f-Similarity Preservation Loss for Soft Labels: A Demonstration on Cross-Corpus Speech Emotion Recognition

Biqiao Zhang *
University of Michigan
2260 Hayward St
Ann Arbor, MI, USA 48109
didizbq@umich.edu

Yuqing Kong *
Peking University
5 Yiheyuan Road
Beijing, China 100871
yuqing.kong@pku.edu.cn

Georg Essl
University of Wisconsin - Milwaukee
2442 E Hartford Ave
Milwaukee, WI, USA 53211
essl@uwm.edu

Emily Mower Provost
University of Michigan
2260 Hayward St
Ann Arbor, MI, USA 48109
emilykmp@umich.edu

Abstract

In this paper, we propose a Deep Metric Learning (DML) approach that supports soft labels. DML seeks to learn representations that encode the similarity between examples through deep neural networks. DML generally presupposes that data can be divided into discrete classes using hard labels. However, some tasks, such as our exemplary domain of speech emotion recognition (SER), work with inherently subjective data, data for which it may not be possible to identify a single hard label. We propose a family of loss functions, f-Similarity Preservation Loss (f-SPL), based on the dual form of f-divergence for DML with soft labels. We show that the minimizer of f-SPL preserves the pairwise label similarities in the learned feature embeddings. We demonstrate the efficacy of the proposed loss function on the task of cross-corpus SER with soft labels. Our approach, which combines f-SPL and classification loss, significantly outperforms a baseline SER system with the same structure but trained with only classification loss in most experiments. We show that the presented techniques are more robust to over-training and can learn an embedding space in which the similarity between examples is meaningful.

1 Introduction

Deep metric learning (DML) aims to use deep neural networks (DNN) to project input data to a learned space, in which the similarity between examples can be directly measured (Lu, Hu, and Zhou 2017). DML has been successfully applied to many visual understanding tasks, such as face verification, image classification, and person re-identification (Taigman et al. 2014; Schroff, Kalenichenko, and Philbin 2015; Hoffer and Ailon 2015; Yi et al. 2014). These tasks often rely on hard class labels to determine the pairwise relationship between data. Yet, soft labels may be preferable to hard labels in some cases: they provide more information for each training example (Hinton, Vinyals, and Dean 2015) and are more robust against label noise (Thiel 2008).

However, the additional information contained in soft labels is not fully exploited in traditional DML approaches. Motivated by this, we propose a family of loss functions, the f-Similarity Preservation Loss (f-SPL), based on the dual form of f-divergence. f-SPL is designed for DML with soft labels, here defined as real-valued labels that are distributed along one or multiple dimensions. f-SPL aims to preserve the label similarities in the learned feature space and can be applied to tasks that require either continuous or discrete (e.g., a class index) test output. Further, we introduce a pair sampling method for the efficient implementation of f-SPL in neural networks.

We evaluate our methods on cross-corpus speech emotion recognition (SER). SER aims to automatically identify human emotions from speech. SER is complicated by the co-modulations that are present in the speech signals (e.g., lexical information and speaker identity). Networks may be unintentionally over-trained to capture signals that are specific to certain speakers or lexical artifacts in the data, resulting in poor generalizability and poor robustness in cross-corpus tasks. DML can be used to generate an embedding space in which distances between examples correspond to the label relationships. This provides a mechanism to reduce the influence of factors other than emotion. However, conventional DML based on hard labels may not be ideal for SER. The labels used in SER systems are usually collected through perceptual experiments. The variability in emotion expression and the subjectivity of emotion perception lead to datasets with uncertain labels. Previous work in SER has demonstrated the efficacy of using soft labels given uncertainty (Steidl et al. 2005; Mower, Mataric, and Narayanan 2011; Fayek, Lech, and Cavedon 2016).

We form the problem as binary classification of soft-labeled valence (positive vs. negative) and activation (calm vs. excited) (Russell 1980). We combine the proposed loss with classification loss in the training of DNN classifiers. Our baseline is the same classifier trained with classification loss only. The results show that our multi-task framework with the added f-SPL statistically significantly increases system performance in the majority of the experiments and is more robust to over-training than the baseline system.

2 Related Works

2.1 Deep Metric Learning

Deep metric learning approaches predominantly focus on hard labels (Lu, Hu, and Zhou 2017). These approaches of-

*Biqiao Zhang and Yuqing Kong contributed equally to this work.

ten rely on loss functions that aim to pull data from the same class closer while pushing data from different classes farther apart. Some works use contrastive loss for pairs of examples through Siamese networks (Bromley et al. 1994; Chopra, Hadsell, and LeCun 2005), identifying "positive pairs" of examples from the same class and "negative pairs" of examples from different classes. This loss then aims to learn a space where the distance between a positive pair is less than a margin τ_+ while the distance between a negative pair is larger than a margin τ_- , where $0 \leq \tau_+ < \tau_-$.

Some works have proposed loss calculations over triplets, defined as sets of three examples: an anchor, a positive example from the anchor's class, and a negative example from a different class (Weinberger and Saul 2009). DNNs with triplet loss (Hoffer and Ailon 2015; Schroff, Kalenichenko, and Philbin 2015) aim to learn an embedding space where the distance between the anchor and the positive example is at least smaller than the distance between that anchor and the negative example by a margin τ.

