AN INTEGRATED MULTI-MODAL MULTI-LABEL FRAMEWORK IN DEEP METRIC LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning models are increasingly applied in domains where supervised performance is not the primary objective. Domains such as healthcare demand machine learning models which provide representations for complex relationships between both heterogeneous modes of data, and multiple co-occurring labels. Previous works have tackled representation learning in such multi-label and multi-modal setting, but have neglected to consider the common requirement of generalization to novel, and unknown, tasks at test-time. In this work, we propose an integrated multi-modal multi-label framework for Deep Metric Learning (DML), which we term 3ML–DML. Our framework extends existing proxy learning losses from DML to the multi-label domain, and provides a novel method for enforcement of label correlations via these proxies. We further introduce a multi-modal component which builds a standard fusion model but draws from DML literature in order to incorporate auxiliary, high-dimensional embeddings and feature spaces from each mode of data as context to match with and refine the output of the fusion model for further improvements in generalization performance. Indeed, when exploring our method in a variety of settings, including on healthcare data, we are able to demonstrate consistent improvements over constructed baselines both in the context of multi-label multi-modal learning but most poignantly, in zero-shot generalization to new labels.

1 INTRODUCTION

Learning from multi-modal data is a current frontier of machine learning research (Baltrušaitis et al., 2018; Nagrani et al., 2021; Zhang et al., 2020). High-impact domains frequently contain vast information spread across several modalities (Beam & Kohane, 2018), and demand robust machine learning solutions that can learn sensible representations of these complex data. Further, representing data that have multiple, simultaneous *labels* exists as an open area of research. In a variety of domains, these two tasks (multi-modal and multi-label learning) coincide. For instance, clinical applications commonly require observation of several distinct data modes, such as vitals, labs, medical imaging, and outcomes range over possible diagnoses that often co-occur (Irvin et al., 2019). Identifying similarity across patients is crucial to healthcare tasks like devising effective treatment plans (Hripcsak et al., 2016), yet defining similarity in high-dimensional, multi-modal data is notoriously challenging (Ghassemi et al., 2020). Despite its importance, quantifying similarity in multi-label and multi-modal data remains largely unstudied.

A powerful approach to quantifying similarity in high-dimensional data is through techniques from Deep Metric Learning (DML). DML aims to learn representation spaces in which semantic similarity between instances is reflected as distance between embeddings. While the idea conceptually lends itself to settings with multi-modal data with multiple labels, DML has been primarily developed for single-label unimodal settings (Schroff et al., 2015; Oh Song et al., 2016; Ge, 2018; Wu et al., 2017; Roth et al., 2020; Musgrave et al., 2020). Much less emphasis and interest has been placed in finding solutions to quantify multi-label similarity via DML, with primary focus on multi-label classification (Liu & Tsang, 2015b; Li et al., 2019; Sun & Zhang, 2021). Unfortunately, a naive classification setup does not port to tasks requiring explicit quantification of similarity, especially for classes unseen during training (Schroff et al., 2015; Wu et al., 2017; Milbich et al., 2021). As we show experimentally, simple extensions of standard DML methods prove insufficient.

Similarly, little research has been done for multi-modal metric learning (Xie & P Xing, 2013; Won et al., 2021), with existing methods focusing on the constrained task of standard metric learning or for highly specific domains. Some recent approaches do seek to learn *cross*-modal similarity, particularly for retrieval tasks (Xu et al., 2019; Huang & Peng, 2017; Zhen et al., 2019; Cao et al., 2019; He et al., 2016b), leveraging multi-modal contrastive and metric learning for tasks such as retrieving the best text for an image. Instead, we aim to develop multi-modal embedding profiles that combine multiple modalities, e.g. both text and images, to quantify similarity between multi-modal data instances towards improved label prediction.

We therefore extend DML to tackle the challenging and important problem of quantifying similarities for multi-label data across multiple modalities. Given a dataset of multi-label instances from multiple modalities, we seek to learn an embedding function to a metric presentation space in which instances with similar *label sets* should themselves be similar. Importantly, for instances with previously unseen label-sets, this property should still hold.

To overcome the challenges unique to multi-label, multi-modal DML, we introduce 3ML-DML, **m**ulti-**m**odal **m**ulti-**l**abel **d**eep **m**etric **l**earning. First, to encourage effective multi-label embeddings, we extend existing proxy-based loss terms in DML such that each label associates with a learned proxy in the (concurrently learned) embedding space. However, we iterate over previous multi-class proxy-based methods by permitting embedded data instances to cluster near *multiple* proxies, representing the multi-label nature of the data. Then, we incorporate an additional loss term that forces the distance space over the learned proxies to match the label correlations between associated labels. Finally, we introduce a unique way to learn on multi-modal data, which concurrently learns a (concatenation) fusion model that is encouraged to represent higher dimensional embeddings of each individual modality via mutual information maximization and self-distillation.

Empirically, we demonstrate the 3ML-DML succeeds in quantification of similarity over multi-modal, multi-label instances. Using two large datasets, we find that 3ML-DML 1) improves performance over relevant baselines in multi-modal multi-label tasks; 2) generalizes exceptionally to unseen test labels and label sets; and 3) unifies its objectives such that each component of the framework stacks in terms of performance gains.

In summary, we show that our approach outperforms existing DML methods for multi-label learning tasks on multi-modal data on impactful real-world tasks, both in healthcare and in other domains. To our knowledge, we are the first to study the integrated multi-modal and multi-label setting for DML. We additionally present a novel method: we are the first to combine recent key ideas from disparate research ares into a new and unified framework.

2 RELATED WORK

Deep Metric Learning Advances in similarity learning are majorly driven by novel insights from Deep Metric Learning (DML), with applications ranging from zeros-shot retrieval (Oh Song et al., 2016; Wu et al., 2017; Roth et al., 2020), clustering (Ge, 2018; Sohn et al., 2019), verification (Deng et al., 2019; Liu et al., 2017), few-shot learning (Snell et al., 2017), and unsupervised representation learning (He et al., 2020; Chen et al., 2020). DML methods are commonly separated based on their use of ranking objectives (Hadsell et al., 2006; Chen et al., 2017; Sohn, 2016) and tuple mining heuristics (Schroff et al., 2015; Wu et al., 2017), proxy- or classification-based training (Movshovitz-Attias et al., 2017; Teh et al., 2020) and orthogonal extensions in multi-task and cross-modal settings (Roth et al., 2022a; Milbich et al., 2020).

Previous works in DML predominantly tackle unimodal, single-class-per-sample setting. While initial work has investigated normal metric learning without deep networks in the multi-label setting (Liu & Tsang, 2015b; Gouk et al., 2016), metric learning for multi-label classification (Li et al., 2019; Sun & Zhang, 2021) and single-label multi-modal retrieval (Xie & P Xing, 2013; Won et al., 2021), we are the first to investigate extending DML to multi-label multi-modal retrieval tasks.

