Polysemous Visual-Semantic Embedding for Cross-Modal Retrieval

Yale Song
Microsoft Cloud & AI
yalesong@microsoft.com

Mohammad Soleymani
USC Institute for Creative Technologies
soleymani@ict.usc.edu

Abstract

Visual-semantic embedding aims to find a shared latent space where related visual and textual instances are close to each other. Most current methods learn injective embedding functions that map an instance to a single point in the shared space. Unfortunately, injective embedding cannot effectively handle polysemous instances with multiple possible meanings; at best, it would find an average representation of different meanings. This hinders its use in real-world scenarios where individual instances and their cross-modal associations are often ambiguous. In this work, we introduce Polysemous Instance Embedding Networks (PIE-Nets) that compute multiple and diverse representations of an instance by combining global context with locally-guided features via multi-head self-attention and residual learning. To learn visual-semantic embedding, we tie-up two PIE-Nets and optimize them jointly in the multiple instance learning framework. Most existing work on cross-modal retrieval focus on image-text pairs of data. Here, we also tackle a more challenging case of video-text retrieval. To facilitate further research in video-text retrieval, we release a new dataset of 50K video-sentence pairs collected from social media, dubbed MRW (my reaction when). We demonstrate our approach on both image-text and video-text retrieval scenarios using MS-COCO, TGIF, and our new MRW dataset.

1. Introduction

Visual-semantic embedding [9, 20] aims to find a joint mapping of instances from visual and textual domains to a shared embedding space so that related instances from source domains are mapped to nearby places in the target space. This has a variety of downstream applications in computer vision including tagging [9], retrieval [11], captioning [20], visual question answering [19].

Formally, the goal of visual-semantic embedding is to learn two mapping functions $f : \mathcal{X} \rightarrow \mathcal{Z}$ and $g : \mathcal{Y} \rightarrow \mathcal{Z}$ jointly, where $\mathcal{X}$ and $\mathcal{Y}$ are visual and textual domains, respectively, and $\mathcal{Z}$ is a shared embedding space. The functions are often designed to be *injective* so that there is a one-to-one mapping from an instance x (or y) to a single point $z \in \mathbb{R}^d$ in the embedding space. They are often optimized to satisfy the following constraint:

$$d(f(x_i), g(y_i)) < d(f(x_i), g(y_j)), \quad \forall i \neq j \qquad (1)$$

where $d(\cdot, \cdot)$ is a certain distance measure, such as Euclidean and cosine distance. This simple and intuitive setup, which we refer to as *injective instance embedding*, is currently the most popular approach in the literature [44].

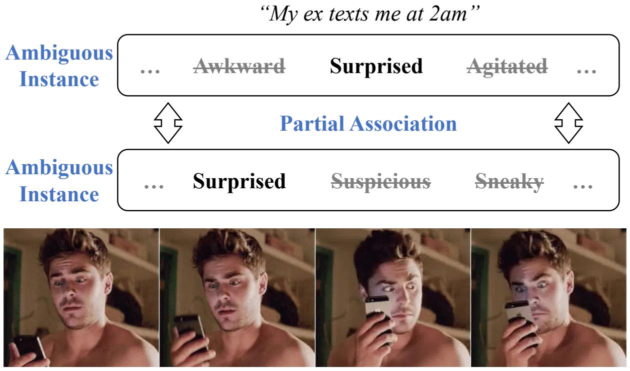

"My ex texts me at 2am"
Ambiguous Instance
… ~~Awkward~~ Surprised ~~Agitated~~ …
Partial Association
Ambiguous Instance
… Surprised ~~Suspicious~~ ~~Sneaky~~ …

Figure 1. Cross-modal retrieval in the real-world could be challenging with *ambiguous instances* (each instance can have multiple meanings/concepts) and their *partial associations* (not all individual meanings/concepts may match). Addressing these two challenges is the focus of this work.

Unfortunately, injective embedding can suffer when there is *ambiguity* in individual instances. Consider an ambiguous instance with multiple meanings/senses, e.g., polysemy words and images containing multiple objects. Even though each of the meanings/senses can map to different points in the embedding space, injective embedding is always forced to find a single point, which could be an (inaccurate) weighted geometric mean of all the desirable points. The issue gets intensified for videos and sentences because the ambiguity in individual images and words can aggregate and get compounded, severely limiting its use in real-world applications such as text-to-video retrieval.

Another case where injective embedding could be problematic is *partial* cross-domain association, a characteristic commonly observed in the real-world datasets. For instance, a text sentence may describe only certain regions of an image while ignoring other parts [47], and a video may contain extra frames not described by its associated sentence [24]. These associations are implicit/hidden, making it unclear which part(s) of the image/video the text description refers to. This is especially problematic for injective embedding because information about any ignored parts will be lost in the mapped point and, once mapped, there is no way to recover from the information loss.

In this work, we address the above issues by (1) formulating instance embedding as a one-to-many mapping task and (2) optimizing the mapping functions to be robust to ambiguous instances and partial cross-modal associations.

To address the issues with ambiguous instances, we propose a novel one-to-many instance embedding model, **P**olysemous **I**nstance **E**mbedding **Net**work (PIE-Net), which extracts K embeddings of each instance by combining global and local information of its input. Specifically, we obtain K *locally-guided* representations by attending to different parts of an input instance (e.g., regions, frames, words) using a multi-head self-attention module [27, 41]. We then combine each of such local representation with global representation via residual learning [15] to avoid learning redundant information. Furthermore, to prevent the K embeddings from collapsing into the mode (or the mean) of all the desirable embeddings, we regularize the K locally-guided representations to be diverse. To our knowledge, we are the first to apply multi-head self-attention with residual learning for the application of instance embedding.