Some works have extended the triplet loss, by considering all positive and negative pairs within a batch (Song et al. 2016), using multiple negative examples in each set (Sohn 2016), or using the cluster center rather than a single example as the anchor (Liu et al. 2016; 2017). Yang et al. (2018) proposed a loss function designed for image sentiment analysis, based on the relationships between neighboring sentiment classes on the Mikels' emotion wheel (Mikels et al. 2005). They added "related" examples, defined as examples from a different class than the anchor but on the same half of the emotion wheel, to triplets. Denoting the distance between anchor and the positive example as anchor-positive, their approach aimed to find a space where anchor-positive is at least smaller than anchor-related by τ_1, and anchor-related is at least smaller than anchor-negative by τ_2. The distance is scaled by class similarity, implemented using a factor proportional to the class distance on the emotion wheel.

Two recent works have used DML for regression. Wang, Wan, and Yuan (2017) combined metric learning for kernel regression with DNN for crowdedness regression. Doumanoglou et al. (2016) proposed a loss function via Siamese network for pose estimation. They compared the distance between labels (d_l) and embeddings (d_f) given pairs of data. Their approach aim to minimize the combination of $d_f - d_l$ and regression loss. However, we note that $d_f - d_l$ is not guaranteed to be non-negative in the loss.

In this work, we propose a family of loss functions for DML with real-valued labels and provide theoretical justifications. We experiment on classification tasks with soft labels. However, the application of the loss functions could also be extended to other tasks, including regression.

2.2 Speech Emotion Recognition with Soft Labels

Emotion expression is subtle, and emotion perception is subjective. This leads to inter-rater variability and uncertainty in emotion labels. One way to take this variability and uncertainty into account is to avoid single hard labels. For example, researchers have represented emotion information using probability distributions over emotion classes (Aldeneh et al. 2017), confidence scores that capture the presence or absence of multiple emotion classes (Mower, Mataric, and Narayanan 2011), or by estimating distributions over evaluator perception (Zhang, Essl, and Mower Provost 2017).

Researchers have investigated the efficacy of using soft labels while training SER systems. They found that training with soft labels increases system performance in terms of standard classification measures (Fayek, Lech, and Cavedon 2016) or an entropy-based measure that takes human confusion into account (Steidl et al. 2005). Lotfian and Busso (2017) proposed considering emotion perception of an utterance as a multidimensional Gaussian distribution over emotion classes. They showed that systems trained using soft labels, calculated by taking the mean of the estimated Gaussian distribution, outperformed systems trained using hard labels. Collectively, the complexity of the task and the efficacy of soft labeling makes SER an ideal task for demonstrating the impact of deep metric learning with soft labels.

2.3 f-Divergence

f-divergence is a family of non-symmetric measures of difference between two distributions, based on the family of convex functions f (Ali and Silvey 1966). These measures are widely used in the learning literature. Common members of the f-divergence family include Kullback-Leibler (KL) divergence and total variation distance. Nguyen, Wainwright, and Jordan (2009) proposed a duality technique of f-divergence, which plays a key role in mutual information estimation (Nguyen, Wainwright, and Jordan 2010), the design of a type of generative adversarial networks, f-GANs (Nowozin, Cseke, and Tomioka 2016), and the design of information elicitation mechanisms and co-training algorithms (Kong and Schoenebeck 2018). Motivated by these works, we use the dual formulation of f-divergence to derive our f-Similarity Preservation Loss.

3 f-Similarity Preservation Loss (f-SPL)

Our goal is to learn an embedding space on which the similarity between examples equals to the label similarity. In Section 3.1, we define a family of loss functions, f-SPL, based on the dual form of f-divergence. Then in Section 3.2, we mathematically prove that we can achieve our goal by minimizing f-SPL. Finally in Section 3.3, we explain how f-SPL can be implemented in a multi-task framework.

3.1 Definition of f-SPL

We denote data and soft labels as $x_1, x_2, ... \in A_X$ and $y_1, y_2, ... \in A_Y$, respectively. The function $C : A_Y \times A_Y \mapsto [0, 2]$ measures label similarity. A feature learning function (i.e., a neural network) $g \in G$, maps inputs from A_X to a new space A_G and $S : A_G \times A_G \mapsto [0, 2]$ measures the similarity on A_G. We seek to find a function, $F(S(g), C)$, that optimizes over g. The optimal solution of F, g^*, satisfies $S(g^*(x_i), g^*(x_j)) = C(y_i, y_j)$ for every $i \neq j$, i.e., the similarity between the examples on the learned space is the same as the similarity between their labels.

We use the dual form of f-divergence to construct $F(S(g), C)$. We name the resulting functions f-Similarity Preservation Gain (f-SPG). We then modify f-SPG to a

f-divergence	$f(t)$	f-SPG(S, C)	f-SPL(S, C)
KL divergence	$t \log t$	$C * (1 + \log S) - S$	$S - C \log(S) - C + C \log(C)$
Reverse KL	$-\log t$	$C * (-\frac{1}{S}) - (\log S - 1)$	$\log(S) + \frac{C}{S} - \log(C) - 1$
Pearson χ^2	$(t-1)^2$	$C * 2(S-1) - (S^2 - 1)$	$(C - S)^2$
Squared Hellinger	$(\sqrt{t} - 1)^2$	$C * (1 - \sqrt{\frac{1}{S}}) - (\sqrt{S} - 1)$	$\sqrt{S} + C\sqrt{\frac{1}{S}} - 2\sqrt{C}$
Jensen-Shannon (JS) Divergence	$-(t+1) \log \frac{t+1}{2} + t \log t$	$C * \log \frac{2S}{1+S} + log(\frac{2}{1+S})$	$(C+1) \log(1+S) - C \log(2S)$ $-(C+1) \log(1+C) + C \log(2C)$