Multi-label Learning Multi-label classification defines the task of predicting a label *set* of a given data instance and is a cornerstone machine learning task (Liu et al., 2021a). A core concept in multi-label learning is reasoning over relationships between labels, a feature missed by multi-class methods. The task is often solved via classifier chains (Chen et al., 2018; Gerych et al., 2021; Hartvigsen et al., 2020), bayesian networks (Zaragoza et al., 2011), multi-task learning (Liu et al., 2018), clustering (Shu et al., 2022), and embedding methods (Bhatia et al., 2015). Embedding

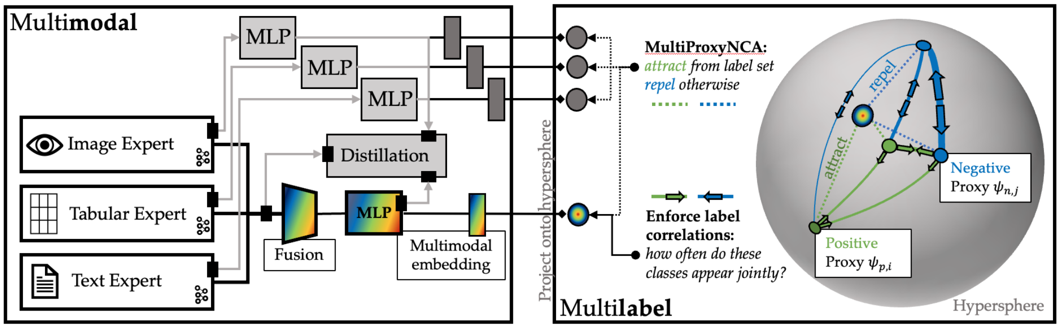

Figure 1: Visualization of our proposed *3ML-DML* framework for effective multimodal multilabel Deep Metric Learning.

methods embed instances into feature vectors for which their k-nearest neighbors have similar label sets, and can also be applied to the labels themselves (Huang & Lin, 2017). Thus, embedding methods are pertinent to DML. However, classic multi-label embedding methods are prohibitively slow (Liu & Tsang, 2015a; Bhatia et al., 2015), so recent works have induced sparsity in embeddings (Shen et al., 2018) or select features directly (Gonzalez-Lopez et al., 2019).

Multi-modal Representation Learning Multimodal representation learning constitutes the task of learning feature representations from multiple modes of data (Guo et al., 2019; Zhang et al., 2020). Prior work has demonstrated the utility of learning from multiple modalities – without overfitting to a specific modality – as opposed to each modality individually (Liang et al., 2021; Wang et al., 2020; Wu et al., 2022). Standardly, multi-modal learning is achieved by learning a *joint* representation that encodes information from all modalities into a single space. This fusion operation can range from concatenation (Anastasopoulos et al., 2019) to more complex model-based (Yang et al., 2016) or attention-based fusion (Nagrani et al., 2021; Jaegle et al., 2021; Pramanik et al., 2019).

A related line of work, *cross-modal* learning, aims to maintain the similarity structure between samples across modalities (Zhang et al., 2019). This cross-modal similarity constraint can be enforced with a static or model-based distance metric (Hsu et al., 2018; Hardoon et al., 2004; Frome et al., 2013), through direct mutual information maximization (Liao et al., 2021), or through contrastive learning (Radford et al., 2021; Yuan et al., 2021; Zolfaghari et al., 2021). Cross- and multi-modal representations have been used in application areas such as zero-shot learning (Radford et al., 2021; Parida et al., 2020), cross modal ranking and retrieval (Parida et al., 2020; Wang et al., 2015), and visual question answering (Cadene et al., 2019; Ben-Younes et al., 2019). Ultimately, our method leverages key insights from fusion models in multi-modal learning and mutual information maximization in cross-modal learning to produce a highly informative multi-modal representation that incorporates mode-specific information beyond fusion.

3 Methods

To quantify similarities on both a multi-modal and multi-label level that allow us to tackle zero-shot similarity tasks, we build on the Deep Metric Learning (DML) framework. DML provides tools to quantify non-linear semantic similarities across data points, and as such is a well suited starting point. We quickly motivate the DML framework in §3.1, before proposing a suitable multi-label DML objective in §3.1.1. To best leverage the multi-label context, we further propose a novel extension based on label correlation in §3.1.2 which allows us to incorporate relations between different labels into the multi-label DML process. To operate on multiple modalities, we propose a multi-modal architecture well suited to be used in conjunction with a DML-based problem setting in §3.2, which together form our multi-modal multi-label DML (3ML-DML) objective §3.3. For a visualization of the complete 3ML-DML framework, we refer to Fig. 1.

3.1 Multi-label Deep Metric Learning

Formally, given training data $\mathcal{X}$, the goal in DML is to find a projection ϕ, commonly parametrized by a deep neural network, into a d-dimensional metric space $\phi : \mathcal{X} \to \Phi \subset \mathbb{R}^d$ such that for data pairs $x_1, x_2 \in \mathcal{X}$, a predefined distance metric $d(\phi(x_1), \phi(x_2))$ (commonly euclidean or cosine distance) between projected ("embedded") data samples quantifies important semantic relations. For regularization purposes, Φ is commonly projected to the unit hypersphere $\mathcal{S}^{d-1}$ (Weisstein, 2002; Wu et al., 2017). In the standard case where samples are only associated with a single label, surrogate training objectives to optimize ϕ are straightforward to define, for example via ranking-based losses (e.g. triplet (Schroff et al., 2015; Wu et al., 2017) or other tuple-based approaches (Chen et al., 2017; Sohn, 2016)) or proxy-based objectives (Movshovitz-Attias et al., 2017; Qian et al., 2019; Teh et al., 2020; Kim et al., 2020). While the latter is considered to generally perform favourably with better convergence properties (Movshovitz-Attias et al., 2017; Kim et al., 2020; Roth et al., 2022b), all of these approaches aim to attract samples from the same while repelling samples from different classes.

However, as we want to quantify similarities between multi-label samples, these standard approaches, which assume single label instances, become insufficient. To tackle this issue, our goal is to provide a suitable *multi-label* DML objective. In particular, as proxy-based objectives provide most of the current state-of-the-art metric learning models (Kim et al., 2020; Teh et al., 2020; Roth et al., 2022b), we aim to adapt the proxy-formulation to the multi-label case.

3.1.1 Multi-proxy Deep Metric Learning

In particular, we extend the ProxyNCA formulation (Movshovitz-Attias et al., 2017), which is the underlying concept behind various recent, more specialized extension Qian et al. (2019); Kim et al. (2020); Teh et al. (2020); Roth et al. (2022b).