To address the partial association issue, we tie-up two PIE-Nets and train our model in the multiple-instance learning (MIL) framework [5]. We call this approach **P**olysemous **V**isual-**S**emantic **E**mbedding (PVSE). Our intuition is: when two instances are only partially associated, the learning constraint of Equation 1 will unnecessarily penalize embedding mismatches because it expects two instances to be perfectly associated. Capitalizing on our one-to-many instance embedding, our MIL objective relaxes the constraint of Equation 1 so that only one of $K \times K$ embedding pairs is well-aligned, making our model more robust to partial cross-domain association. We illustrate this intuition in Figure 2. This relaxation, however, could cause a discrepancy between two embedding distributions because $(K \times K - 1)$ embedding pairs are left unconstrained. We thus regularize the learned embedding space by minimizing the discrepancy using the Maximum Mean Discrepancy (MMD) [13], a popular technique for determining whether two sets of data are from the same probability distribution.

We demonstrate our approach on two cross-modal retrieval scenarios: image-text and video-text. For image-text retrieval, we evaluate on the MS-COCO dataset [25]; for video-text retrieval, we evaluate on the TGIF dataset [24] as well as our new MRW (my reaction when) dataset, which we collected to promote further research in cross-modal video-text retrieval under ambiguity and partial association. The dataset contains 50K video-sentence pairs collected from social media, where the videos depict physical or emotional reactions to certain situations described in text. We compare our method with well-established baselines and carefully conduct an ablation study to justify various design choices. We report strong performance on all three datasets, and achieve the state-of-the-art result on image-to-text retrieval task on the MS-COCO dataset.

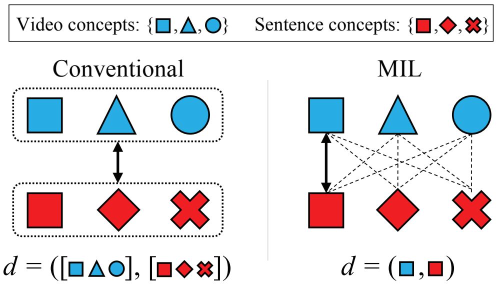


Figure 2. We represent each instance with k embeddings, each representing different parts of the instance, e.g., regions of an image, frames of a video, or words of a sentence. Conventional approaches measure the visual-semantic distance by considering all k embeddings, and thus would suffer when not all concepts are related. We instead assume there is a partial match and measure the distance between only the most related combination (squares).

2. Related Work

Here we briefly review some of the most relevant work on instance embedding for cross-modal retrieval.

Correlation maximization: Most existing methods are based on one-to-one mapping of instances into a shared embedding space. One popular approach is maximizing correlation between related instances in the embedding space. Rasiwasia *et al.* [33] use canonical correlation analysis (CCA) to maximize correlation between images and text, while Gong *et al.* [11] extend CCA to a triplet scenario, e.g., images, tags, and their semantic concepts. Most recent methods incorporate deep neural networks to learn their embedding models in an end-to-end fashion. Andrew *et al.* [2] propose deep CCA (DCCA), and Yan *et al.* [48] apply it to image-to-sentence and sentence-to-image retrieval.

Triplet ranking: Another popular approach is based on triplet ranking [9, 21, 45, 49], which encourages the distance between positive pairs (e.g., ground-truth pairs) to be closer than negative pairs (e.g., randomly selected pairs). Frome *et al.* [9] propose a deep visual-semantic embedding (DeViSE) model, using a hinge loss to implement triplet ranking. Faghri *et al.* [7] extend this with the idea of hard negative mining, which focuses on maximum violating negative pairs, and report improved convergence rates.

Learning with auxiliary tasks: Several methods learn the embeddings in conjunction by solving auxiliary tasks, e.g., signal reconstruction [8, 6, 40], semantic concept categorization [33, 18], and minimizing the divergence between embedding distributions induced by different modalities [40, 50]. Adversarial training [12] is also used by many: Wang *et al*. [43] encourage the embeddings from different modalities to be indistinguishable using a domain discriminator, while Gu *et al*. [14] learn the embeddings with image-to-text and text-to-image synthesis tasks in the adversarial learning framework.

Attention-based embedding: All the above approaches are based on one-to-one mapping and thus could suffer from polysemous instances. To alleviate this, recent methods incorporate cross-attention mechanisms to selectively attend to local parts of an instance *given the context of* a conditioning instance from another modality [17, 23], e.g., attend to different image regions given different text queries. Intuitively, this can resolve the issues with ambiguous instances and their partial associations because the same instance can be mapped to different points depending on the presence of the conditioning instance. However, such approach comes with computational overhead at inference time because each query instance needs to be encoded as many times as the number of references instances in the database; this severely limits its use in real-world applications. Different from previous approaches, our method is based on multi-head self-attention [27, 41] which does not require a conditioning instance when encoding, and therefore each instance is encoded only once, significantly reducing computational overhead at inference time.

Beyond injective embedding: Similar to our motivation, some attempts have been made to go beyond the injective mapping. One approach is to design the embedding function to be stochastic and map an instance to a certain probability distribution (e.g., Gaussian) instead of a single point [35, 30, 31]. However, learning distributions is typically difficult/expensive and often lead to approximate solutions such as Monte Carlo sampling.