Table 1: Reference for common f-divergences, their corresponding convex functions f (Nowozin, Cseke, and Tomioka 2016), f-SPG(S, C), and f-SPL(S, C).

family of loss functions, f-SPL, such that: (1) f-SPL is always non-negative and (2) maximizing f-SPG is equivalent to minimizing f-SPL.

f-SPG Given a convex function f, a feature learning function $g \in G$, and a pair of examples $p = [(x, y), (x', y')]$, we define f-SPG based on the dual formulation of f-divergence (Section 3.2, Lemma 2) as:

$$\begin{aligned} f\text{-SPG}(p; g) :=& f\text{-SPG}(S_p(g), C_p) \\ :=& C_p * \partial f(S_p(g)) - f^\star(\partial f(S_p(g))), \end{aligned}$$

where $C_p := C(y, y')$, $S_p(g) := S(g(x), g(x'))$, ∂f is the subdifferential of f, and $f^\star$ is the convex conjugate of f (formally defined in Section 3.2).

Given a set of pairs $I = \{[(x, y), (x', y')], ...\}$, we define the total f-SPG as the sum of the individual f-SPG:

$$f\text{-SPG}(I; g) := \sum_{p \in I} f\text{-SPG}(p; g).$$

Fixing the set I, we seek g that maximizes f-SPG$(I; g)$. When the convex function f is differentiable and ∂f is invertible, and the set, I, satisfies a *balance condition*,

$$\sum_{p \in I} (C_p - 1) = 0,$$

our main theorem (Theorem 3) in Section 3.2 will show that: (1) the maximizer of f-SPG, g^*, preserves the pairwise similarity, that is, for every $p = [(x, y), (x', y')] \in I$, $C(y, y') = S(g^*(x), g^*(x'))$; (2) the maximum of f-SPG represents the amount of information contained in the pairs.

f-SPL We convert f-SPG to a loss function, f-SPL, so that it can be used as a component of neural network training. To do this, we identify the maximal point of f-SPG at which the label similarity is equal to the feature similarity, f-SPG(C_p, C_p), and subtract from it f-SPG$(S_p(g), C_p)$:

$$\begin{aligned} &f\text{-SPL}(p; g) := f\text{-SPG}(C_p, C_p) - f\text{-SPG}(S_p(g), C_p) \\ &\text{and} f\text{-SPL}(I; g) := \sum_{p \in I} f\text{-SPL}(p; g). \end{aligned}$$

As a result, f-SPL has the following properties: (1) f-SPL is always non-negative; (2) minimizing f-SPL$(I; g)$ over g is equivalent to maximizing f-SPG$(I; g)$ over g. Table 1 shows five special cases of f-SPL based on the convex functions corresponding to common f-divergence measures.

3.2 Theoretical Justifications

We will show the feature learning function, g^*, that minimizes f-SPL$(I; g)$ to zero and maximizes f-SPG$(I; g)$ to the amount of information contained in the set I, also preserves the pairwise similarity of I.

To give the theoretical justification, we first give the formal definition of f-divergence and its dual form.

f-divergence (Ali and Silvey 1966; Csiszár, Shields, and others 2004) Given set Σ and the set of all possible distributions over Σ, Δ_Σ, f-divergence $D_f : \Delta_\Sigma \times \Delta_\Sigma \mapsto \mathbb{R}$ is a non-symmetric measure of the difference between two distributions, $\mathbf{p}, \mathbf{q} \in \Delta_\Sigma$, and is defined as

$$D_f(\mathbf{p}, \mathbf{q}) = \sum_{\sigma \in \Sigma} \mathbf{p}(\sigma) f\left(\frac{\mathbf{q}(\sigma)}{\mathbf{p}(\sigma)}\right)$$

where $f : \mathbb{R} \mapsto \mathbb{R}$ is a convex function and $f(1) = 0$.

Definition 1 (Fenchel Duality (Rockafellar and others 1966))**.** *Given any function $f : \mathbb{R} \mapsto \mathbb{R}$, we define its convex conjugate $f^\star$ as a function that also maps $\mathbb{R}$ to $\mathbb{R}$ such that*

$$f^\star(x) = \sup_t tx - f(t).$$

Lemma 2 (Dual form of f-divergence (Nguyen, Wainwright, and Jordan 2009; 2010))**.**

$$\begin{aligned} D_f(\mathbf{p}, \mathbf{q}) &\geq \sup_{u \in \Sigma} \mathbb{E}_{\mathbf{p}} u - \mathbb{E}_{\mathbf{q}} f^\star(u) \\ &= \sup_{u \in \mathcal{G}} \sum_\sigma u(\sigma)\mathbf{p}(\sigma) - \sum_\sigma f^\star(u(\sigma))\mathbf{q}(\sigma) \end{aligned}$$

where $\mathcal{G}$ is a set of functions that map Σ to $\mathbb{R}$. The equality holds if and only if $u(\sigma) = u^*(\sigma) \in \partial f\left(\frac{\mathbf{p}(\sigma)}{\mathbf{q}(\sigma)}\right)$, i.e., the subdifferential of f on value $\frac{\mathbf{p}(\sigma)}{\mathbf{q}(\sigma)}$.

We define $\mathbf{C}$ and $\mathbf{1}$ as distributions over the pairs in I such that $\mathbf{C}(p) = \frac{C_p}{\sum_{p \in I} C_p}$ and $\mathbf{1}(p) = \frac{1}{|I|}$ for all $p \in I$.