In the single label setting, given class labels $c \in \mathcal{C}$, associated proxy representations $\psi_c \in \mathbb{R}^d$, and a minibatch of samples $\mathcal{B}$, the ProxyNCA loss is defined as

$$\mathcal{L}_{\text{PNCA}} = -\frac{1}{b} \sum_{x_i \in \mathcal{B}} \log \left(\frac{\exp\left(-d\left(\phi(x_i), \psi_{c(x_i)}\right)\right)}{\sum_{c^* \in \mathcal{C} \setminus \{c(x_i)\}} \exp\left(-d\left(\phi(x_i), \psi_{c^*}\right)\right)} \right) \tag{1}$$

where $c(x_i)$ denotes the class associated with sample x_i. This formulations allows for a straightforward extensions into the multi-label setting, where every sample is now assigned a label set instead of a single label. As such, let $\ell \in \mathcal{Y}$ be the set of *unique* labels, to each of which we assign a proxy $\psi_\ell \in \mathbb{R}^d$. Then, for each sample with label set $Y(x_i)$ in the minibatch $\mathcal{B}$, we define the multi-label ProxyNCA objective as

$$\mathcal{L}_{\text{MPNCA}} = -\frac{1}{b} \sum_{x_i \in \mathcal{B}} \log \left(\frac{\sum_{\ell \in Y(x_i)} \exp\left(-d\left(\phi(x_i), \psi_\ell\right)\right)}{\sum_{\ell \in \mathcal{Y} \setminus \{Y(x_i)\}} \exp\left(-d\left(\phi(x_i), \psi_\ell\right)\right)} \right) \tag{2}$$

Within this framework, we aim to minimize the distance between a sample embedding $\phi(x_i)$ and all associated *positive* proxies, i.e. those part of the associated label set $Y(x_i)$, while repelling the remaining proxies associated with labels not part of $Y(x_i)$.

3.1.2 Label Correlation Enforcement

The standalone multi-label ProxyNCA formulation does not account for any relations between different labels, which is a crucial element in multi-label-based representation learning; the benefit of having multiple labels is an additional axis of information regarding the interplay between different class concepts (Tsoumakas & Katakis, 2007; Maxwell et al., 2017; Liu et al., 2021b), as in practice, labels often appear in correlation. This has to be reflected in the metric representation space. In particular, proxies that belong to label concepts that share strong correlations should be explicitly close in the representation space, and vice versa. Unfortunately, the standalone ProxyNCA treats every label as an independent entity.

We therefore propose a novel label correlation enforcement, which allows us to re-align proxies based on their associations in the multi-label context - i.e. if two proxies belong to labels that often appear jointly, the should thus be reasonably close in the final metric representation space. To achieves this, we minimize the mean squared loss between the pairwise distances over the learned proxies, and the Pearson correlation coefficient over label variables observed on each batch.

This works, as for real-valued, standardized vectors u, v (zero mean, unit length), their Pearson correlation coefficient r is directly related to their Euclidean distance d via

$$r = (1 - d^2/2) \tag{3}$$

See Supplemental B.2 for derivation.

If we are now given access to a batch $\Phi^{\mathcal{B}} \in \mathbb{R}^{b \times d}$ of size b with d-dimensional embeddings generated from our DML network, where each entry $\Phi_i^{\mathcal{B}}$ is associated with a multi-hot, c-dimensional label vector $y_i \in \mathbf{Y} \in \{0,1\}^{b \times c}$ associated with each samples respective label set.

The pairwise Pearson correlation coefficient $\texttt{PCC}$ between pairs of standardized columns in the multi-label tensor $\mathbf{Y}$, denoted as $\mathbf{r}$, can then be computed as $\mathbf{r}_{ij} = \texttt{PCC}(\mathbf{Y}_{:,i}, \mathbf{Y}_{:,j})$, which gives an indicator on the relation between two classes c_i and c_j. If we now compute the euclidean distance $d(\cdot, \cdot)$ between respective proxy pairs ψ_i and ψ_j such that we are given a distance matrix $d_{ij} = d(\psi_i, \psi_j)$, we can align the proxy distances with the associated class correlations $\mathbf{r}_{ij}$ by using the relation noted in Eq. 3, such that our label correlation enforcement term can be defined as

$$\mathcal{L}_{\ell corr} = \mathcal{L}_{MSE}(\mathbf{r}, 1 - \mathbf{d}^2/2) \tag{4}$$

which gives the final, label-correlated multi-label multi-proxy objective as

$$\mathcal{L}_{\text{lcMPNCA}} = \mathcal{L}_{\text{MPNCA}} + \gamma \cdot \mathcal{L}_{\ell corr} \tag{5}$$

3.2 Metric Learning from Multiple Modalities

To incorporate context from multiple different modalities into a single metric representation space, we require a mechanism that unifies features extracted across modalities.

For that, we begin by adopting a standard concatenation framework for the fusion of multiple modalities following Nagrani et al. (2021); Zhang et al. (2020). In particular, given separate representations from mode-specific models for n modalities, $\omega_1^m \in \mathbb{R}^{k_1}, \ldots, \omega_n^m \in \mathbb{R}^{d_n}$, we concatenate the representations to create a joint multi-modal representation $\psi^m = [\omega_1^m, \ldots, \omega_n^m] \in \mathbb{R}^{d_1 \bullet \cdots \bullet d_m}$. The resulting representation ω^m is then passed into the primary part of the fusion model, where the number of layers before and after the fusion quantify the *fusion depth*. We elaborate on implementation details in the Supplemental.

However, simple fusion on its own is insufficient for good multi-modal similarity learning, as in practice we not only want to merge representation, but we also want relations between samples within each modality to be retained or at the very least actively used to inform the arrangement in the final multi-modal metric space.

3.2.1 Multi-Modal S2SD

Therefore, to better integrate multiple modalities in a way more suitable for the particular task of similarity learning, we extend the vanilla fusion approach with a simultaneous similarity-based self-distillation (S2SD) setup introduced in Roth et al. (2021) (see Supplemental B.1 for vanilla S2SD). We reframe S2SD for our multi-modal setup by attaching a mode-specific, high-dimensional MLP $\phi_{g_i}^m$ to the output of each modality expert model ω_i^m, as well as the fused, lower dimensional multi-modal output ω^m, giving ϕ^m. Similar to the standard S2SD setup, every singly modality expert is optimized with its on instantiation of the multi-label MultiProxyNCA objective, s.t. every modality on its own also learns to arrange and relate different samples.