The work most similar to ours is by Ren *et al*. [36], where they compute multiple representations of an image by extracting local features using the region proposal method [10]; text instances are still represented by a single embedding vector. Different from theirs, our method computes multiple and diverse representations from both modalities, where each representation is a combination of global context and locally-guided features, instead of just a local feature. Song *et al*. [39], a prequel to this work, also compute multiple representations of each instance using multi-head self-attention. We extend their approach by combining global and locally-guided features via residual learning. We also extend the preliminary version of the MRW dataset with an increased number of sample pairs. Lastly, we report more comprehensive experimental results, adding results on the MS-COCO [25] dataset for image-text cross-retrieval.

3. Approach

Our Polysemous Visual-Semantic Embedding (PVSE) model, shown in Figure 3, is composed of modality-specific feature extractors followed by two sub-networks with an identical architecture; we call the sub-network Polysemous Instance Embedding Network (PIE-Net). The two PIE-Nets are independent of each other and do not share the weights.

The PIE-Net takes as input a global context vector and multiple local feature vectors (Section 3.1), computes locally-guided features using the local feature transformer (Section 3.2), and outputs K embeddings by combining the global context vector with locally-guided features (Section 3.3). We train the PVSE model in the Multiple Instance Learning (MIL) [5] framework. We explain how we make our model robust to ambiguous instances and partial cross-modal associations via our loss functions (Section 3.4) and finish with implementation details (Section 3.5).

3.1. Modality-Specific Feature Encoder

Image encoder: We use the ResNet-152 [15] pretrained on ImageNet [38] to encode an image x. We take the feature map before the final average pooling layer as local features $\Psi(x) \in \mathbb{R}^{7\times7\times2048}$. We then apply average pooling to $\Psi(x)$ and feed the output to one fully-connected layer to obtain global features $\phi(x) \in \mathbb{R}^{H}$.

Video encoder: We use the ResNet-152 to encode each of T frames from a video x, taking the 2048-dim output from the final average pooling layer, and use them as local features $\Psi(x) \in \mathbb{R}^{T\times2048}$. We then feed $\Psi(x)$ into a bidirectional GRU (bi-GRU) [4] with H hidden units, and take the final hidden states as global features $\phi(x) \in \mathbb{R}^{H}$.

Sentence encoder: We encode each of L words from a sentence x using the GloVe [32] pretrained on the CommonCrawl dataset, producing L 300-dim vectors, and use them as local features $\Psi(x) \in \mathbb{R}^{L\times300}$. We then feed them into a bi-GRU with H hidden units, and take the final hidden states as global features $\phi(x) \in \mathbb{R}^{H}$.

3.2. Local Feature Transformer

The local feature transformer takes local features $\Psi(x)$ and transforms them into K locally-guided representations $\Upsilon(x)$. Our intuition is that different combinations of local information could yield diverse and refined representations of an instance. We implement this intuition by employing a multi-head self-attention module to obtain K attention maps, prepare K combinations of local features by attending to different parts of an instance, and apply non-linear transformations to obtain K locally-guided representations.

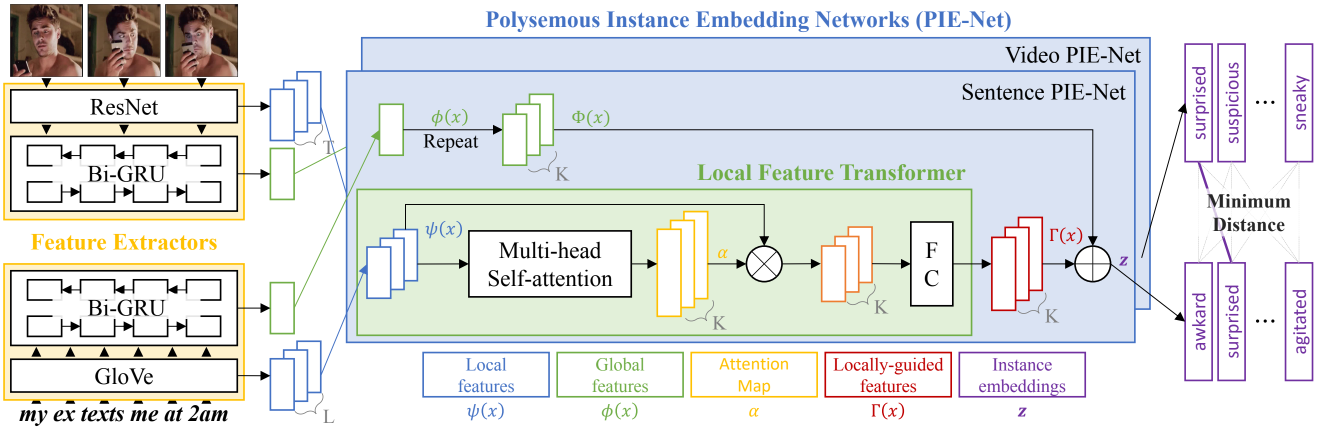


Figure 3. The architecture of Polysemous Visual-Semantic Embedding (PVSE) for video-sentence data.

We use a two-layer perceptron to implement the multi-head self-attention module.[1] Given local features $\Psi(x) \in \mathbb{R}^{B\times D}$[2], it computes K attention maps $\alpha \in \mathbb{R}^{K\times B}$:

$$\alpha = \text{softmax}\left(w_2 \tanh\left(w_1 \Psi(x)^\intercal\right)\right) \tag{2}$$

where $w_2 \in \mathbb{R}^{K\times A}$, $w_1 \in \mathbb{R}^{A\times D}$; we set $A = D/2$ per empirical evidence. The softmax is applied row-wise so that each of the K attention coefficients sum up to one.