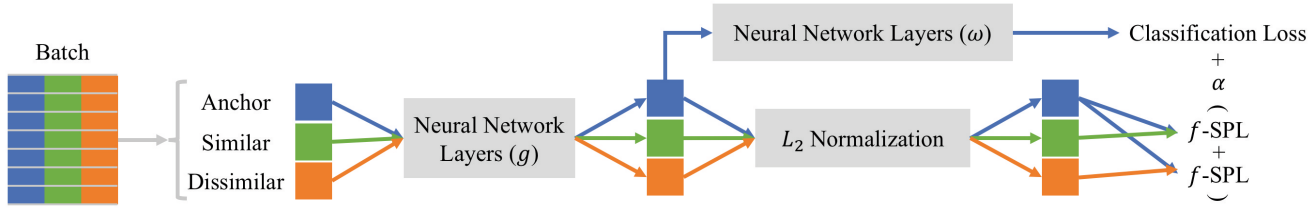

Figure 1: The proposed multi-task framework for training. The inputs are batches of triplets. The loss is the combination of the classification loss calculated on the anchor a only and the f-SPL between a and a similar example s plus a and a dissimilar example d (s.t., $C(y_a, y_s) + C(y_a, y_d) \approx 2$). α is the weighting term for f-SPL. The test phase does not depend on triplets.

$D_f(\mathbf{C}, \mathbf{1})$ measures the amount of information contained in the chosen pairs: when all chosen pairs are neither very similar nor very dissimilar, $\mathbf{C}$ is close to $\mathbf{1}$, which implies a small amount of information contained in the pairs; when all chosen pairs are either very similar or very dissimilar, $\mathbf{C}$ is far away from $\mathbf{1}$, which implies a large amount of information contained in the pairs.

We use the above lemma to show our main theorem:

Theorem 3. *Given a convex function f, a balanced set I, when f is differentiable and ∂f is invertible, for every minimizer g^* of f-SPL$(I; g)$, for every $[(x, x'), (y, y')] \in I$,*

$$S(g^*(x), g^*(x')) = C(y, y').$$

g^* minimizes f-SPL$(I; g)$ to zero and maximizes f-SPG$(I; g)$ to $D_f(\mathbf{C}, \mathbf{1})$.

Proof. The balance condition implies that

$$\sum_{p \in I} C_p = |I|$$

Thus, by dividing $|I|$, we can rewrite f-SPG$(I; g)$ as

$$\sum_{p \in I} \partial f(S_p(g))) * \mathbf{C}(p) - f^\star(\partial f(S_p(g))) * \mathbf{1}(p).$$

Based on Lemma 2, for every maximizer g^* of f-SPG$(I; g)$/minimizer of f-SPL$(I; g)$, we have

$$\partial f(S_p(g^*)) = \partial f\left(\frac{\mathbf{C}(p)}{\mathbf{1}(p)}\right) = \partial f(C_p)$$

for every $p \in I$ and the maximum of f-SPG is $D_f(\mathbf{C}, \mathbf{1})$, which also implies the minimum of f-SPL is zero.

Therefore, when f is differentiable and ∂f is invertible, g^* preserves the pairwise similarity of the pairs in I. □

3.3 Multi-Task Framework

In this work, we use a multi-task framework that jointly reduces classification loss and f-SPL, as shown in Figure 1. The first block of neural network layers corresponds to g and the second block of layers is denoted as ω. Previous work has demonstrated the efficacy of using DML loss with hard labels within multi-task frameworks (Liu et al. 2017; Yang et al. 2018). We hypothesize that DML loss will also enhance classification performance given soft labels. The classification loss provides direction for the optimization, while f-SPL, calculated on the output of an intermediate layer, enforces that the learned representation preserves pairwise similarity.

Recall that the theoretical guarantee of the f-SPL is subject to a *balanced condition*:

$$\sum_{p \in I} (C_p - 1) = 0,$$

where $C : A_Y \times A_Y \mapsto [0, 2]$ is the label similarity. We wish to satisfy this condition regardless of data shuffling or the selection of batch size, while still allowing for randomness. Therefore, we generate the pairs in a triplet form. For each anchor (x_a, y_a), we pick a similar example (x_s, y_s) and a dissimilar example (x_d, y_d) that satisfy $C(y_a, y_s) - 1 \approx 1 - C(y_a, y_d)$. Specifically, we calculate the *label similarity* between the anchor and all other examples (can be reduced to a subset of examples, if the training set is very large) and generate a dictionary with unique similarity values (rounded to two decimal point) as keys and utterance indices as values. We keep a key only if 2-key is also in the dictionary. When generating a triplet, we randomly select a key c, and two examples, each from c and $2 - c$, respectively. As a result, every batch of triplets

$$T = \{tri = [(x_a, y_a), (x_s, y_s), (x_d, y_d)], ...\}$$

naturally implies a balanced set

$$I_T = \{s = [(x_a, y_a), (x_s, y_s)], d = [(x_a, y_a), (x_d, y_d)], ...\}.$$

The overall loss function for each triplet, tri, is

$$L(tri; g, \omega) = L_{cls}(y_a, \hat{y_a}) + \alpha(f\text{-SPL}(S_s(g), C_s) + f\text{-SPL}(S_d(g), C_d)),$$

where $\hat{y_a} = \omega(g(x_a))$ is the prediction over classes.

In the loss function, L_{cls} is the classification loss calculated on the anchor only, and α is the trade off between L_{cls} and f-SPL. C_s and C_d are the label similarity between y_a and y_s, y_a and y_d, respectively. $S_s(g)$ and $S_d(g)$ are the similarity between $g(x_a)$ and $g(x_s)$, $g(x_a)$ and $g(x_d)$, respectively.

