LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-CONF-788499

From Intra-Modal to Inter-Modal Space: Multi-Task Learning of Shared Representations for Cross-Modal Retrieval

J. Choi, M. Larson, G. Friedland, A. Hanjalic

August 29, 2019

IEEE International Conference on Multimedia Big Data
Singapore, Singapore
September 11, 2019 through September 13, 2019

Disclaimer

FROM INTRA-MODAL TO INTER-MODAL SPACE: MULTI-TASK LEARNING OF SHARED REPRESENTATIONS FOR CROSS-MODAL RETRIEVAL

Jaeyoung Choi[1,3], Martha Larson[1,2], Gerald Friedland[4], Alan Hanjalic[1]

[1]Delft University of Techonology, Netherlands
[2] Radboud University, Netherlands
[3] International Computer Science Institute, USA
[4]University of California at Berkeley, USA

ABSTRACT

Learning a robust shared representation space is critical for effective multimedia retrieval, and is increasingly important as multimodal data grows in volume and diversity. The labeled datasets necessary for learning such a space are limited in size and also in coverage of semantic concepts. These limitations constrain performance: a shared representation learned on one dataset may not generalize well to another. We address this issue by building on the insight that, given limited data, it is easier to optimize the semantic structure of a space within a modality, than across modalities. We propose a two-stage shared representation learning framework with intra-modal optimization and subsequent cross-modal transfer learning of semantic structure that produces a robust shared representation space. We integrate multi-task learning into each step, making it possible to leverage multiple datasets, annotated with different concepts, as if they were one large dataset. Large-scale systematic experiments demonstrate improvements over previously reported state-of-the-art methods on cross-modal retrieval tasks.

Index Terms— cross-modal retrieval, multi-task learning, video retrieval, image retrieval

1. INTRODUCTION

Cross-modal representation learning utilizes semantic correlation among multiple modalities in multiple datasets (e.g., containing videos, images, and text). Cross-modal retrieval between image and text modalities have been studied extensively [1, 2, 3, 4] and video-text cross-modal retrieval has recently become a hot research topic [5, 6, 7, 8]. These approaches try to learn projections of data of different modalities into a shared space. This space simplifies retrieval because cross-modal similarity is calculated as the similarity between shared representations.

Cross-modal representation learning leverages semantic correlation among multiple modalities, hence the quality of alignment in the dataset is crucial. However, the alignment between different modalities in multimodal datasets is often imperfect. For instance, sentences describing an image, often do not convey a complete 'picture' of the image; some elements of the image may be missing in the annotations, and some annotations may fail to be well-represented in the image. These issues can occur even when expert annotators attempt to formulate sentences that capture image content as fully as possible.

The lack of truly large-scale aligned, multimodal datasets presents a further challenge. Models for learning cross-modal representations are often trained and evaluated on labeled datasets that are small, in either the number of data samples that they contain, the number of semantic labels that they cover, or both. In such cases, the model that is learned can easily overfit to one specific dataset that was used to train it. Such a model would be limited in its ability to generalize, since fine-tuning the model to another domain would cause it to lose the structure of the original space. The high-cost of annotating aligned, cross-modal datasets makes it unlikely that the problem of limited data availability will be solved anytime soon.

In order to address these issues, we propose, in this paper, a two-stage shared representation space optimization strategy. Under our strategy, the semantic space of each modality is first optimized individually. This intra-modal optimization is followed by inter-modal optimization, which cross-transfers intra-modal semantic structure to a joint semantic space. To address the challenge of the imperfect cross-modal alignment of paired data, our model uses a bi-directional quadruplet loss function that takes two pairs of aligned data as input and jointly optimizes the cross-modal semantic relationship and the inter-modal invariance in the joint space. The net effect is that the approach cumulatively compensates for imperfect alignments between modalities. We show that modality-wise pre-optimization of semantic space is a crucial step for learning a more discriminative joint semantic structure.

Further, in order to address the challenge of overfitting and loss of generalizability stemming from the shortage of aligned data, we integrate multi-task learning in each optimization step. Multi-task learning is inspired by the way in

which humans learn as a natural activity. Specifically, people apply knowledge learned from previous tasks to help learn a new task [9]. Rather than learning representations based on a single dataset with single task, representations should become more general as more data are used to learn them. Our proposed multi-task learning approach for representation learning leverages supervised data from cross-task datasets with multiple modalities. This effectively gives the model access to a larger pool of collective data, which leads to learning a more generalized discriminative single-modal semantic space for each modality. Learning of the cross-modal representation space can, in turn, benefit from transferring such generalized intra-modal semantic structure. In our work, the learning of the weighting of individual task loss is inspired by Bayesian deep learning [10, 11]. The uncertainty of the multi-task loss is homoscedastic in nature. Since the homoscedasticity is task (or dataset) dependent, we can infer the weighting of the task loss from the observable uncertainty, which takes the form of noise.

In sum, the main contributions of this work are as follows:

1. We propose a novel shared representation space optimization strategy. The semantic space of each modality is optimized beforehand, followed by inter-modal optimization with the objective of transferring intra-modal semantic structure.