Finally, we multiply the attention map with local features and further apply a non-linear transformation to obtain K locally-guided representations $\Upsilon(x) \in \mathbb{R}^{K\times H}$:

$$\Upsilon(x) = \sigma\left((\alpha\Psi(x))w_3 + b_3\right) \tag{3}$$

where $w_3 \in \mathbb{R}^{D\times H}$ and $b_3 \in \mathbb{R}^H$. We use the sigmoid as our activation function $\sigma(\cdot)$.

3.3. Feature Fusion With Residual Learning

The fusion block combines global features $\phi(x)$ and locally-guided features $\Upsilon(x)$ to obtain the final K embedding output. We note that there is an inherent information overlap between the two features (both are derived from the same instance). To prevent $\Upsilon(x)$ from becoming redundant with $\phi(x)$ and encourage it to learn only locally-specific information, we cast the feature fusion as a residual learning task. Specifically, we consider $\phi(x)$ as input to the residual block and $\Upsilon(x)$ as residuals with its own parameters to optimize (w_1, w_2, w_3, b_3). As shown in [15], this residual mapping makes it easier to optimize the parameters associated with $\Upsilon(x)$, helping us find meaningful locally-specific information; in the extreme case, if global features $\phi(x)$ were the optimal, the residuals will be pushed to zero and the approach will fall back to the standard injective embedding.

We compute K embedding vectors $z \in \mathbb{R}^{K\times H}$ as:

$$z = \text{LayerNorm}\left(\Phi(x) + \Upsilon(x)\right) \tag{4}$$

where $\Phi(x) \in \mathbb{R}^{K\times H}$ is K repetitions of $\phi(x)$. Following [41], we apply the layer normalization [3] to the output.

3.4. Optimization and Inference

Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ with N instance pairs (x are either images or videos, y are sentences), we optimize our PVSE model to minimize a learning objective:

$$\mathcal{L} = \mathcal{L}_{mil} + \lambda_1 \mathcal{L}_{div} + \lambda_2 \mathcal{L}_{mmd} \tag{5}$$

where λ_1 and λ_2 are scalar weights that balance the influence of the loss terms. We describe each loss term below.

MIL Loss: We train our model in the Multiple Instance Learning (MIL) framework [5], designing a learning constraint for the cross-modal retrieval scenario:

$$\min_{p,q} d(z^x_{i,p}, z^y_{i,q}) < d(z^x_{i,p}, z^y_{j,q}), \quad \forall i \neq j, \ \forall p, q \tag{6}$$

where z^x and z^y are the PIE-Net embeddings of x and y, respectively, and $p, q = 1, \cdots, K$. We use the cosine distance as our distance metric, $d(a, b) = (a \cdot b)/(\|a\|\|b\|)$.

Making an analogy to the MIL for binary classification [1], the left side of the constraint is the "positive" bag where at least one of $K \times K$ embedding pairs is assumed to be positive (match), while the right side is the "negative" bag containing only negative (mismatch) pairs. Optimizing under this constraint allows our model to be robust to partial cross-modal association because it can ignore mismatching embedding pairs of partially associated instances.

We implement the above constraint by designing our MIL loss function $\mathcal{L}_{mil}$ to be:

$$\frac{1}{N^2}\sum_{i,j}^{N} \max\left(0, \rho - \min_{p,q} d(z^x_{i,p}, z^y_{j,q}) + \min_{p,q} d(z^x_{i,p}, z^y_{i,q})\right)$$

[1] We have experimented with a more sophisticated version of the multi-head self-attention [41], but it did not improve performance further.

[2] B is 49 ($= 7 \times 7$) for images, T for videos, and L for sentences; D is 2048 for images and videos, and 300 for sentences

where ρ is a margin parameter. Notice that we have the min operator for $d(z^x_{i,p}, z^y_{j,q})$, similar to [36]; this can be seen as a form of hard negative mining, which we found to be effective and accelerate the convergence.

Diversity Loss: To ensure that our PIE-Net produces diverse representations of an instance, we design a diversity loss $\mathcal{L}_{div}$ that penalizes the redundancy among K locally-guided features. To measure the redundancy, we compute a Gram matrix of $\Upsilon(x)$ (and of $\Upsilon(y)$) that encodes the correlations between all combinations of locally-guided features, i.e., $G_{i,j} = \sum_h \Upsilon(x)_{ih}\Upsilon(x)_{jh}$. We normalize each $\Upsilon(x)_i$ prior to the computation so that they are on an l_2 ball.

The diagonal entries in G are always one (they are on a unit ball); the off-diagonals are zero iff two locally-guided features are orthogonal to each other. Therefore, the sum of off-diagonal entries in G indicates the redundancy among K locally-guided features. Based on this, we define our diversity loss as:

$$\mathcal{L}_{div} = \frac{1}{K^2}\left(\|G^x - I\|_2 + \|G^y - I\|_2\right) \qquad (7)$$

where G^x and G^y are the gram matrices of $\Upsilon(x)$ and $\Upsilon(y)$, respectively, and $I \in \mathbb{R}^{K\times K}$ is an identity matrix.

Note that we do not compute the diversity loss on the final embedding representations z^x and z^y because they already have global information baked in, making the orthogonality constraint invalid. This also ensures that the loss gets back-propagated through appropriate parts in the computational graph, and does not affect the global feature encoders, i.e., the FC layer for the image encoder, and the bi-GRUs for the video and sentence encoders.