With the new mode-specific, high-dimensional embedding spaces, we finally perform distillation via the contrastive S2SD distillation objective onto our target fusion embedding ϕ^m. The complete multi-modal S2SD objective then reads

$$\begin{aligned}\mathcal{L}_{\text{mmS2SD}} =& \frac{1}{2} \cdot \left[\mathcal{L}_{\text{MPNCA}}(\Phi^m) + \frac{1}{n}\textstyle\sum_{i=1}^{n} \mathcal{L}_{\text{MPNCA}}(\Phi_{g_i}^m)\right] \\ &+ \frac{\gamma}{n}\sum_{i=1}^{n} \mathcal{L}_{\text{dist}}(D^{\Phi^m}, D^{\Phi_{g_i}^m}) + \gamma \mathcal{L}_{\text{dist}}(D^{\Phi^m}, D^{\Omega^m})\end{aligned} \tag{6}$$

with n different modalities. We use large greek letters to denote the respective batch of embeddings generated by the different parts of our multi-modal architecture, with $\Phi^m := \phi^m(\mathcal{B})$ comprising our main lower dimensional multi-modal target embedding space.

3.3 MULTI-LABEL MULTI-MODAL METRIC LEARNING

The complete multi-modal multi-label objective, 3ML-DML, thus comprises:

$$\mathcal{L}_{\text{3ML-DML}} = \mathcal{L}_{\text{mmS2SD}} + \gamma \mathcal{L}_{\ell corr} \tag{7}$$

with $\gamma \in \mathbb{R}$ a weighting term for the label correlation loss (optimized via simple grid search), which is only applied on the primary multi-modal embedding space Φ^m. Conceptually, the complete objective can be understood as **(1)** learning a multi-label suitable embedding space through our MultiProxyNCA objective (Eq. 2), **(2)** jointly learning and distilling more finegrained, higher-dimensonal unimodal relations between samples into the main multi-modal space via the multi-modal S2SD distillation objective (Eq. 6) and finally **(3)** leveraging interclass relations through our label correlation enforcement objective (Eq. 4).

4 EXPERIMENTS

4.1 DATASETS

We benchmark our method on the datasets shown in Table 3. The `MIMIC-III` database Johnson et al. (2016) consists of clinical time series and clinical notes for over 30,000 patients from an ICU in Boston. Here, we follow data processing steps defined in prior work to obtain a cohort where the task is to predict patient membership in one or more 25 HCUP CCS ICD-9 code groups Harutyunyan et al. (2019). In addition, we use the `MM-IMDB` Liang et al. (2021); Arevalo et al. (2017) dataset, where the goal is to determine the genre of a movie using its poster and plot summary.

4.2 IMPLEMENTATION DETAILS

We use the ResNet-50 (He et al., 2016a) architecture for the image encoder. We use $\text{BERT}_{\text{BASE}}$ (Devlin et al., 2018) as the text encoder. Further details regarding dataset statistics can be found in Supplemental A. To build representations on time-series data, we use a Gated Recurrent Unit (GRU) network (Chung et al., 2014). We split samples from each dataset into 70% training, 15% validation, 15% test sets. We vary the learning rate, embedding dimension, batch size, label correlation loss weight term, S2SD embedding dimensions, and select the model with the best downstream linear AUROC on the validation set.

4.3 EVALUATION METRICS

To comprehensively measure the performance of our models, we pull from previous works in multi-label learning, healthcare, and capitalize on standard metrics in deep metric learning. In particular, we focus on area-under-receiver-operator-curve (AUROC, or AUC) via linear classification downstream , mean average precision (mAP) , and recall@k (see Supplemental for precise definitions). AUROC constitutes a commonly used metric in multi-label learning, while mAP is standardly used in both multi-label learning and deep metric learning. We extend standard recall@k in deep metric learning to the multi-label setting.

In order to adapt recall@k for multi-label tasks, we change the local measure of recall@k from *existence* of a positive sample in a point's k-nearest neighbors to *the highest overlap in proportion of labels* between a data point and its k-nearest neighbors.

Definition 1 (Multi-Label Recall@k). *Given $k \in \{1, \ldots, |X|\}$, define $NN_k : \Phi \subset \mathcal{S}^{D-1} \to \mathcal{P}(X)$ as a function that receives a point $\phi(x) \in \Phi$ and returns a set in the powerset of X, $\mathcal{P}(X)$, containing points in X that map to the k nearest neighbors of $\phi(x)$ in Φ. Let $\texttt{MultiHot} : \mathcal{P}(\mathcal{Y}) \mapsto [0,1]^{|\mathcal{Y}|}$ be a function that encodes label sets as multi-hot vectors. Then, multi-label recall@k is defined as:*

$$\textit{Recall@k} = \frac{1}{|X|} \sum_{x \in X} \max_{\tilde{x} \in NN_k(x)} 1 - \frac{\|\texttt{MultiHot}(Y(x)) - \texttt{MultiHot}(Y(\tilde{x}))\|_1}{|\mathcal{Y}|}$$

where $Y(x)$ indicates the label set for point $x \in X$.

We extrapolate on additional adjustments to the proposed multi-label recall@k metric in Supplemental B.1.

4.4 BASELINES

In order to compare our method as described in Section 3 to relevant constructed baselines, we propose the following baseline methods for multi-label learning:

- `powerset`: Given the available labels for a dataset, we construct a proxy for each possible combination of labels, resulting in $2^{|\mathcal{Y}|}$ proxies, and train a DML model with standard ProxyNCA loss Movshovitz-Attias et al. (2017).
- `supervised`: We train the network by adding on a linear layer mapping the embedding space to $\mathbb{R}^{|\mathcal{Y}|}$, and minimize the unweighted sum of binary cross-entropy for each label.

The `powerset` method presents a naive multi-label method in which label correlations and dependencies are largely ignored and each label set is treated as a separate class, ported to the deep metric learning space. Comparison with the `supervised` method establishes differentiation between the most commonly used multi-label *classification* training algorithm and our method. Note that both of these baselines are expected to perform well: the `powerset` method leverages the ability of deep metric learning to learn informative embeddings on a large number of classes and generalize to novel classes; whereas the supervised method is the most commonly used method for multi-label classification in practice.

To facilitate comparison between our adapted S2SD multi-modal approach, we benchmark against simple fusion-based multi-modal learning wherein representations from each modality are concatenated and a multi-layer dense network is learned on these concatenated representations to obtain the desired DML embedding space. We use hyperparameter search to optimize the depth at which fusion occurs, and compare fusion to unimodal models with each modality.

4.5 EXPERIMENTAL SETUP

Given our datasets, we run multiple experiments in order to (1) compare the performance of our method to relevant constructed baselines; (2) understand the ability of our method to generalize to novel labels and label sets; and (3) assess the impact of each component of our method.

Novel Label Generalization For each dataset, we train a model on the entire label set for that dataset and evaluate metrics on the test set for the identical label set for a simple benchmark. However, as zero-shot generalization is a motivating principle for the proposed method, we build zero-shot generalization tasks for each of our datasets in order to evaluate the ability of our method to generalize to novel labels (and thus, novel label sets) at test-time. For each dataset in Table 3, we first remove all labels with less than 5% prevalence. Then, with the exception of `COCO`, we randomly select 50% of labels to train on, and test on the remaining 50% of labels, as is standard in multi-class deep metric learning. We average metrics over two randomly chosen sets of labels for train / test label split.