The total loss of the batch is the mean of all triplets' losses: $L(T; g, \omega) := \frac{1}{N} \sum_{tri \in T} L(tri; g, \omega)$, where N is the batch size. Note that the f-SPL portion of $L(T; g, \omega)$ equals $\frac{\alpha}{N} f\text{-SPL}(I_T; g)$.

The multi-task framework is only used in the training phase. In the testing phase, the trained network takes batches of individual examples as the input.

4 Experiments

We experiment on IEMOCAP (Busso et al. 2008) and MSP-Improv (Busso et al. 2017). We select these datasets because: (1) they are relatively large, which allows us to train neural networks; (2) they provide ordinal evaluations of valence and activation; (3) they use similar emotion elicitation methods, but differ in speakers, lexical content, recording conditions, and the number of evaluations per utterance.

All experiments use cross-corpus evaluation. This results in four experiments (2 training-testing combination $\times$ 2 dimensions). We introduce the data, model, and experimental settings in more detail in the following subsections.

4.1 Data

IEMOCAP The IEMOCAP dataset consists of five dyadic sessions, each between a male and a female actor, resulting in 12 hours of recordings in total. Interactive scenarios, both scripted and improvised, were used to elicit the emotions of the speakers. The dataset was segmented into 10,039 utterances according to speaker turns. The valence and activation levels of each utterance were assessed by at least two evaluators using a 5-point Likert scale (Busso et al. 2008).

MSP-Improv The MSP-Improv dataset contains six dyadic sessions, each between a male and a female actor, resulting in nine hours of speech. The emotion elicitation methods include both improvisations and target sentences embedded in interactive scenes. Similar to IEMOCAP, the dataset was segmented into 8,438 utterances. Each utterance was evaluated by at least five annotators for valence and activation using a 5-point Likert scale (Busso et al. 2017).

Labels We focus on predicting binary valence and activation, where the classifiers are trained using soft labels. We consider each evaluation as a vote to the two classes, weighted by the distance to the opposite class. For example, an evaluation value of 2 on the 5-point scale is converted to [0.75, 0.25]. For each utterance, we average over the converted evaluations and use the resulting two-dimensional vector as the final soft label. The vector representing a soft label always sums to one. The label similarity, $C \in [0, 2]$, is calculated by $2 - 2d$, where d is the total variation distance ($\in [0, 1]$) between a pair of labels. Given the way we generate the soft labels, d is equivalent to the scaled Euclidean distance between the average of the raw evaluations on the one-dimensional space.

Features We preprocess the data such that the audio sampling rate is 16,000 Hz for both datasets. We then extract 40-dimensional log Mel-frequency Filterbank energy (MFB) using Kaldi (Povey et al. 2011), with a frame size of 25ms and a step size of 10ms, as in (Aldeneh and Mower Provost 2017; Zhang, Essl, and Mower Provost 2017; Aldeneh et al. 2017). We perform z-normalization for each feature dimension at the frame-level over each dataset, individually.

4.2 Classification Model

We use temporal Convolutional Neural Networks with global pooling (*Conv-Pool*) as our model. The Conv-Pool structure has been demonstrated to be the state-of-the-art on categorical emotion recognition in (Aldeneh and Mower Provost 2017), and has shown good performance on predicting the distribution of emotion perception in (Zhang, Essl, and Mower Provost 2017). Figure 2 shows the architecture the network. It consists of a 1D convolutional layer over time with 128 kernels and a kernel width of 16, a global max pooling, two fully-connected layers with a layer size of 128, and a final fully connected softmax layer. These hyper-parameters are selected according to (Aldeneh and Mower Provost 2017; Zhang, Essl, and Mower Provost 2017). The inputs to the network are the variable-length MFBs. The global max-pooling layer summarizes the output of the 1D convolutional layer and generates a fixed-length representation. This representation is then fed into the fully-connected layers. We use Rectified Linear Units (ReLU) as the activation functions, except in the last fully-connected layer, where softmax is used instead.

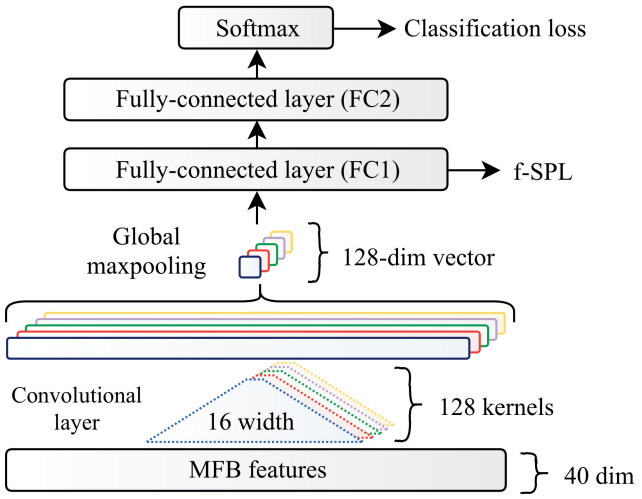

Figure 2: The Conv-Pool network structure.

We calculate f-SPL on the output of FC1 (see Figure 2). In this way, we allow room for modeling non-linearity on both sides of the intermediate representation. We first normalize the output of FC1 to unit vectors and then calculate the Euclidean distance between the embeddings. It is worth noting that although the distance, D, between two unit vectors has a range of $[0, 2]$, our embeddings have non-negative entries due to ReLU and thus $D \in [0, \sqrt{2}]$. Therefore, we scale the distance and convert it to the embedding similarity $S \in [0, 2]$ by $2 - \sqrt{2}D$.