Domain Discrepancy Loss: Optimizing our model under the MIL loss has one drawback: two distributions induced by z^x and z^y, which we denote by Z^x and Z^y, respectively, may diverge quickly because we only consider the minimum distance pair, $\min_{p,q} d(z^x_p, z^y_q)$, in loss computation and let the other $(K \times K - 1)$ pairs left to be unconstrained. It is therefore necessary to regularize the discrepancy between the two distributions.

One popular way to measure the discrepancy between two probability distributions is the Maximum Mean Discrepancy (MMD) [13]. The MMD between two distributions P and Q over a function space $\mathcal{F}$ is

$$\text{MMD}(P, Q) = \sup_{f\in\mathcal{F}} \left(\mathbb{E}_{X\sim P}\left[f(X)\right] - \mathbb{E}_{Y\sim Q}\left[f(Y)\right]\right) \quad (8)$$

When $\mathcal{F}$ is a reproducing kernel Hilbert space (RKHS) with a kernel $\kappa : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ that measures the similarity between two samples, Gretton *et al.* [13] showed that the supremum is achieved at $f(x) = \mathbb{E}_{X'\sim P}[\kappa(x, X')] - \mathbb{E}_{X'\sim Q}[\kappa(x, X')]$. Substituting this to Equation 8 and squaring the result, and approximating the expectation over our empirical distributions Z^x and Z^y, we have our domain discrepancy loss $\mathcal{L}_{mmd}$ defined as

$$\frac{\sum \kappa(z^x_{i,p}, z^x_{j,q}) - 2\sum \kappa(z^x_{i,p}, z^y_{j,q}) + \sum \kappa(z^y_{i,p}, z^y_{j,q})}{K^2N^2}$$

where the summation in each term is taken over all pairs of embeddings $(i, j, p, q) \in [1, \cdots, K^2N^2]$. We use a radial basis function (RBF) kernel as our kernel function.

Inference: At test time, we assume a database of M instances (e.g., videos) and their KM embedding vectors. Given a query instance (e.g., a sentence), we compute K embedding vectors and find the best matching instance in the database by comparing the cosine distances between all K^2M combinations of embeddings.

3.5. Implementation Details

We subsample frames at 8 FPS and store them in a binary storage format.[3] We set the maximum length of video to be 8 frames; for videos longer than 8 frames we select random subsequences during training, while during inference we sample 8 frames evenly spread across each video. We do not limit the sentence length as it has a minimal effect on the GPU memory footprint. We cross-validate the optimal hyper-parameter settings, varying $K \in [1 : 8]$, $H \in [512, 1024, 2048], \rho \in [0.1 : 1.0], \lambda_1, \lambda_2 \in [0.1, 0.01, 0.001]$. We use the AMSGRAD optimizer [34] with an initial learning rate of 2e-4 and reduce it by half when the loss stagnates. We train our model end-to-end, except for the pretrained CNN weights, for 50 epochs with a batch of 128 samples. We then finetune the whole model (including the CNN weights) for another 50 epochs.

4. MRW Dataset

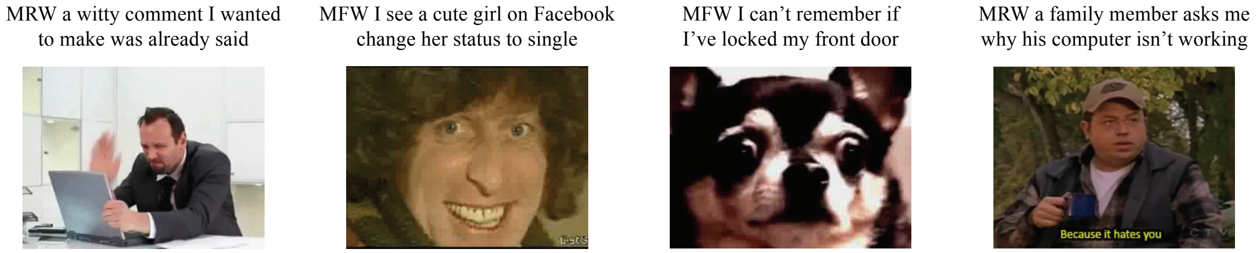
To promote future research in video-text cross-modal retrieval, especially with ambiguous instances and their partial cross-domain association, we release a new dataset of 50K video-sentence pairs collected from social media; we call our dataset MRW (my reaction when).

Table 1 provides descriptive statistics of several video-sentence datasets. Most existing datasets are designed for video captioning [37, 46, 24], with sentences providing textual descriptions of visual content in videos (video → text relationship). Our dataset is unique in that it provides videos that display physical or emotional reactions to the given sentences (text → video relationship); these are called *reaction GIFs*. According to a subreddit `r/reactiongif`[4]:

> *A reaction GIF is a physical or emotional response that is captured in an animated GIF which you can link in response to someone or something on the Internet. The reaction must not be in response to something that happens within the GIF, or it is considered a "scene".*

[3] https://github.com/TwentyBN/GulpIO

[4] https://www.reddit.com/r/reactiongifs


MRW a witty comment I wanted to make was already said

MFW I see a cute girl on Facebook change her status to single

MFW I can't remember if I've locked my front door

MRW a family member asks me why his computer isn't working

(a) Physical Reaction (b) Emotional Reaction (c) Animal Reaction (d) Lexical Reaction (Caption)

Figure 4. Our dataset contains videos depicting *reactions* to the situations described in the corresponding sentences. Here we show the four most common reaction types: (a) physical, (b) emotional, (c) animal, (d) lexical.