2. We integrate multi-task learning into our proposed framework to leverage multiple data sets as if they were one large data set, in order to learn a robust joint semantic representation for video, image and text.

3. We show that bi-directional quadruplet loss is effective at cross-modal transfer of intra-modal semantic structure.

4. We demonstrate the effect of deep contextualized word representations [12] in cross-modal retrieval task.

The remainder of the paper is organized as follows. We overview the related works on cross-modal retrieval and multi-task learning in Section 2. Section 3 elaborates the proposed approach. Section 4 describes the experimental setting and presents the results, amd we analyze the result on section 5. The conclusions are made in Section 6.

2. RELATED WORK

In this section, we introduce the two lines of research that are most-closely related to our work: cross-modal retrieval and multi-task learning.

2.1. Cross-modal Retrieval

Image-Text Retrieval

The general strategy of cross-modal retrieval is to measure similarities among data of different modality or media type and to attain a representation space in which the intra-class variation is minimized, the inter-class variation is maximized, and the difference of each data pair captured from two modalities of the same class is minimized. There are Canonical correlation analysis (CCA)-based approaches [13, 1] that use the principle of learning a common space that maximizes the pairwise correlation between bi-modal pairs of data (e.g., image and text). Recently, DNN-based approaches have become popular in the research community. Inputs from different modalities are coded to obtain a shared representation through a shared layer in deep neural networks [2, 3]. Fast0Tag [14] projects an image by identifying a principal direction in the space and targeting that principal direction when learning to project the image. [15] uses noise contrastive estimation on a noisy web-scale dataset [16] to learn projection from image to word embeddings space. VSE++ [17] proposes a modified pairwise ranking loss weighted by violation caused by hard-negatives. ACMR [18] introduced adversarial loss to cross-modal retrieval to learn an embedding that is ignorant of the input modality. Cross-modal Transfer Learning [19, 20, 21] tries to leverage a auxiliary large-scale single modality dataset and transfer its semantic structure to cross-modal training.

Video-Text Retrieval

When compared to image datasets, the size of video datasets with supervised aligned captions available for cross-modal training is far from being sufficient. MSR-VTT [22], which has the largest number of aligned captions (200,000), has only 10,000 videos. LSMDC [23], the largest video dataset with aligned caption, has 118,081 video clips. The scale is an order of magnitude smaller when compared to image datasets. Word2VisualVec [7] projects sentences into visual space with mean squared loss under an assumption that visual space is semantically better structured. [24] uses web image search results of the input text. Other works [7, 24] uses mean-pooled features from video frames. [25] propose an LSTM with visual-semantic embedding that jointly optimizes a contextual loss to learn the relationship of words and a relevance loss to create a visual-semantic embedding space by reflecting the relationship between the semantics of the sentence and visual content. [26] learns a shared space across image, text, and sound modality by using student-teacher model and ranking loss. The work most related to ours is by Mithun et al [5], where they propose a modified bi-directional pairwise ranking loss and multi-modal features by leaning deep representations for object-text and activity-text subtasks and applying fusion strategy. Different from theirs, our input feature is a combination of object, activity, and sound feature and we learn a single shared space between visual and text modality. We also apply

2.2. Multi-task Learning

In much previous work, a neural network is trained on a large dataset first and then fine tuned on smaller ones for specific tasks. For an industrial or practical use cases, having a single model that can perform well on datasets from many different domain is very important.

Multi-task learning (MTL) has been successfully used across many applications of machine learning [27] to address the above-mentioned issues: loss of generalizability and shortage of annotated datasets. A growing amount of recent work has shown the effectiveness of multi-task learning for representation learning in many domains including computer vision [11, 28, 29], NLP [30, 31, 32] and other domains [27, 33, 19]. Improved generalization comes from leveraging domain-specific information in the training data of related tasks [9]. MTL biases the model to prefer representations that other tasks also prefer [34, 31, 30, 28, 33, 29].

The scope of multi-task learning can be very broad. As long as the model optimizes more than a single loss function, such as when optimizing ranking loss and contrastive loss together, it is implicitly doing MTL. In cross-modal retrieval domain, Huang et al. [19] integrates ranking loss and contrastive loss for modeling cross-modal semantic similarity. Wang et al. [18] uses weighted loss consisting of inter-modal triplet loss, intra-modal contrastive loss, and adversarial loss as another implicit MTL.

However, applying multi-task learning is not trivial. For instance, naively combining different datasets would not be effective as some concepts or labels are biased and represented with more examples, while some datasets are noisier than others due to the inherent nature of the source, varying quality of annotators, etc. How to weigh different loss functions is one of several challenges we face when applying MTL. [11] proposes to use uncertainty to weigh losses in MTL by learning another noise parameter that is integrated in the loss function for each task. This makes it possible to use multiple tasks, possibly regression and classification, and to bring all losses to the same scale.

The main novelty of our work that differentiates us is that we utilize three modalities—video, image, and text—to simultaneously optimize intra-modal and inter-modal semantic structure using uncertainty-based weighting to handle loss scale issue between datasets and tasks.