The structure of the model is kept the same in all experiments. We use cross-entropy computed using the soft labels as the classification loss. We weigh the two classes using $N/\left(2\sum_{i=1}^{N} y_i^c\right)$ in the loss calculation to reduce the influence of data imbalance. Here, N is the total number of training utterances, y_i^c is the value for class c in the label vector of data point i. We consider a loss function containing only the cross-entropy classification loss as the baseline. For the multi-task loss, we select f-SPL based on the con-

Train	Test	Dim	Epoch	Chance	CE	CE+KL	CE+RKL	CE+PS	CE+HLG	CE+JS
MSP	IEMOCAP	Valence	23	58.41	64.45	**66.01***	65.73*	65.55*	65.76*	65.94*
		Activation	33	62.74	80.79	**81.67***	81.16	81.53*	81.10	81.41
IEMOCAP	MSP	Valence	20	54.43	**61.31**	60.81	60.71	60.48	60.78	60.44
		Activation	16	54.21	72.74	74.17*	**74.30***	74.17*	74.04*	74.03*

Table 2: UAR (%) for the four cross-corpus experiments. The best performance in each experiment is marked by bold and underline. Epoch: the number of epochs trained; CE: cross-entropy loss only (baseline); CE+KL, CE+RKL, CE+PS, CE+HLG, and CE+JS: multi-task with cross-entropy and f-SPL, where f is the convex function that corresponds to KL divergence, Reserve-KL, Pearson χ^2, Squared Hellinger, and JS divergence, respectively. "*" indicates that the marked performance is significantly better than CE, where significance is assessed at p <0.05 using the Tukey's honest test on the ANOVA statistics.

vex functions corresponding the five common f-divergence measures, as shown in Table 1. An epsilon value of 1e-12 is added to the denominators and the input of log in f-SPL in implementation for numerical stability.

For the multi-task frameworks, every training example is used as the anchor once in each epoch. Therefore, the classification loss is calculated over the same data as in the baseline. The triplets are randomly generated using the similarity dictionary (Section 3.3) at the beginning of each epoch. Empirical results show that the values of f-SPL is about a magnitude smaller than the classification loss, because triplets with extreme similarity values are rare in our data. Therefore, we use a α value of 10 in the loss function.

4.3 Performance Measure and Cross-Validation

In the testing phase, we convert the output of the network to a class prediction. We use Unweighted Average Recall (UAR) as the performance measure due to data imbalance, as discussed in (Rosenberg 2012). In the case that the ground truth labels are tied (i.e., [0.5, 0.5]), we consider predictions for either class as correct, as in (Aldeneh et al. 2017). This is true for both the baseline CNN and f-SPL approaches. As a result, the chance performance calculated by generating predictions uniformly at random is higher than 50%.

We experiment using PyTorch version 0.2.0, using a learning rate of 0.0001 with the Adam optimizer (Kingma and Ba 2015) and a batch size of 100. We select weight decay in $\{0, 0.0001, 0.001, 0.01\}$ and the number of epochs to train in $[1, 50]$ by leave-one-session-out cross-validation (LOSOCV) on the training dataset. In each experiment (e.g., valence, train on IEMOCAP and test on MSP-Improv), the weight decay and number of training epochs that lead to the highest LOSOCV UAR of the baseline model (averaged over three runs) are used for all models. In the cross-corpus training and testing, we run each experiment 30 times to reduce performance fluctuations. We report the average UAR and conduct significance tests using all the results.

5 Results and Discussion

5.1 Performance Comparison

We present the UAR of the four experiments (2 training-testing combinations $\times$ 2 dimensions) of all the models in Table 2. Each reported UAR is averaged over 30 runs. All cross-validation experiments selected the same weight-decay value of 0.001. The models include:

- CE: Conv-Pool network (Figure 2) with only cross-entropy classification loss. This is used as the baseline.
- CE+f, where $f \in$ KL, RKL, PS, HLG, JS: Conv-Pool network using the multi-task framework illustrated in Figure 1, with the convex functions corresponding to KL divergence, Reserve-KL, Pearson χ^2, Squared Hellinger, and JS divergence as f for f-SPL.

For each experiment, we first test if the influence of model is significant, using a one-way Analysis of variance (ANOVA) test and asserting significance at p <0.05. We treat the result of each run as a random example, and group them by the model. This results in 180 examples (30 runs$\times$6 models) in each test. We find that the influence of model is significant for valence when training on MSP-Improv and testing on IEMOCAP (denoted as *MSP→IEMOCAP Valence*), and for activation with both training-testing combinations. The statistics are F(5,174)=8.1, p=6.9e-7 for *MSP→IEMOCAP Valence*, F(5,174)=3.7, p=0.0033 for *MSP→IEMOCAP Activation*, and F(5,174)=9.7, p=3.3e-8 for *IEMOCAP→MSP Activation*, respectively.

We find that in three out of four experiments, all the five CE+f models show consistent performance improvement over the baseline CE model, with the only exception of *IEMOCAP→MSP Valence*. For the experiments where the influence of model is significant, we conduct pairwise comparisons using the Tukey's honest test on the statistics of the ANOVA and assert significance at p <0.05. We find that in *MSP→IEMOCAP Valence*, all five CE+f models are significantly better than CE, with p = 6.9e-7, 1.0e-4, 0.0017, 6.9e-5, and 2.6e-6 for CE+KL, CE+RKL, CE+PS, CE+HLG, and CE+JS, respectively. In *MSP→IEMOCAP Activation*, CE+KL and CE+PS has significantly higher UAR than CE (p=0.0028 and 0.022, respectively). In *IEMOCAP→MSP Activation*, all the five CE+f models have significantly better performance than CE. The p-values are 9.1e-7, 6.7e-8, 8.8e-7, 1.2e-5, and 1.3e-5 for CE+KL, CE+RKL, CE+PS, CE+HLG, and CE+JS, respectively. We do not observe any significant difference between the performances of the five CE+f models in any experiments.