	#clips	#sentences	vocab	text source
LSMDC16 [37]	128,085	128,085	22,898	DVS
MSR-VTT [46]	10,000	200,000	29,316	AMT
TGIF [24]	100,000	125,781	11,806	AMT
DiDeMo [16]	26,982	40,543	7,785	AMT
MRW	50,107	50,107	34,835	In-the-wild

Table 1. **Descriptive statistics** of our dataset compared to existing video-sentence datasets.

This definition clearly differentiates ours from existing datasets: There is an inherently weaker association of concepts between video and text; see Figure 4. This introduces several additional challenges to cross-modal retrieval, part of which are the focus of this work, i.e., dealing with ambiguous instances and partial cross-domain association. We provide detailed data analyses and compare it with existing video captioning datasets in the supplementary material.

5. Experiments

We evaluate our approach on image-text and video-text cross-modal retrieval scenarios. For image-text cross-retrieval, we evaluate on the MS-COCO dataset [25]; for video-text we use the TGIF [24] and our MRW datasets.

For MS-COCO we use the data split of [21], which provides 113,287 training, 5K validation and 5K test samples; each image comes with 5 captions. We report results on both 1K unique test images (averaged over 5 folds) and the full 5K test images. For TGIF we use the original data split [24] with 80K training, 10,708 validation and 34,101 test samples; since most test videos come with 3 captions, we report results on 11,360 unique test videos. For MRW, we use a data split of 44,107 training, 1K validation and 5K test samples; all the videos come with one caption.

Following the convention in cross-modal retrieval, we report results using Recall@k (R@k) at $k = 1, 5, 10$, which measures the the fraction of queries for which the correct item is retrieved among the top k results. We also report the median rank (Med R) of the closest ground truth result in the list, as well as the normalized median rank (nMR) that divides the median rank by the number of total items. For cross-validation, we select the best model that achieves the highest $rsum = R@1 + R@5 + R@10$ in both directions (visual-to-text and text-to-visual) on a validation set.

While we report quantitative results in the main paper, our supplementary material contains qualitative results with visualizations of multi-head self-attention maps.

5.1. Image-Text Retrieval Results

Table 2 shows the results on MS-COCO. To facilitate comprehensive comparisons, we provide previously reported results on this dataset.[5] Our approach outperforms most of the baselines, and achieves the new state-of-the-art on the image-to-text task on the 5K test set. We note that both GXN [14] and SCO [18] are trained with multiple objectives; in addition to solving the ranking task, GXN performs image-text cross-modal synthesis as part of training, while SCO performs classification of semantic concepts and their orders as part of training. Compared to the two methods, our model is trained with a single objective (ranking) and thus could be considered as a simpler model.

The most direct comparison to ours would be with VSE++ [7]. Both our model and VSE++ share the same image and sentence encoders. When we let our PIE-Net to produce single embeddings for input instances (K=1), the only difference becomes that VSE++ directly uses our global features as their embedding representations, while we use the output from our PIE-Nets. The performance gap between ours (K=1) and VSE++ shows the effectiveness of our PIE-Net, which combines global context with locally-guided features produced by our local feature transformer.

5.2. Video-Text Retrieval Results

Table 3 and Table 4 show the results on TGIF and MRW datasets. Because there is no previously reported results on these datasets for the cross-model retrieval scenario, we run the baseline models and report their results. We can see that our method show strong performance compared to all the baselines. We provide implementation details of the baseline models in the supplementary material.

[5] We omit results from cross-attention models [17, 23] that require a pair of instances (e.g., image and text) when encoding each instance.

Method	1K Test Images						5K Test Images					
	Image-to-Text			Text-to-Image			Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
DVSA [20]	38.4	69.9	80.5	27.4	60.2	74.8	16.5	39.2	52.0	10.7	29.6	42.2
GMM-FV [22]	39.4	67.9	80.9	25.1	59.8	76.6	17.3	39.0	50.2	10.8	28.3	40.1
m-CNN [29]	42.8	73.1	84.1	32.6	68.6	82.8	-	-	-	-	-	-
Order [42]	46.7	-	88.9	37.9	-	85.9	23.3	-	65.0	18.0	-	57.6
DSPE [45]	50.1	79.7	89.2	39.6	75.2	86.9	-	-	-	-	-	-
VQA-A [26]	50.5	80.1	89.7	37.0	70.9	82.9	23.5	50.7	63.6	16.7	40.5	53.8
2WayNet [6]	55.8	75.2	-	39.7	63.3	-	-	-	-	-	-	-
RRF-Net [28]	56.4	85.3	91.5	43.9	78.1	88.6	-	-	-	-	-	-
CMPM [50]	56.1	86.3	92.9	44.6	78.8	89.0	31.1	60.7	73.9	22.9	50.2	63.8
VSE++ [7]	64.6	90.0	95.7	52.0	84.3	92.0	41.3	71.1	81.2	30.3	59.4	72.4
GXN [14]	68.5	-	**97.9**	56.6	-	94.5	-	-	-	-	-	-
SCO [18]	**69.9**	**92.9**	97.5	**56.7**	**87.5**	**94.8**	42.8	72.3	83.0	**33.1**	62.9	**75.5**
PVSE (K=1)	66.7	91.0	96.2	53.5	85.1	92.7	41.7	73.0	83.0	30.6	61.4	73.6
PVSE	69.2	91.6	96.6	55.2	86.5	93.7	**45.2**	**74.3**	**84.5**	32.4	**63.0**	75.0

Table 2. **MS-COCO results.** Besides our results, we also provide previously reported results to facilitate comprehensive comparisons.