Ablation Study In order to further our understanding of which components of our model contribute to distinct aspects of performance, we perform the following experiments with each dataset. In the following settings, we perform runs for the entire label set case and in the zero-shot generalization case:

- S2SD multi-modal loss with multi-label ProxyNCA loss and label correlation (`3ML-DML`)
- S2SD multi-modal loss with multi-label ProxyNCA loss and no label correlation ($-\mathcal{L}_{corr}$)
- Standard multi-modal concatenation fusion (no S2SD) with our multi-label ProxyNCA loss and label correlation (−S2SD)
- Standard multi-modal concatenation fusion (no S2SD) with our multi-label ProxyNCA loss and no label correlation (−S2SD, $\mathcal{L}_{corr}$)

In our experiments with label correlation, we additionally vary the weighting of the label correlation term in order to determine to what extent label correlation impacts the performance of the model – particularly, in the entire label set versus the zero-shot generalization case.

Dataset	Eval Set	Loss	Metrics		
			Linear AUROC	mAP	Recall@1
MMIMDB	train	3ML-DML	**0.868**	0.089	0.884
		mmS2SD	0.866	0.085	0.883
		supervised	0.856	**0.110**	0.887
		powerset	0.858	0.102	**0.891**
		MPNCA⁻	0.820	0.099	0.880
	test	3ML-DML	**0.844**	0.068	0.873
		mmS2SD	0.842	0.065	0.872
		supervised	0.808	0.062	0.865
		powerset	0.835	**0.077**	**0.876**
		MPNCA⁻	0.780	0.065	0.868
MIMIC-III	train	3ML-DML	**0.723**	**0.154**	0.793
		mmS2SD	0.720	0.145	**0.794**
		supervised	0.713	0.130	0.791
		MPNCA⁻	0.665	0.114	0.782
	test	3ML-DML	**0.702**	0.139	0.789
		mmS2SD	0.697	**0.141**	0.788
		supervised	0.699	0.118	**0.790**
		MPNCA⁻	0.642	0.094	0.780

Table 1: Comparison of performance metrics on baseline methods. Here, MPNCA⁻ is a baseline using $\mathcal{L}_{MPNCA}$ with both positive and negative proxies. Note that the powerset method is computationally prohibitive as the number of proxies in the powerset grows exponentially with the label set size. Due to computational constraints, we exclude the “multiproxyncapowerset" results for the “mimiciii" dataset due to its larger label sizes.

Dataset	Eval Set	Loss	Metrics		
			Linear AUROC	mAP	Recall@1
MMIMDB	train	3ML-DML	0.868	0.089	0.884
		$-\mathcal{L}_{corr}$	0.866	0.085	0.883
		–S2SD	**0.871**	**0.102**	**0.888**
		–S2SD, $\mathcal{L}_{corr}$	**0.871**	0.099	0.887
	test	3ML-DML	0.844	0.068	0.873
		$-\mathcal{L}_{corr}$	0.842	0.065	0.872
		–S2SD	**0.848**	**0.075**	**0.876**
		–S2SD, $\mathcal{L}_{corr}$	0.846	0.073	0.874
MIMIC-III	train	3ML-DML	**0.723**	**0.154**	0.793
		$-\mathcal{L}_{corr}$	0.720	0.145	**0.794**
		–S2SD	0.717	0.150	0.792
		–S2SD, $\mathcal{L}_{corr}$	0.715	0.150	0.793
	test	3ML-DML	**0.702**	0.139	**0.789**
		$-\mathcal{L}_{corr}$	0.697	**0.141**	0.788
		–S2SD	0.699	0.139	**0.789**
		–S2SD, $\mathcal{L}_{corr}$	0.699	0.137	**0.789**

Table 2: Comparison of performance metrics when we ablate components from our method 3ML-DML using the procedure described in Section 4.5. Here, $-\mathcal{L}_{corr}$ indicates ablating the label correlation loss, and –S2SD indicates using standard fusion for multi-modal learning.

5 Results

5.1 Baseline Comparison

3ML-DML improves metrics across key baselines In Table 1, we present results of our methods (3ML-DML and mmS2SD) against the baselines described above. Looking at the case where we evaluate on the same label set as training, we find that on both MMIMDB and MIMIC-III, our method performs the best in terms of downstream mean linear AUROC, even outperforming the supervised baseline. In terms of the quality of the embedding space learnt, we find that all methods are fairly comparable in terms of Recall@1. We also find that our method outperforms the baselines on mAP in `MIMIC-III`, though supervised learning does do better on `MMIMDB`.

Generalization to novel labels In the Eval Set = test setting, all models are evaluated on novel labels not seen during training. Across all metrics, we observe that the model performances generally all drop slightly when generalizing from the source to target label set distributions, as expected. To compare our methods (`3ML-DML` and `mmS2SD`) with the baselines, we analyze the percent drop in performance for each method and metric after generalization. In the `MIMIC-III` dataset, our methods perform on par with baseline models across all metrics. In the `MMIMDB` dataset, our methods perform noticeably better than the baseline `supervised` and `MPNCA`$^{-}$ methods with respect to mAP and marginally better for the two other metrics. The percent drop in mAP is 43.63% for `supervised` and 34.3% for `MPNCA`$^{-}$, while only 23.6% for `3ML-DML` and 23.5% for `mmS2SD`.

5.2 3ML-DML Ablation study

Impact of each component on performance metrics According to Table 2, we observe varied results with respect to differing components of the method. In the case of `MIMIC-III`, which serves as our primary application dataset (healthcare), we observe that the unified framework `3ML-DML` provides better AUROC than each of its relative components. In particular, we see that: removing the label correlation term results in a degradation of 3%, using standard fusion as opposed to multi-modal S2SD results in a drop of 6% and both in tandem results in degradation of 8%. Thus, we conclude in our primary application dataset, we observe that each component provides improvement in the standard and zero-shot generalization case. In the `MMIMDB`, we observe that S2SD marginally degrades the overall performance compared to utilizing standard fusion and our proposed multi-label ProxyNCA alone in terms of AUROC. We hypothesize that such an occurrence arises from greater informative value in one of the modalities. In the standard fusion case, the fusion layer should down-weight the less useful modality. However, in S2SD, we ensure maximal mutual information between *both* modalities and the ultimate fusion model. Perhaps additionally we are overfitting to the weaker modality by learning a higher dimensional embedding than necessary in S2SD. We propose remedying this by further exploration of differing representation dimensionalities for each modality expert in multi-modal S2SD.

Limits of weighted label correlation In Appendix C.1, we present plots showing the effect of varying γ (the label correlation weight) on downstream performance metrics on `MMIMDB`. We find that, although $\mathcal{L}_{corr}$ does give marginal improvements as evident in Table 2, there is no clear trend showing a significant improvement in performance as γ is varied in a grid for the two losses examined. Exploring the effect of $\mathcal{L}_{corr}$ and how it varies during the training process, as well as potentially devising a method to choose γ automatically, is an area of future work.