3. PROPOSED APPROACH

3.1. Problem Formulation

In our work, we use three modalities but this can be extended to any number of modalities. For the sake of simplicity, and without losing generality, we show detailed formulation of handling video and text pairs.

Let $V = [v_1, v_2, ..., v_n] \in \mathbb{R}^{d_v \times n}$ be a collection of video features, and $T = [t_1, t_2, ..., t_n] \in \mathbb{R}^{d_t \times n}$ be the associated text features where v_i and t_i form an input pair and d_v and d_t are the dimensions of the video and text features, respectively. Datasets with supervised semantic labels have a label vector denoted as $y_i = [y_{i1}, y_{i2}, ..., y_{iC}] \in \mathbb{R}^C$, where $c \in C$ denotes semantic label class where

$$y_{ij} = \begin{cases} 1 & \text{if the pair}(v_i, t_i) \text{ is assigned with class label } j \\ 0 & \text{otherwise} \end{cases}$$

Our objective is to learn a shared representation space where the semantic similarity between projected features from V and T can be directly compared with some distance metric in the learned space.

3.2. Intra-modal Representation Space Optimization

We first optimize each modality's semantic structure before optimizing cross-modal shared representation space. Siamese and triplet networks have shown to be useful for learning mappings from image to a compact Euclidean space where distances correspond to a measure of similarity [35, **?**]. Embeddings trained in such way can be used as features vectors for classification or few-shot learning tasks.

For the sake of simplicity, we give detailed problem formulation in image modality only. Other modalities are optimized in a similar fashion. Given an anchor image i_a and its associated label vector, y_a, we can find positive sample $i_p \; s.t. \; y_a \cdot y_p \neq 0$, and negative sample $i_n \; s.t. \; y_a \cdot y_n = 0$. The objective is to learn embeddings such that the anchor image i_a is closer to the positive example i_p than it is to the negative example i_n by some margin m. The loss function is formulated as:

$$L(i_a, i_p, i_n) = \max(0, m + |f(i_a) - f(i_p)|_2^2 - |f(i_a) - f(i_n)|_2^2) \quad (1)$$

where $f(\cdot)$ denotes nonlinear projection of the input modality.

Note that the number of possible triplets grows cubically with the number of examples, and is therefore infeasible to train using all combinations. Using hard-negatives has shown to be effective in many embedding tasks [17]. For a positive pair (i_a, i_p), the hardest negative sample is defined as $\hat{i} = \arg\max_{i_n} D(i_a, i_n)$, where $D(\cdot, \cdot)$ is the distance between two inputs in the projected space, i.e., $D(i_a, i_n) = |f(i_a) - f(i_n)|_2^2)$.

The optimization of parameters of the model (θ) can be written as following:

$$\min_\theta \sum_{i_a} [m - D(i_a, i_p) + D(i_a, \hat{i})] \quad (2)$$

In practice, we use a semi-hard negative sampling [17] by finding the hardest-negative sample within a mini-batch at each iteration instead of comparing against the entire training set due to computational efficiency. Semi-hard negative training has shown to provide some regularization effect as well.

3.3. Cross-modal Optimization with Bi-directional Quadruplet Loss

We have already optimized intra-modal semantic structure using triplet loss with semi-hard negative sampling, and now we optimize the inter-modal semantic structure using quadruplet loss allowing us to utilize the learned semantic similarity in the previous step. Without losing generality, we use video and text modalities to formulate the problem. Here, we use the cosine similarity function $S(\cdot,\cdot)$ between two vectors in the shared space. Given a pair of videos $(v_i, v_j) \in V \times V$ and their corresponding pair of text annotation $(t_i, t_j) \in T \times T$ where $i \neq j$, we utilize the intra-modal semantic structure within V by minimizing the distance between the difference between two similarities; similarity from t_i to v_i and v_j, respectively, and similarity from v_i to v_i(itself) and v_j, respectively. The intuition is that we want t_i to be projected in the shared space where its semantic relationship to projections of v_i and v_j is similar to intra-modal relationship between v_i and v_j. By applying this similarly to T and using t_j as the anchor, the bi-directional quadruplet loss is defined as:

$$\begin{aligned} L(v_i, v_j, t_i, t_j) = \\ |(S(v_i, t_i) - S(v_j, t_i)) - (S(v_i, v_i) - S(v_i, v_j))| \\ + |(S(t_j, v_j) - S(t_j, v_i)) - (S(t_j, t_j) - S(t_j, t_i))| \end{aligned} \quad (3)$$

where $S(v, v) = 1$. This can be re-written in a more intuitive form :

$$\begin{aligned} L(v_i, v_j, t_i, t_j) = |(S(v_i, t_i) - 1) + (S(v_i, v_j) - S(t_i, v_j))| \\ + |(S(t_j, v_j) - 1) + (S(t_j, t_i) - S(t_j, v_i))| \end{aligned} \quad (4)$$

The loss function is pushing v_i and t_i together while making v_i and t_i to have the same similarity with v_j, and pushing v_j and t_j together while making t_i and v_i to have the same similarity with t_j.

