref. - (Wu et al., 2017)</i>	55.4%	53.4%	88.10%
<i>Ref. + (Wu et al., 2017)</i>	69.0%	69.1%	90.70%
<i>E2E</i>	60.5% (64.1%)	59.9% (63.5%)	89.2% (92.9%)
<i>E2E (Ablation)</i>	58.5% (62.4%)	60.4% (63.3%)	88.1% (92.7%)