6 Conclusion

In conclusion, we demonstrate significant value in the application of deep metric learning to the task of multi-modal multi-label learning. We propose a novel method that draws from previous work in proxy-based learning and relates learned distance functions to label correlations. We additionally propose a concurrent multi-modal method which utilizes direct distillation from high dimensional unimodal embeddings to enhance the fusion model. We show improvement in performance over relevant baselines in the standard and zero-shot generalization settings; and analyze the contribution of each component of 3ML-DML. To future work, we consider reasoning over label correlations directly via modifications to the mixture model – for example, via classifier chains Chen et al. (2018); Gerych et al. (2021); Hartvigsen et al. (2020) on von Mises Fisher conditional distributions; learnable and anisotropic concentration parameters Roth et al. (2022b); or other geometric characterizations of the feature space. We push towards further exploration in the space of similarity learning in multi-modal multi-label settings.

REFERENCES

Antonios Anastasopoulos, Shankar Kumar, and Hank Liao. Neural language modeling with visual features. *arXiv preprint arXiv:1903.02930*, 2019.

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443, 2018.

Andrew L Beam and Isaac S Kohane. Big data and machine learning in health care. *Jama*, 319(13): 1317–1318, 2018.

Hedi Ben-Younes, Remi Cadene, Nicolas Thome, and Matthieu Cord. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 8102–8109, 2019.

Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, and Prateek Jain. Sparse local embeddings for extreme multi-label classification. *Advances in neural information processing systems*, 28, 2015.

Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1989–1998, 2019.

Wenming Cao, Qiubin Lin, Zhihai He, and Zhiquan He. Hybrid representation learning for cross-modal retrieval. *Neurocomputing*, 345:45–57, 2019. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2018.10.082. URL `https://www.sciencedirect.com/science/article/pii/S0925231219301407`. Deep Learning for Intelligent Sensing, Decision-Making and Control.

Shang-Fu Chen, Yi-Chen Chen, Chih-Kuan Yeh, and Yu-Chiang Wang. Order-free rnn with visual attention for multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Everest Hinton. A simple framework for contrastive learning of visual representations. 2020. URL `https://arxiv.org/abs/2002.05709`.

Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4685–4694, 2019. doi: 10.1109/CVPR.2019.00482.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013.

Weifeng Ge. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 269–285, 2018.

Walter Gerych, Thomas Hartvigsen, Luke Buquicchio, Emmanuel Agu, and Elke A Rundensteiner. Recurrent bayesian classifier chains for exact multi-label classification. *Advances in Neural Information Processing Systems*, 34:15981–15992, 2021.

Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020:191, 2020.

Jorge Gonzalez-Lopez, Sebastián Ventura, and Alberto Cano. Distributed selection of continuous features in multilabel classification using mutual information. *IEEE transactions on neural networks and learning systems*, 31(7):2280–2293, 2019.

Henry Gouk, Bernhard Pfahringer, and Michael Cree. Learning distance metrics for multi-label classification. In Robert J. Durrant and Kee-Eung Kim (eds.), *Proceedings of The 8th Asian Conference on Machine Learning*, volume 63 of *Proceedings of Machine Learning Research*, pp. 318–333, The University of Waikato, Hamilton, New Zealand, 16–18 Nov 2016. PMLR. URL `https://proceedings.mlr.press/v63/Gouk8.html`.

Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

Thomas Hartvigsen, Cansu Sen, Xiangnan Kong, and Elke Rundensteiner. Recurrent halting chain for early multi-label classification. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1382–1392, 2020.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016a.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Yonghao He, Shiming Xiang, Cuicui Kang, Jian Wang, and Chunhong Pan. Cross-modal retrieval via deep and bidirectional representation learning. *IEEE Transactions on Multimedia*, 18(7): 1363–1377, 2016b. doi: 10.1109/TMM.2016.2558463.

George Hripcsak, Patrick B Ryan, Jon D Duke, Nigam H Shah, Rae Woong Park, Vojtech Huser, Marc A Suchard, Martijn J Schuemie, Frank J DeFalco, Adler Perotte, et al. Characterizing treatment pathways at scale using the ohdsi network. *Proceedings of the National Academy of Sciences*, 113(27):7329–7336, 2016.

Tzu-Ming Harry Hsu, Wei-Hung Weng, Willie Boag, Matthew McDermott, and Peter Szolovits. Unsupervised multimodal representation learning across medical images and reports. *arXiv preprint arXiv:1811.08615*, 2018.

Kuan-Hao Huang and Hsuan-Tien Lin. Cost-sensitive label embedding for multi-label classification. *Machine Learning*, 106(9):1725–1746, 2017.

Xin Huang and Yuxin Peng. Cross-modal deep metric learning with multi-task regularization. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 943–948, 2017. doi: 10.1109/ICME.2017.8019340.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, pp. 4651–4664. PMLR, 2021.

Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):117–128, 2011.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Changsheng Li, Chong Liu, Lixin Duan, Peng Gao, and Kai Zheng. Reconstruction regularized deep metric learning for multi-label image classification. *IEEE transactions on neural networks and learning systems*, 31(7):2294–2303, 2019.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. *arXiv preprint arXiv:2107.07502*, 2021.

Ruizhi Liao, Daniel Moyer, Miriam Cha, Keegan Quigley, Seth Berkowitz, Steven Horng, Polina Golland, and William M Wells. Multimodal representation learning via maximization of local mutual information. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 273–283. Springer, 2021.

Weiwei Liu and Ivor W Tsang. Large margin metric learning for multi-label prediction. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015a.

Weiwei Liu and Ivor Wai-Hung Tsang. Large margin metric learning for multi-label prediction. In *AAAI*, 2015b.

Weiwei Liu, Donna Xu, Ivor W Tsang, and Wenjie Zhang. Metric learning for multi-output tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):408–422, 2018.

Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor Tsang. The emerging trends of multi-label learning. *IEEE transactions on pattern analysis and machine intelligence*, 2021a.

Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor Tsang. The emerging trends of multi-label learning. *IEEE transactions on pattern analysis and machine intelligence*, 2021b.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Andrew Maxwell, Runzhi Li, Bei Yang, Heng Weng, Aihua Ou, Huixiao Hong, Zhaoxian Zhou, Ping Gong, and Chaoyang Zhang. Deep learning architectures for multi-label classification of intelligent health risk prediction. *BMC bioinformatics*, 18(14):121–131, 2017.

T. Milbich, K. Roth, B. Brattoli, and B. Ommer. Sharing matters for generalization in deep metric learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020. doi: 10.1109/TPAMI.2020.3009620.