5.2 Analysis of Results

We further analyze the results to better understand the reasons behind the improvement in performance. We plot the test UAR against the number of training epochs in Figure 3 for the two experiments where the CE+f models achieved

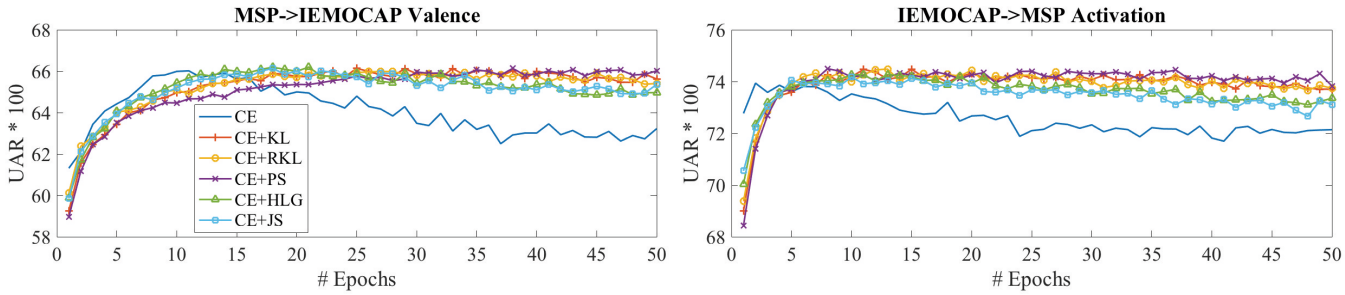

Figure 3: Test UAR against the number of training epochs for *MSP→IEMOCAP Valence* and *IEMOCAP→MSP Activation*.

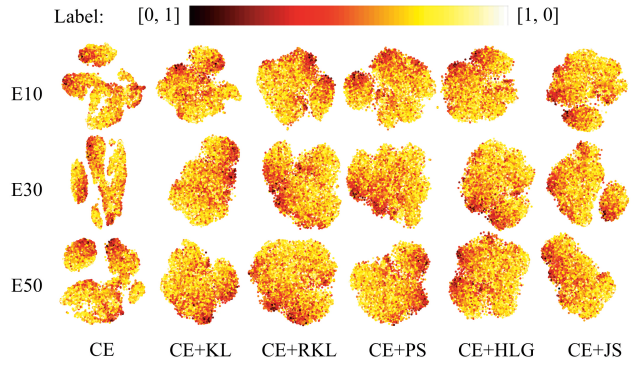

Figure 4: *IEMOCAP→MSP Activation* embedding visualization. Dots are data points in MSP. Colors represent the activation labels, the darkest are [0, 1] and the lightest are [1,0]. The rows are the embeddings at epoch 10, 30, 50 (e.g., E50). The columns correspond to the six models.

the highest performance gain over the baseline CE model. We find that while the optimal results from different models do not differ much, the CE+f models are more stable over time. More specifically, the CE model reaches the best UAR around epoch 10 in *MSP→IEMOCAP Valence* and within 5 epochs in *IEMOCAP→MSP Activation*. It starts to show signs of over-training after that, even before reaching the number of epochs to train we set and with weight-decay, when both hyper-parameters are selected by cross-validation. In contrast, the proposed CE+f models with the exact same hyper-parameters do not show too much performance decline after reaching the highest UAR.

We visualize the learned feature embeddings at epoch 10, 30, and 50 for *IEMOCAP→MSP Activation* with t-Distributed Stochastic Neighbor Embedding in Figure 4. The color of the dots in the figure represents the soft labels. The dark end of the color gradient represents [0, 1] and the light end represents [1, 0]. We find that the baseline CE models lead to several clusters, but the clusters do not correspond to labels. On the other hand, the CE+f models often lead to a single cluster where the opposite labels are more well separated and the data that are more uncertain (e.g., label $\sim$ [0.5, 0.5]) are in between. This shows that we can learn an embedding that has emotional meaning using a multi-task framework combining classification loss and f-SPL.

6 Conclusions

In this paper, we propose a family of loss functions, f-Similarity Preservation Loss, based on the dual form of f-divergence. These loss functions are designed for deep metric learning with soft labels, i.e., labels with continuous values along one or multiple dimensions. We prove mathematically that the minimizer of the proposed loss functions, a set of nonlinear mappings through neural networks, preserves the pairwise label similarities in the learned feature embeddings when the pairs of data satisfy a balanced condition. We propose a pair sampling method that guarantees the balanced condition regardless of shuffling and batch size without losing randomness. Finally, we introduce a framework that combines f-SPL with the traditional classification loss.

We apply the proposed methods on the task of cross-corpus speech emotion recognition with dimensional emotion descriptors. We show that our methods significantly outperform the baseline model, which uses only the classification loss for optimization. This demonstrates the efficacy of our f-SPL in the multi-task framework. Further analysis shows that our methods are more robust to over-training and are able to learn an emotionally-meaningful embedding space. In the future, we are interested to explore whether f-SPL can also be effectively applied to transfer learning, with a small set of labeled data from the target domain.

Acknowledgements

This material is based in part upon work supported by the Michigan Institute for Data Science ("MIDAS"), by Toyota Research Institute ("TRI"), and by the National Science Foundation (NSF CAREER 1651740). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the MIDAS, NSF, TRI, or any other Toyota entity.