3.4. Multi-task Loss

The multi-task loss function is defined as follows:

$$L(x; \theta; \lambda) = \sum_{i=0}^{T} \lambda_i \, L_i(x; \theta) \quad (5)$$

where x is a set of training data, θ is the network parameters learned by minimizing $L()$, T is the total number of tasks (or datasets). We want to optimize λ_i which controls the weight of each loss. The naive approach would be to set the weighting to be equal. Much existing work uses exhaustive hyper-parameter search or a heuristic approach [36] to find weights when combining multiple losses. Since we need to handle many datasets and tasks in our multi-task learning framework, it would become very expensive to find the weighting using a greedy or grid search.

Recently, Bayesian deep learning approaches [10, 11] have shown that it is possible to learn another noise parameter that is integrated into the loss function for each task. This makes it possible to bring all losses to the same scale.

The classification likelihood of a Bayesian probabilistic model output is (cross-entropy loss for classification)

$$p(y|f^W(x), \sigma) = Softmax(1/\sigma^2 f^W(x)) \quad (6)$$

where $f^W(x)$ is the output of the neural network and W is the weights on input x. σ is the observation noise. The log likelihood of Eq. 6 is

$$\begin{aligned} log(p(y = c|f^W(x), \sigma)) = 1/\sigma^2 f_c^W(x) - \\ \log(\sum_{i=0}^{C} \exp(1/\sigma^2 f_i^W(x)))) \end{aligned} \quad (7)$$

where C is the number of classes. We can derive similarly for regression loss as shown in [11].

Converting k-Tuple Loss as Regression Loss

We can convert triplet loss and quadruplet loss to a k-way (k = 3, 4) regression loss function. Given a training dataset containing N triplets $(x_1, y_1, z_1), (x_2, y_2, z-2), ...(x_n, y_n, z_n)$ and their corresponding outputs $(d_1, d_2, ...d_n)$, the triplet loss can be formulated as trivariate regression function as following:

$$f(x_i, y_i, z_i) = d_i \in [0, 2+m] = [m+D(x_i, y_i) - D(x_i, z_i)]_+ \quad (8)$$

where $D(\cdot,\cdot)$ is the distance function, $[\cdot]_+$ is a soft margin function, and m is the margin between embeddings.

3.5. Training Procedure

The training procedure of our framework consists of two stages: intra-modal optimization and inter-modal optimization. One of the main objectives of our method is to leverage larger number of samples from multiple datasets. Within each modality, and within each dataset, we group samples into mini-batches and merge them together in one queue $\bigcup D_t$. In each epoch, we iterate minibatch b_t from $\bigcup D_t$ and compute loss for dataset t.

In the multi-task inter-modal optimization stage, we use mini-batch based stochastic gradient descent to learn the parameters of all the shared layers and dataset-specific layers as shown in Algorithm 1. First, similar to intra-modal optimization, samples from each dataset are packed into mini-batches. Then we pack these mini-batches into a chunk so that each chunk has one mini-batch from each dataset. In each iteration, we go through mini-batches in each chunk, and the model parameters are updated by weighted sum of loss. As an ablation study, we also train the model by jointly optimizing intra-modal and inter-modal loss, training with triplet loss instead of quadruplet loss for inter-modal optimization; and without multi-task learning.

Algorithm 1 Multi-Task Learning of shared representation Space

Initialize the parameters with random values
for m in $1, 2, ..., M$ **do** ▷ M modalities
  for t in $1, 2, ..., T_M$ **do** ▷ T_m tasks in modality m
    Pack the dataset t into mini-batch D_t
  for $epoch$ in $1, 2, ..., epoch_{max}$ **do**
    for b_t in $\bigcup D_t$ **do** ▷ b_t is a minibatch of dataset t
      Compute Weighted Loss : $L(\Theta)$ = Eq. 1
      Compute gradient: $\nabla(\Theta)$
      Update model parameters : $\Theta = \Theta - \epsilon\nabla(\Theta)$
for t in $1, 2, ..., T$ **do** ▷ Prepare the data for inter-modal opt. with T datasets
  Pack the dataset t into mini-batch D_t
for $epoch$ in $1, 2, ..., epoch_{max}$ **do**
  for b_t in $\bigcup D_t$ **do** ▷ b_t is a minibatch of dataset t
    Compute Weighted Loss: $L(\Theta)$
      $L(\Theta)$ = Eq. 4
    Compute gradient: $\nabla(\Theta)$
    Update model parameters: $\Theta = \Theta - \epsilon\nabla(\Theta)$

4. EXPERIMENTS AND RESULTS

4.1. Datasets

We perform our experiments on four images and two video datasets widely used in the related work and for benchmarking. We summarize the characteristics of the image-text datasets and video-text datasets and give an overview in Table 1 and 2. In the training phase, only the training and validation sets are used for generating vocabulary dictionary and learning the models. Test images, videos, and text are only for test and not used in any stage of the training.