Timo Milbich, Karsten Roth, Samarth Sinha, Ludwig Schmidt, Marzyeh Ghassemi, and Björn Ommer. Characterizing generalization under out-of-distribution shifts in deep metric learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=_KqWSCu566`.

Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 360–368, 2017.

Kevin Musgrave, Serge J. Belongie, and Ser-Nam Lim. A metric learning reality check. *CoRR*, abs/2003.08505, 2020. URL https://arxiv.org/abs/2003.08505.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34, 2021.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4004–4012, 2016.

Kranti Parida, Neeraj Matiyali, Tanaya Guha, and Gaurav Sharma. Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3251–3260, 2020.

Subhojeet Pramanik, Priyanka Agrawal, and Aman Hussain. Omninet: A unified architecture for multi-modal multi-task learning. *arXiv preprint arXiv:1907.07804*, 2019.

Qi Qian, Lei Shang, Baigui Sun, Juhua Hu, Hao Li, and Rong Jin. Softtriple loss: Deep metric learning without triplet sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Karsten Roth, Timo Milbich, Samarth Sinha, Prateek Gupta, Bjorn Ommer, and Joseph Paul Cohen. Revisiting training strategies and generalization performance in deep metric learning. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8242–8252. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/roth20a.html.

Karsten Roth, Timo Milbich, Bjorn Ommer, Joseph Paul Cohen, and Marzyeh Ghassemi. Simultaneous similarity-based self-distillation for deep metric learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9095–9106. PMLR, 18–24 Jul 2021. URL http://proceedings.mlr.press/v139/roth21a.html.

Karsten Roth, Oriol Vinyals, and Zeynep Akata. Integrating language guidance into vision-based deep metric learning, 2022a. URL https://arxiv.org/abs/2203.08543.

Karsten Roth, Oriol Vinyals, and Zeynep Akata. Non-isotropy regularization for proxy-based deep metric learning, 2022b. URL https://arxiv.org/abs/2203.08547.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 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.

Xiaobo Shen, Weiwei Liu, Yong Luo, Yew-Soon Ong, and Ivor W Tsang. Deep binary prototype multi-label learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 2675–2681, 2018.

Senlin Shu, Fengmao Lv, Yan Yan, Li Li, Shuo He, and Jun He. Incorporating multiple cluster centers for multi-label learning. *Information Sciences*, 2022.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning, 2017.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, pp. 1857–1865, 2016.

Kihyuk Sohn, Wenling Shang, Xiang Yu, and Manmohan Chandraker. Unsupervised domain adaptation for distance metric learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BklhAj09K7.

Yan-Ping Sun and Min-Ling Zhang. Compositional metric learning for multi-label classification. *Frontiers Comput. Sci.*, 15:155320, 2021.

Eu Wern Teh, Terrance DeVries, and Graham W. Taylor. Proxynca++: Revisiting and revitalizing proxy neighborhood component analysis. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIV*, volume 12369 of *Lecture Notes in Computer Science*, pp. 448–464. Springer, 2020. doi: 10.1007/978-3-030-58586-0_27. URL https://doi.org/10.1007/978-3-030-58586-0_27.

Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

Kaiye Wang, Ran He, Liang Wang, Wei Wang, and Tieniu Tan. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2010–2023, 2015.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12695–12705, 2020.

Eric W. Weisstein. Hypersphere, 2002.

Minz Won, Sergio Oramas, Oriol Nieto, Fabien Gouyon, and Xavier Serra. Multimodal metric learning for tag-based music retrieval. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 591–595. IEEE, 2021.

Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2840–2848, 2017.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pp. 24043–24055. PMLR, 2022.

Pengtao Xie and Eric P Xing. Multi-modal distance metric learning. 2013.

Xing Xu, Li He, Huimin Lu, Lianli Gao, and Yanli Ji. Deep adversarial metric learning for cross-modal retrieval. *World Wide Web*, 22(2):657–672, March 2019. ISSN 1386-145X. doi: 10.1007/s11280-018-0541-x. URL https://doi.org/10.1007/s11280-018-0541-x.

Xiaodong Yang, Pavlo Molchanov, and Jan Kautz. Multilayer and multimodal fusion of deep neural networks for video classification. In *Proceedings of the 24th ACM international conference on Multimedia*, pp. 978–987, 2016.

Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6995–7004, 2021.

Julio Cesar Zaragoza, Enrique Sucar, and Eduardo Morales. A two-step method to learn multidimensional bayesian network classifiers based on mutual information measures. In *Twenty-fourth International FLAIRS Conference*, 2011.

Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, 2020.

Su-Fang Zhang, Jun-Hai Zhai, Bo-Jun Xie, Yan Zhan, and Xin Wang. Multimodal representation learning: Advances, trends and challenges. In *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, pp. 1–6. IEEE, 2019.

Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, and Thomas Brox. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1450–1459, 2021.

A DATASET METADATA

Dataset	Modalities	# Samples	# Labels	% Exclusive	Prediction Task
MIMIC-III	{T, TS}	30,558	25	9.8%	clinical phenotypes
MM-IMDB	{I, T}	25,959	23	23.0%	movie genre

Table 3: The table contains useful metadata for the datasets used in experiments. Experiments are conducted with the displayed four binary multi-label datasets: an unimodal dataset, MIMIC-CXR Johnson et al. (2019) and three multimodal datasets. In the above table, I denotes image data, T, text data, TS, time series data, and % exclusive describes the percentage of samples with exactly one positive label.

B BACKGROUND

B.1 DEEP METRIC LEARNING BACKGROUND

B.1.1 EVALUATION METRICS IN DEEP METRIC LEARNING

Definition 2 (Recall@k). *Jegou et al. (2011) Given $k \in \{1, \ldots, |X|\}$, define $NN_k : \Phi \subset \mathcal{S}^{D-1} \to \mathcal{P}(X)$ as a function that receives a point $\phi(x) \in \Phi$ and returns a set in the powerset of X, $\mathcal{P}(X)$, containing points in X that map to the k nearest neighbors of $\phi(x)$ in Φ. Then, Recall@k is measured as:*

$$\textit{Recall@k} = \frac{1}{|X|} \sum_{x \in X} \begin{cases} 1 & \exists \tilde{x} \in NN_k(x) : Y(\tilde{x}) = Y(x) \\ 0 & \textit{else} \end{cases}$$

B.1.2 PROXY LEARNING AS LIKELIHOOD MAXIMIZATION UNDER VON MISES-FISHER MIXTURE MODELS

The probability density function of the von Mises-Fischer distribution:

$$f_p(\mathbf{x}; \mu, \kappa) = C_p(\kappa) \exp(\kappa \mu^\intercal \mathbf{x}) \tag{8}$$

where $\kappa \geq 0$ the concentration constant, $\|\mu\| = 1$ and normalization constant

$$C_p(\kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\kappa)} \tag{9}$$

where I_v denotes the modified Bessel function of the first kind at order v. Idea: project ϵ-balls on the hypersphere (modulate via concentration parameter κ and normalization constant). Note that when $\kappa = 1$ and normalization constant $C = 1/N$, where $N = |\mathcal{Y}|$ denotes the number of classes, the distribution expresses Proxy-NCA loss via a maximum-likelihood problem under the von Mises-Fischer distribution.