References

Aldeneh, Z., and Mower Provost, E. 2017. Using regional saliency for speech emotion recognition. In *ICASSP*, 2741–2745.

Aldeneh, Z.; Khorram, S.; Dimitriadis, D.; and Mower Provost, E. 2017. Pooling acoustic and lexical features for the prediction of valence. In *ICMI*, 68–72.

Ali, S. M., and Silvey, S. D. 1966. A general class of coefficients of divergence of one distribution from another. *J. R. Stat. Soc. Series B Stat. Methodol.* 131–142.

Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; and Shah, R. 1994. Signature verification using a "siamese" time delay neural network. In *NIPS*, 737–744.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42(4):335.

Busso, C.; Parthasarathy, S.; Burmania, A.; AbdelWahab, M.; Sadoughi, N.; and Mower Provost, E. 2017. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Trans. Affect. Comput.* 8(1):67–80.

Chopra, S.; Hadsell, R.; and LeCun, Y. 2005. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, volume 1, 539–546.

Csiszár, I.; Shields, P. C.; et al. 2004. Information theory and statistics: A tutorial. *Found. Trends Commun. Inf. Theory* 1(4):417–528.

Doumanoglou, A.; Balntas, V.; Kouskouridas, R.; and Kim, T.-K. 2016. Siamese regression networks with efficient mid-level feature extraction for 3d object pose estimation. *arXiv:1607.02257*.

Fayek, H. M.; Lech, M.; and Cavedon, L. 2016. Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels. In *IJCNN*, 566–570.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. In *NIPS*.

Hoffer, E., and Ailon, N. 2015. Deep metric learning using triplet network. In *SIMBAD*, 84–92.

Kingma, D. P., and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Kong, Y., and Schoenebeck, G. 2018. Water from two rocks: Maximizing the mutual information. In *Economics and Computation*, 177–194.

Liu, H.; Tian, Y.; Yang, Y.; Pang, L.; and Huang, T. 2016. Deep relative distance learning: Tell the difference between similar vehicles. In *CVPR*, 2167–2175.

Liu, X.; Kumar, B. V.; You, J.; and Jia, P. 2017. Adaptive deep metric learning for identity-aware facial expression recognition. In *CVPR*, 522–531.

Lotfian, R., and Busso, C. 2017. Formulating emotion perception as a probabilistic model with application to categorical emotion classification. In *ACII*, 415–420.

Lu, J.; Hu, J.; and Zhou, J. 2017. Deep metric learning for visual understanding: An overview of recent advances. *IEEE Signal Process. Mag.* 34(6):76–84.

Mikels, J. A.; Fredrickson, B. L.; Larkin, G. R.; Lindberg, C. M.; Maglio, S. J.; and Reuter-Lorenz, P. A. 2005. Emotional category data on images from the international affective picture system. *Behav. Res. Methods.* 37(4):626–630.

Mower, E.; Mataric, M. J.; and Narayanan, S. 2011. A framework for automatic human emotion classification using emotion profiles. *IEEE Audio, Speech, Language Process.* 19(5):1057–1070.

Nguyen, X.; Wainwright, M. J.; and Jordan, M. I. 2009. On surrogate loss functions and f-divergences. *Ann. Stat.* 876–904.

Nguyen, X.; Wainwright, M. J.; and Jordan, M. I. 2010. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Inf. Theory* 56(11):5847–5861.

Nowozin, S.; Cseke, B.; and Tomioka, R. 2016. f-gan: Training generative neural samplers using variational divergence minimization. In *NIPS*, 271–279.

Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; Silovsky, J.; Stemmer, G.; and Vesely, K. 2011. The kaldi speech recognition toolkit. In *ASRU*.

Rockafellar, R. T., et al. 1966. Extension of fenchel'duality theorem for convex functions. *Duke Math. J.* 33(1):81–89.

Rosenberg, A. 2012. Classifying skewed data: Importance weighting to optimize average recall. In *Interspeech*.

Russell, J. A. 1980. A circumplex model of affect. *J. Pers. Soc. Psychol.* 39(6):1161.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 815–823.

Sohn, K. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, 1857–1865.

Song, H. O.; Xiang, Y.; Jegelka, S.; and Savarese, S. 2016. Deep metric learning via lifted structured feature embedding. In *CVPR*, 4004–4012.

Steidl, S.; Levit, M.; Batliner, A.; Noth, E.; and Niemann, H. 2005. "of all things the measure is man" automatic classification of emotions and inter-labeler consistency. In *ICASSP*.

Taigman, Y.; Yang, M.; Ranzato, M.; and Wolf, L. 2014. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 1701–1708.

Thiel, C. 2008. Classification on soft labels is robust against label noise. In *KES*, 65–73.

Wang, Q.; Wan, J.; and Yuan, Y. 2017. Deep metric learning for crowdedness regression. *IEEE Trans. Circuits Syst. Video Technol.*

Weinberger, K. Q., and Saul, L. K. 2009. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.* 10(Feb):207–244.

Yang, J.; She, D.; Lai, Y.; and Yang, M.-H. 2018. Retrieving and classifying affective images via deep metric learning. In *AAAI*.

Yi, D.; Lei, Z.; Liao, S.; and Li, S. Z. 2014. Deep metric learning for person re-identification. In *ICPR*, 34–39.

Zhang, B.; Essl, G.; and Mower Provost, E. 2017. Predicting the distribution of emotion perception: capturing inter-rater variability. In *ICMI*, 51–59.