4.1.1. Image-Text Datasets

Wikipedia dataset [37] is a widely-used cross-modal dataset, which includes text and images selected from featured articles in Wikipedia. Articles are filtered to have only sections that contain a single image and at least 70 words. There are 2,866 documents in total, with 2,173 pairs as training set and 693 pairs as testing set. The median text length is 200 words.
IAPRTC-12 dataset [38] consists of 19,805 images featuring various domains, such as landscapes, portraits, indoor and sports scenes. Each image is annotated with a description of one to three sentences. We used 17,825 images for training, and 1,980 images for testing.
NUS-WIDE NUS-WIDE [39] is a web image dataset that includes 269,648 images and corresponding tags from Flickr. There are 5,018 unique tags. The dataset has six different types of low-level features extracted from the images as well as bag-of-words SIFT descriptions for the tags. We use the official train/test split provided by authors: 161,789 images for training and 107,859 for testing.
XMediaNet dataset [40] is a large-scale dataset of texts, images, videos and audios and 3D models. Categories in the dataset are chosen from WordNet and thus they have semantic hierarchy structure. The categories are divided into two types: animals, such as elephant, owl, bee, and artifacts, such as violin, airplane or camera. We used 32,000 texts and images for training and 8,000 texts and images for testing.

4.1.2. Video-Text Datasets

MSR-VTT dataset [22, 41] is a large-scale video dataset with text description annotation. The dataset contains 10,000 video clips with a split of 6,513 for training, 2,990 for testing, and 497 for validation. Each video has 20-sentences descriptions.
MSVD dataset [42] is a another video dataset that contains 1,970 YouTube video clips. Each video has around 40 sentences as annotation. We used English descriptions only. We used the split of 1,200 for training, 100 for validation, and 670 for testing.

4.2. Input Feature Representation

Next, we present our features, providing brief descriptions.

Text Features For initial input embedding, recent work uses pre-trained dense word embeddings, such as fastText [43]. Recently released pre-trained neural models such as ELMo [44] and BERT (Bidirectional Encoder Representations from Transformer) [12] have shown impressive results on how well models can handle various language-based tasks. Traditional word embeddings provide context-free embeddings (static) whereas ELMo or BERT gives contextualized embeddings (dynamic). For instance, given two sentences, "Am I supposed to wear a belt with a suit?", "We found a new asteroid belt", FastText would give the same embedding for the word *belt* in both sentences whereas BERT will give different ones depending on the context. We used BERT [12] to encode our text annotations, and compare its performance against FastText which is a static embedding. We take the second-to-last hidden layer of all of the tokens in the sentence and do average pooling for sentence level embedding. For an image-text pair that has multiple sentences, its document-level embedding is obtained by average pooling over sentence-level embeddings. We use second-to-last layer instead of the last one, as the last layer may be biased because of its proximity to the target functions (masked language model and next sentence prediction) during the pre-training of the model.
Image Features We used a convolutional neural network (CNN) pre-trained on ImageNet dataset to encode images. For extracting features from images, off-the-shelf pre-trained networks [45] have been widely used. These features

Table 1. Comparison of the image datasets used in our experiments.

Dataset	# of Images (train/test)	Task Type	Image Type	Text Type	Img-Txt Alignment (# of text instances per image)	Label Type	#Label (average number per image if multi-label)
Wikipedia	2,173 / 693	Single-label	Wikipedia Image	sentence from Wikipedia article	22.61	Wikipedia Categories	10
NUS-WIDE	161,789/ 107,859	Multi-label	Flickr	noisy tags annotated by user	11.8 (avg), 9 (median)	Semantic Concepts	81
IAPRTC-12	17,646 / 1,980	Multi-label	still, natural images	sentences that, describe the image	1.76	Image Segmentation Labels	249 (avg. 4.14)
XMediaNet	32,000 / 8,000	Single-label	Flickr	sentences from Wikipedia article	10	Animals (48), Artifacts (152)	194

Table 2. Comparison of the video datasets used in our experiments.

Dataset	# of Videos (train/val/test)	# sentences per video	context
MSR-VTT	6513 / 497 / 2990	20	20 categories
MSVD	1200 / 100 / 670	around 40	multi-category

represent the objects and their relationships in an image. Specifically, image features were extracted directly from last average pooling layer of ResNet [45] with 101 layers (ResNet-101). Images were rescaled to 224x224 for input. The dimension of the image feature is 2048.

Video Features We used Inception-v1 I3D model trained on the Kinetics dataset [46] to encode activity depicted in videos and mean-pooled global average layer of ResNet-101 from frames as object feature. We also extract audio features using attention neural networks [47] trained on two million samples (527 classes) of Audio Set [48]. The dimension of the audio feature is 128. In this work, we use the fusion of activity, object and audio features as an video input embedding.

4.3. Network Architecture

Sub-networks convert input feature representations from each modality (text, image, and video) to representations of the same dimensions, and this shared representation layer is used as input to the metric learning network.

For encoding the image, we take CNN feature and use two fully connected layers. Concatenated fusion of activity, object, and audio feature goes through two fully connected layers. Average pooled input text representation goes through one fully connected layer.