Assuming μ_k to denote some class concept or class prototype, the probability of assigning a sample representation ϕ_i to μ_k is given as $p(\phi_i|\mu_k) = f_p(\phi(\mathbf{x}_i); \mu_k, \kappa_k)$. This can be extended to the vMF mixture model when multiple classes (class prototypes) are available:

$$p_{\text{vMFmm}}(\phi_i|\mu_k) = \frac{\pi_k C_d(\kappa_k) e^{\kappa_k s(\phi_i, \mu_k)}}{\sum_{\mu^* \in \mathcal{M}} \pi_{k^*} C_d(\kappa_{k^*}) e^{\kappa_{k^*} s(\phi_i, \mu_{k^*})}} \tag{10}$$

$$C_d(\kappa) = \kappa^{d/2-1} \cdot \left[(2\pi)^{d/2} I_{d/2-1}(\kappa)\right]^{-1} \tag{11}$$

where each class is defined by a unique prototype μ_k and some (potentially fixed) concentration κ_k, with overall probability of assignment controlled by the mixture degrees π_k. Assuming indeed a fixed κ_k and constant mixture π_k, it is easy to recover the Proxy-NCA loss through likelihood maximization of the mixture model.

Definition 3 (Proxy-NCA). *Kim et al. (2020) ProxyNCA learns class proxies, or class centers, which each represent a class in the set of unique classes $\mathcal{Y}$. Then, each anchor from the batch is sampled and a positive or negative proxy $\psi_c \in \mathbb{R}^d$ per class $c \in \mathcal{Y}$ is introduced in lieu of a positive or negative sample, respectively, giving:*

$$\mathcal{L}_{proxy} = -\frac{1}{b}\sum_{x_i \in \mathcal{B}} \log\left(\frac{\exp\left(-d\left(\phi(x_i), \psi_{Y(x_i)}\right)\right.}{\sum_{c\in\mathcal{Y}\setminus\{Y(x_i)\}} \exp\left(-d\left(\phi(x_i), \psi_c\right)\right.}\right)$$

B.1.3 Simultaneous Similarity-based Self-Distillation for Multimodal Metric Learning

S2SD was proposed as an extension to standard metric learning, which utilizes a multi-level distillation setup. Assuming a low-dimensional target dimensionality of the main embedding space $\Phi \in \mathbb{R}^d$ and access to a higher-dimensional, shared feature extraction network and respective (batched) feature representations $\Phi_\text{f} \in \mathbb{R}^{d_f}$, S2SD generates a sequences of n increasingly higher-dimensional embedding vectors $\{\Phi_{g_i}\}_{i=[0,n-1]}$ with $|\Phi_{g_i}| < |\Phi_{g_j}|$ for $i < j$. To maximize the transfer capabilities of the primary, low-dimensional embedding space Φ, the S2SD objective is then defined as

$$\mathcal{L}_\text{S2SD} = \frac{1}{2}\cdot\left[\mathcal{L}_\text{DML}(\Phi) + \frac{1}{n}\textstyle\sum_{i=1}^n \mathcal{L}_\text{DML}(\Phi_{g_i})\right] + \frac{\gamma}{n}\sum_{i=1}^{n}\mathcal{L}_\text{dist}(D^\Phi, D^{\Phi_{g_i}}) + \gamma\mathcal{L}_\text{dist}(D^\Phi, D^f) \quad (12)$$

with $D^f, D^{\Phi_{g_i}}, D^\Phi$ denoting batch-wise similarity matrices, e.g. $D^\Phi_{i,j} = \Phi_i^T\Phi_j$ over the high-dimensional feature vectors Φ_f, the additional S2SD embeddings Φ_{g_i} and the main embedding space Φ, respectively.

B.2 Pearson Correlation Coefficient and Euclidean Distance

Let $u, v \in \mathbb{R}^k$ be vectors with zero mean and unit length (*i.e.,* $\|u\| = \|v\| = 1$, $\overline{u} = \overline{v} = 0$). The Pearson Correlation Coefficient r between u and v can thus be written as:

$$r = \sum_{i=1}^{k}(u_i - \overline{u})(v_i - \overline{v})\sqrt{\sum_{i=1}^{k}(u_i - \overline{u})^2}\sqrt{\sum_{i=1}^{k}(v_i - \overline{v})^2} = \frac{\sum_{i=1}^k u_i v_i}{\sqrt{\sum_{i=1}^k u_i^2}\sqrt{\sum_{i=1}^k v_i^2}} = \sum_{i=1}^{k} u_i v_i$$

We can further rewrite the euclidean distance between u and v as:

$$d = \sqrt{\sum_{i=1}^{k}(u_i - v_i)^2} = \sqrt{\sum_{i=1}^{k} u_i - 2\sum_{i=1}^{k} u_i v_i + \sum_{i=1}^{k} v_i} = \sqrt{2 - 2\sum_{i=1}^{k} u_i v_i}$$

Thus, $(1 - d^2/2) = 1 - \frac{2-2\sum_{i=1}^k u_i v_i}{2} = \sum_{i=1}^k u_i v_i = r$.

C Additional Results

C.1 Varying Label Correlation Weight

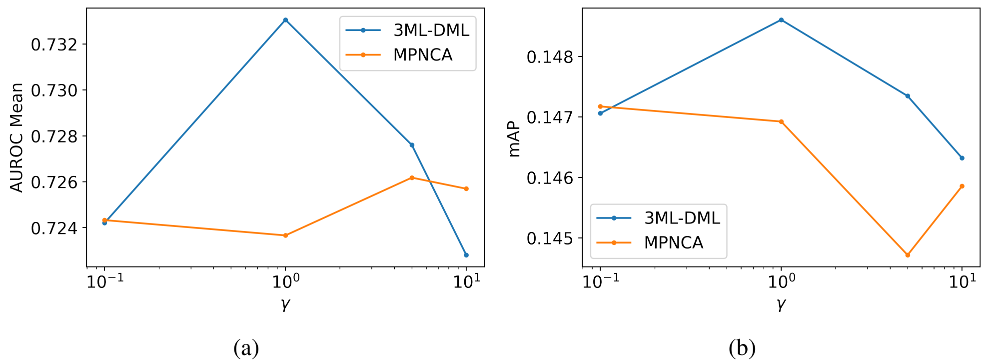

Figure 2: Comparison of model performance for `3ML-DML` and regular `MPNCA` as a function of the label correlation loss weight γ, as measured by (a) mean linear AUROC, and (b) mAP.