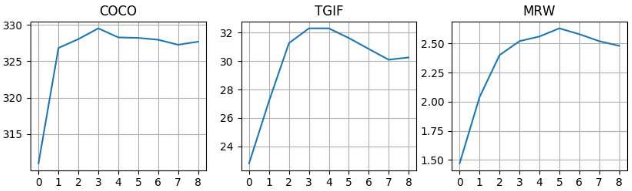


Figure 5. Performance (*rsum*) with different numbers of embeddings, $K = [0 : 8]$. The results at $K = 0$ is when we take out the PIE-Net and use the global feature as the embedding output.

We notice is that the overall performance is much lower than the results from MS-COCO. This shows how challenging video-text retrieval is (and video understanding in a broader context), and calls for further research in this task. We can also see that there is a large performance gap between the two datasets. This suggests the two datasets have significantly different characteristics: the TGIF contains sentences describing visual content in videos, while our MRW dataset contains videos showing one of possible reactions to certain situations described in sentences. This makes the association between video and text modalities much weaker for the MRW than for the TGIF.

5.3. Ablation Results

The number of embeddings K**:** Tables 2, 3, 4 show that computing multiple embeddings per instance improves performance compared to just a single embedding (see the last two rows in each table). To better understand the effect of K, we vary it from 1 to 8, and also compare with $K = 0$, a baseline where we bypass our Local Feature Transformer and simply use the global feature as the final embedding representation. Figure 5 shows the performance on all three datasets based on the *rsum* metric (R@1 + R@5 + R@10 for image/video-to-text and back). The results are from the models before fine-tuning the ResNet-152 weights. We can see that there is a significant improvement from $K = 0$ to $K = 1$; this shows the effectiveness of our Local Feature Transformer. We can make an interesting observation by comparing the optimal K settings across different datasets: $K = 3$ for COCO and TGIF, and $K = 5$ for MRW. While this cannot be used as strong evidence, we believe this shows the level of ambiguity is higher on MRW than the other two datasets.

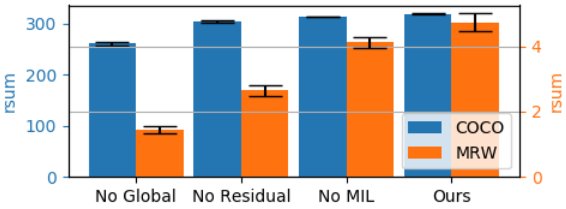


Figure 6. Performance (*rsum*) on MS-COCO and MRW with different ablative settings. The error bars are obtained from multiple runs over $K = [1 : 8]$.

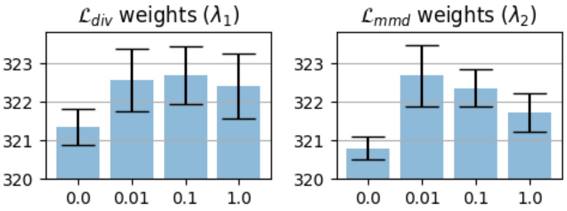


Figure 7. Performance (*rsum*) on MS-COCO with different loss weights for $\mathcal{L}_{div}$ and $\mathcal{L}_{mmd}$. The error bars are obtained from multiple runs of $K = [2 : 4]$ and $\lambda_{(\cdot)} = [0.0, 0.01, 0.1, 1.0]$.

Method	Video-to-Text				Text-to-Video			
	R@1	R@5	R@10	Med R (nMR)	R@1	R@5	R@10	Med R (nMR)
DeViSE [9]	0.84	3.53	6.02	379 (0.03)	0.83	3.38	5.99	378 (1.03)
VSE++ [7]	0.42	1.63	3.60	692 (0.09)	0.55	1.89	3.77	620 (0.09)
Order [42]	0.51	2.09	3.80	500 (0.04)	0.48	2.13	3.86	478 (0.04)
Corr-AE [8]	0.89	3.41	5.61	365 (0.03)	0.90	3.48	5.97	352 (0.03)
PVSE (K=1)	1.51	5.67	8.75	292 (0.03)	1.61	5.23	8.51	284 (0.03)
PVSE	**2.32**	**7.49**	**11.94**	**162 (0.01)**	**2.17**	**7.76**	**12.25**	**155 (0.01)**

Table 3. Experimental results on the TGIF dataset.

Method	Video-to-Text				Text-to-Video			
	R@1	R@5	R@10	Med R (nMR)	R@1	R@5	R@10	Med R (nMR)
DeViSE [9]	0.02	0.18	0.56	1917 (0.38)	0.10	0.38	0.54	1917 (0.38)
VSE++ [7]	0.12	0.38	0.82	1781 (0.36)	0.14	0.44	0.88	1767 (0.35)
Order [42]	0.04	0.14	0.40	1771 (0.35)	0.02	0.14	0.32	1780 (0.36)
Corr-AE [8]	0.14	0.54	1.06	1605 (0.32)	0.04	0.26	0.60	1614 (0.37)
PVSE (K=1)	0.10	0.40	0.76	1595 (0.32)	0.10	0.38	0.66	1619 (0.32)
PVSE	**0.16**	**0.68**	**1.80**	**1586 (0.32)**	**0.16**	**0.60**	**1.83**	**1573 (0.37)**

Table 4. Experimental results on the MRW dataset.

Global vs. locally-guided features: We analyze the importance of global and locally-guided features, as well as different strategies to combine them. Figure 6 shows results on several ablative settings: `No Global` is when we use locally-guided features alone (discard global features); `No Residual` is when we simply concatenate global and locally-guided features, instead of combining them via residual learning. We report results on both MS-COCO and MRW because the two datasets exhibit the biggest difference in the level of ambiguity.