4.4. Implementation Details

Our implementation used PyTorch [49]. Representations are indexed with FAISS [50] with 'IndexFlatL2' index class for exact nearest neighbor search with L2 distance. The model was trained with two p3.2xlarge instances (NVIDIA Tesla V100 GPU). For pre-trained BERT models, we used an open-source Pytorch implementation available online[1].

4.5. Metric

We use mean average precision (mAP) and precision@k for evaluation of Image-Text retrieval. Given a set of queries, mAP is defined as:

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q} \tag{9}$$

where Q is the number of queries in the set and $AP(q)$ is the average precision

$$AP(q) = \frac{\sum_{k=1}^{R} P(k)\delta(k)}{\sum_{j=1}^{R} \delta(j)} \tag{10}$$

where R is the size of the retrieved result, and $\delta(k) = 1$ if the k-th result is relevant, otherwise 0. $P(k)$ is the precision of the result at k-th position. Following standard practices, a result is considered relevant to a query if at least one class label is shared between them [51]. The similarity between the query and each test sample is computed with the cosine similarity. The intuition behind AP is to penalize models that are not able to sort true positives to lead at the top of the retrieved ranked list. We report the mAP performance of both retrieval directions, image-to-text ($I \rightarrow T$) and text-to-image ($T \rightarrow I$).

For video-text retrieval, we adopt rank-based metric, R@K, Median Rank and Mean Rank. R@K (Recall@K) measures the percentage of test samples for which the correct result is found in the top-K retrieved points to the query. Median Rank is the median of the ground-truth results in the ranking and Mean Rank is the mean rank of all correct results.

[1] https://github.com/huggingface/pytorch-pretrained-BERT

	Wikipedia		IAPRTC		NUS-WIDE		XMediaNet	
	I→T	T→I	I→T	T→I	I→T	T→I	I→T	T→I
Wang et al. 2014	0.187	0.179	-	-	-	-	-	-
Wang et al. 2017	0.468	0.412	-	-	0.519	0.542	-	-
Peng et al. 2018	0.537	**0.485**	-	-	0.556	0.584	-	-
Ours (Sta. Embedding)	0.430	0.365	0.465	0.516	0.493	0.507	0.378	0.395
Ours (Sta. + Pre-opt.)	0.482	0.392	0.489	0.522	0.548	0.532	0.416	0.436
Ours (Sta. + Pre-opt + MTL)	0.502	0.444	**0.523**	0.535	**0.586**	**0.603**	0.467	0.498
Ours (Dyn. Embedding)	0.423	0.381	0.426	0.519	0.432	0.398	0.472	0.402
Ours (Dyn. + Pre-opt)	0.527	0.435	0.466	0.522	0.469	0.430	0.479	0.488
Ours (Dyn. + Pre-opt + MTL)	**0.541**	0.473	0.478	**0.536**	0.481	0.483	**0.497**	**0.517**

Table 3. mAP comparison with existing image-text retrieval methods. *Sta. Embedding* and *Sta.* denotes the use of static text embedding [43], *Dyn. Embedding* and *Dyn.* denotes the use of dynamic text embedding [12], and *Pre-opt.* denotes pre-optimization of intra-modal semantic space

4.6. Results

Image-Text Retrieval Results In Table. 3, we compare our approach with existing image-text retrieval approaches [51, 18, 40] on Wikipedia and NUS-WIDE dataset. Our approach with combinations of either the static or dynamic embedding, pre-optimization and multi-task learning shows better performance in image-to-text (I→T) retrieval task on Wikipedia and NUS-WIDE dataset, and in text-to-image (T→I) retrieval task on NUS-WIDE dataset. No single setup outperforms on every task and datasets. Peng et al [40] showed better score on Wikipedia dataset in text-to-image retrieval task.

Video-Text Retrieval Results In Table. 4, we compare our approach with three existing video-text retrieval approaches [26, 24, 5] along with a baseline approach (mean-pooled ResNet). On MSR-VTT dataset, our approach with a combination of dynamic embedding, pre-optimization and multi-task learning outperforms all existing works in both video-to-text and text-to-video retrieval task. On MSVD dataset, our approach performs better or equally well with existing works on both video-to-text and text-to-video retrieval task, except for Recall@1 metric on video-to-text task.

The effect of multi-task learning, and optimization strategy is discussed below in Section 5.

5. DISCUSSION

In this section, we discuss how the specific aspects of our experimental results support the contributions of the paper, which were given in Section 1.

5.1. Pre-optimization of intra-modal space

Pre-optimization of intra-modal space outperformed the joint optimization of modalities by a large margin in all datasets with both static and dynamic embedding for textual modality. With intra-modal pre-optimization, embeddings in each modality are maximally separated between class and clustered within class before learning the cross-modal transfer mapping. Since intra-modal discriminability is already optimized, the cross-modal transfer learning becomes easier to train as it only needs to optimize a good projection into the shared space without having to additionally optimize the discriminability of mapped items in the shared space.

It seems that the benefit of separate optimization can be maximized with many other datasets that have supervised semantic class labels that directly refers to contents. Datasets that do not have manual annotations such as YFCC100M [16] may see limited effect from using separate optimization.

5.2. Effect of bi-directional quadruplet loss for inter-modal optimization

Our bi-directional quadruplet loss outperforms triplet loss in all datasets with pre-optimization of intra-modal space. It shows that our proposed quadruplet loss effectively exploits optimized intra-modal space. We try to attain a representation space which the intra-class variation is minimized, the inter-class variation is maximized, and the difference of each data pair captured from two modalities of the same class is minimized. With our two-stage optimization strategy, the first two conditions are optimized with the pre-optimization of intra-modal space, and the proposed quadruplet loss for inter-modal optimization addresses the third condition. In addition, negative effect of semantic misalignment often present in paired dataset is mitigated from pushing embeddings of two modalities from a pair to have the same similarity with embeddings of both modalities from another pair.

5.3. Multi-task learning for intra-modal semantic structure optimization

The result shows that leveraging more data with MTL consistently gives better performance. MTL helps the model to

	MSR-VTT								MSVD							
	Video-to-Text				Text-to-Video				Video-to-Text				Text-to-Video			
	R@1	R@5	MedianR	MeanR	R@1	R@5	MedianR	MeanR	R@1	R@5	MedianR	MeanR	R@1	R@5	MedianR	MeanR
Baseline (mean-pooled ResNet)	10.5	26.7	25	266.6	5.8	17.6	61.0	296.6	18.1	44.2	11.3	56.3	14.8	39.2	8.5	45.0
Aytar et al. 2017	8.7	22.4	31.0	225.8	4.8	15.3	73.0	313.6	-	-	-	-	-	-	-	-
Otani et al. 2016	-	-	-	-	-	-	-	-	9.85	27.1	19.0	75.2	7.7	23.4	21.0	49.1
Mithun et al. 2018	12.5	32.1	16.0	134.0	7.0	20.9	38.0	213.8	**25.5**	51.3	**5.0**	32.5	20.2	47.5	**6.0**	29.0
Ours (Dyn. Embedding)	11.6	28.2	26.0	158.8	6.1	18.3	77.0	281.2	20.8	42.5	9.0	49.2	17.4	39.3	10.0	38.6
Ours (Dyn. + Pre-opt.)	12.5	31.8	15.0	135.2	6.9	21.1	38.0	214.5	24.9	48.6	6.0	36.1	19.5	46.8	6.0	29.2
Ours (Dyn. + Pre-opt. + MTL)	**12.7**	**33.0**	**15.0**	**128.4**	**7.3**	**21.5**	**37.0**	**203.2**	25.4	**51.6**	**5.0**	**32.3**	**20.8**	**48.3**	**6.0**	**28.4**

Table 4. Comparison of video-text retrieval methods. *Dyn. Embedding* and *Dyn.* denotes the use of dynamic text embedding [12], and *Pre-opt.* denotes pre-optimization of intra-modal semantic space.

generalize better when used with pre-optimization of intra-modal semantic structure. The performance gain from using MTL is smaller in video-text retrieval result when compared to image-text retrieval result and the most plausible explanation is that we only used two datasets, compared to four datasets in image-text retrieval.

5.4. Effect of using dynamic text embedding with multi-task learning

Neither of the text encoders (static and dynamic embedding) clearly outperformed one another, but rather each dominated in different datasets. Despite its simplicity, the simple strategy of average pooling static embeddings to encode text performed well and sometimes better than dynamic text embedding with IAPRTC-12 and NUS-WIDE as seen in Table. 3. Especially, NUS-WIDE did not benefit from using dynamic text embedding in both image-to-text and text-to-image retrieval. We believe that noisy unordered textual modality input (user annotated tags) in NUS-WIDE are not a good form of input to BERT, which typically expects a phrase or sentence.

6. CONCLUSION

In this paper, we propose an effective two-stage optimization strategy for learning a robust representation space for cross-modal retrieval task, and assess its effectiveness. We notice that substantial performance improvements can be obtained by pre-optimizing an intra-modal space before cross-modal transfer learning using bi-directional quadruplet loss. Multi-task learning also plays a crucial role in the learning of more robust embedding by allowing the model to leverage more data. We also find that dynamic text embedding can benefit the retrieval performance when the textual modality of the data is well-formulated phrases or sentences. Our approach is able to attain competitive performance on various datasets, and in most cases outperforms the state of the art.

Acknowledgment

Parts of this work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 and was supported by the LLNL-LDRD Program under Project No. 17-SI-003. Computation resources used in this work were partially supported by AWS Cloud Credits for Research. Any findings and conclusions are those of the authors, and do not necessarily represent the views of the funders.

7. REFERENCES

[1] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *International Journal of Computer Vision*, vol. 106, no. 2, 2014.

[2] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML)*, 2011.

[3] Nitish Srivastava and Ruslan R Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2012.

[4] Li He, Xing Xu, Huimin Lu, Yang Yang, Fumin Shen, and Heng Tao Shen, "Unsupervised cross-modal retrieval through adversarial learning," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2017.

[5] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury, "Learning joint embedding with multimodal cues for cross-modal video-text retrieval," in *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*, 2018.