We notice that the performance drops significantly on both datasets when we discard global features. Together with $K = 0$ results in Figure 5 (discard locally-guided features), this shows the importance of balancing global and local information in the final embedding. We also see that simply concatenating the two features (no residual learning) hurts the performance, and the drop is more significant on the MRW dataset. This suggests our residual learning setup is especially crucial for highly ambiguous data.

MIL objective: Figure 6 also shows the result of `No MIL`, which is when we concatenate the K embeddings and optimize the standard triplet ranking objective [9, 21, 7], i.e., the "Conventional" setup in Figure 2. While the differences are relatively smaller than with the other ablative settings, there are statistically significant differences between the two results on both datasets ($p = 0.046$ on MS-COCO and $p = 0.015$ on MRW). We also see that the difference between `No MIL` and `Ours` on MRW is more pronounced than on MS-COCO. This suggests thed MIL objective is especially effective for highly ambiguous data.

Sensitivity analysis on different loss weights: Figure 7 shows the sensitivity of our approach when we vary the relative loss weights, i.e., λ_1 and λ_2 in Equation 5. Note that the weights are relative, not absolute, e.g., instead of directly multiplying $\lambda_1 = 1.0$ to $\mathcal{L}_{div}$, we first scale it to $\lambda_1 \times (\mathcal{L}_{mil}/\mathcal{L}_{div})$ and then multiply it to $\mathcal{L}_{div}$. The results show that both loss terms are important in our model. We can see, in particular, that $\mathcal{L}_{mmd}$ plays an important role in our model. Without it, the two embedding spaces induced by different modalities may diverge quickly due to the MIL objective, which may result in a poor convergence rate. Overall, our results suggests that the model is not much sensitive to the two relative weight terms.

6. Conclusion

Ambiguous instances and their partial associations pose significant challenges to cross-modal retrieval. Unlike the traditional approaches that use injective embedding to compute a single representation per instance, we propose a Polysemous Instance Embedding Network (PIE-Net) that computes *multiple and diverse* representations per instance. To obtain visual-semantic embedding that is robust to partial cross-modal association, we tie-up two PIE-Nets, one per modality, and jointly train them using the Multiple Instance Learning objective. We demonstrate our approach on the image-text and video-text cross-modal retrieval scenarios and report strong results compared to several baselines.

Part of our contribution is also in the newly collected MRW dataset. Unlike existing video-sentence datasets that contain sentences describing visual content in videos, ours contain videos illustrating *one of possible reactions* to certain situations described in sentences, which makes video-sentence association somewhat ambiguous. This poses new challenges to cross-modal retrieval; we hope there will be further progress on this challenging new dataset.

References

[1] Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*, 2013.

[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *ICML*, 2013.

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[4] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, 2014.

[5] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 1997.

[6] Aviv Eisenschtat and Lior Wolf. Linking image and text with 2-way nets. In *CVPR*, 2017.

[7] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: improved visual-semantic embeddings. In *BMVC*, 2017.

[8] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *ACM Multimedia*, 2014.

[9] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

[10] Ross Girshick. Fast r-cnn. In *ICCV*, 2015.

[11] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 2014.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[13] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *NIPS*, 2007.

[14] Jiuxiang Gu, Jianfei Cai, Shafiq Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *CVPR*, 2018.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[16] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pages 5803–5812, 2017.

[17] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *CVPR*, 2017.

[18] Yan Huang and Qi Wu. Learning semantic concepts and order for image and sentence matching. In *CVPR*, 2018.

[19] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2017.

[20] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[21] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[22] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *CVPR*, 2015.

[23] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. 2018.

[24] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. TGIF: A new dataset and benchmark on animated GIF description. In *CVPR*, 2016.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[26] Xiao Lin and Devi Parikh. Leveraging visual question answering for image-caption ranking. In *ECCV*, 2016.

[27] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In *ICLR*, 2017.

[28] Yu Liu, Yanming Guo, Erwin M Bakker, and Michael S Lew. Learning a recurrent residual fusion network for multimodal matching. In *ICCV*, 2017.

[29] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *ICCV*, 2015.

[30] Tanmoy Mukherjee and Timothy Hospedales. Gaussian visual-linguistic embedding for zero-shot recognition. In *EMNLP*, 2016.

[31] Seong Joon Oh, Kevin Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew Gallagher. Modeling uncertainty with hedged instance embedding. *arXiv preprint arXiv:1810.00319*, 2018.

[32] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[33] 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 *ACM Multimedia*, 2010.

[34] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *ICLR*, 2018.

[35] Zhou Ren, Hailin Jin, Zhe Lin, Chen Fang, and Alan Yuille. Joint image-text representation by gaussian visual-semantic embedding. In *ACM Multimedia*, 2016.

[36] Zhou Ren, Hailin Jin, Zhe Lin, Chen Fang, and Alan Yuille. Multiple instance visual-semantic embedding. In *BMVC*, 2017.

[37] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *IJCV*, 2017.

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[39] Yale Song and Mohammed Soleymani. Cross-modal retrieval with implicit concept association. *arXiv preprint arXiv:1804.04318*, 2018.

[40] Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov. Learning robust visual-semantic embeddings. In *ICCV*, 2017.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[42] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In *ICLR*, 2016.

[43] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *ACM Multimedia*, 2017.

[44] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.

[45] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016.

[46] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016.

[47] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[48] Fei Yan and Krystian Mikolajczyk. Deep correlation for matching images and text. In *CVPR*, 2015.

[49] Keren Ye and Adriana Kovashka. Advise: Symbolism and external knowledge for decoding advertisements. In *ECCV*, 2018.

[50] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *ECCV*, 2018.