[6] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang, "Dual encoding for zero-example video retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[7] Jianfeng Dong, Xirong Li, and Cees GM Snoek, "Word2VisualVec: Image and video to sentence match-

ing by visual feature prediction," *IEEE Transactions on Multimedia*, 2018.

[8] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[9] Rich Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, 1997.

[10] Alex Kendall and Yarin Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2017.

[11] Alex Kendall, Yarin Gal, and Roberto Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019.

[13] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proceedings of ACM International Conference on Multimedia*, 2010.

[14] Yang Zhang, Boqing Gong, and Mubarak Shah, "Fast zero-shot image tagging," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] Karl Ni, Kyle Zaragoza, Alexander Gude, Yonas Tesfaye, Carmen Carrano, Charles Foster, and Barry Chen, "Sampled image tagging and retrieval methods on user generated content," in *British Machine Vision Conference (BMVC)*, 2017.

[16] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li, "YFCC100M: The New Data in Multimedia Research," *Commun. ACM*, vol. 59, no. 2, Jan. 2016.

[17] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler, "VSE++: Improved visual-semantic embeddings," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[18] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen, "Adversarial cross-modal retrieval," in *Proceedings of ACM international conference on Multimedia*, 2017.

[19] Xin Huang and Yuxin Peng, "Cross-modal deep metric learning with multi-task regularization," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2017.

[20] Xin Huang and Yuxin Peng, "Deep cross-media knowledge transfer," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[21] Xiangbo Shu, Guo-Jun Qi, Jinhui Tang, and Jingdong Wang, "Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation," in *Proceedings of ACM international conference on Multimedia*, 2015.

[22] Jun Xu, Tao Mei, Ting Yao, and Yong Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[23] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Chris Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele, "Movie description," *International Journal of Computer Vision*, 2017.

[24] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya, "Learning joint representations of videos and sentences with web image search," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

[25] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui, "Jointly modeling embedding and translation to bridge video and language," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[26] Yusuf Aytar, Carl Vondrick, and Antonio Torralba, "See, hear, and read: Deep aligned representations," *arXiv preprint arXiv:1706.00932*, 2017.

[27] Sebastian Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[28] Wu Liu, Tao Mei, Yongdong Zhang, Cherry Che, and Jiebo Luo, "Multi-task deep visual-semantic embedding for video thumbnail selection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[29] Yongxin Yang and Timothy Hospedales, "Deep multi-task representation learning: A tensor factorisation approach," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[30] Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser, "Multi-task sequence to sequence learning," in *Proceedings of International Conference on Learning Represenations (ICLR)*, 2015.

[31] Marek Rei, "Semi-supervised multitask learning for sequence labeling," in *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, 2017.

[32] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang, "Adversarial multi-task learning for text classification," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

[33] Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang, "Representation learning using multi-task deep neural networks for semantic classification and information retrieval," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2015.

[34] Jonathan Baxter, "A model of inductive bias learning," *Journal of artificial intelligence research*, vol. 12, 2000.

[35] 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 (CVPR)*, 2015.

[36] Keke He, Zhanxiong Wang, Yanwei Fu, Rui Feng, Yu-Gang Jiang, and Xiangyang Xue, "Adaptively weighted multi-task deep network for person attribute classification," in *Proceedings of the ACM international conference on Multimedia*, 2017.

[37] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R.G. Lanckriet, R. Levy, and N. Vasconcelos, "A New Approach to Cross-Modal Multimedia Retrieval," in *Proocedings of ACM International Conference on Multimedia*, 2010.

[38] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel, "Multimodal neural language models," in *Proceedings of International Conference on Machine Learning (ICML)*, 2014.

[39] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng, "Nus-wide: A real-world web image database from national university of singapore," in *Proceedings of ACM Conference on Image and Video Retrieval (CIVR)*, 2009.

[40] Yuxin Peng, Xin Huang, and Yunzhen Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges," *IEEE Transactions on circuits and systems for video technology*, vol. 28, no. 9, 2018.

[41] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui, "Jointly modeling embedding and translation to bridge video and language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[42] David L Chen and William B Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proceedings of Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, 2011.

[43] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin, "Pre-training distributed word representations," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2018.

[44] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, "Deep contextualized word representations," in *Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2018.

[45] 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 (CVPR)*, 2016.

[46] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[47] Changsong Yu, Karim Said Barsim, Qiuqiang Kong, and Bin Yang, "Multi-level attention model for weakly supervised audio classification," in *In Proceedings of DCASE2018 Workshop on Detection and Classification of Acoustic Scenes and Events*, 2018.

[48] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[49] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in PyTorch," in *Proceedings of Advances in Neural Information Processing Systems (NIPS) Autodiff Workshop*, 2017.

[50] Jeff Johnson, Matthijs Douze, and Hervé Jégou, "Billion-scale similarity search with gpus," *arXiv preprint arXiv:1702.08734*, 2017.

[51] Wei Wang, Beng Chin Ooi, Xiaoyan Yang, Dongxiang Zhang, and Yueting Zhuang, "Effective multi-modal retrieval based on stacked auto-encoders," in *Proceedings of the VLDB Endowment*, 2014.