Do Cross Modal Systems Leverage Semantic Relationships?

Shah Nawaz[1], M. Kamran Janjua[2], Ignazio Gallo[1], Arif Mahmood[3], Alessandro Calefati[1], and Faisal Shafait[2]

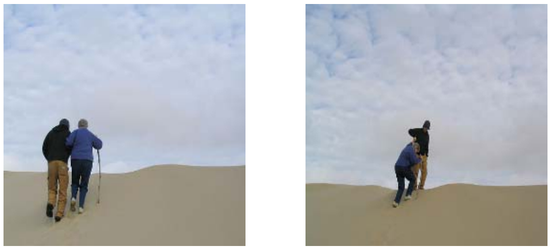
[1]Department of Theoretical and Applied Science, University of Insubria, Varese, Italy
[2]SEECS, National University of Sciences and Technology, Islamabad, Pakistan
[3]Department of Computer Science, Information Technology University, Lahore, Pakistan
{snawaz,ignazio.gallo,a.calefati}@uninsubria.it,
{mjanjua.bscs16seecs,faisal.shafait}@seecs.edu.pk, arif.mahmood@itu.edu.pk

Abstract

Current cross modal retrieval systems are evaluated using $R@K$ measure which does not leverage semantic relationships rather strictly follows the manually marked image text query pairs. Therefore, current systems do not generalize well for the unseen data in the wild. To handle this, we propose a new measure `SemanticMap` *to evaluate the performance of cross modal systems. Our proposed measure evaluates the semantic similarity between the image and text representations in the latent embedding space. We also propose a novel cross modal retrieval system using a single stream network for bidirectional retrieval. The proposed system is based on a deep neural network trained using extended center loss, minimizing the distance of image and text descriptions in the latent space from the class centers. In our system, the text descriptions are also encoded as images which enabled us to use single stream network for both text and images. To the best of our knowledge, our work is the first of its kind in terms of employing a single stream network for cross modal retrieval systems. The proposed system is evaluated on two publicly available datasets including MSCOCO and Flickr30K and has shown comparable results to the current state-of-the-art methods.*

1. Introduction

The mapping of multiple modalities to a shared latent space has resulted in improved scene understanding because the retrieved text/audio help in understanding the visual content. Due to recent advancements in this area [46, 12, 45, 31], we see a surge in multimodal tasks such as image caption generation [11, 25, 49, 55, 33], visual question answering [14, 1], and audio-visual correspondence [2, 4, 36, 6]. One of the drawback of these techniques is that they rely on manually marked image-text pairs. If two different images have similar text descriptions, the pairwise loss functions usually end up wrongly associating text neighbors. The training objective of using these raw pairs is to adhere image and its text description in a pair and to ignore those descriptions which might be semantically similar but in different pairs, see Figure 1 where images (a) and (b) are semantically similar but occur in different pairs.



(a) A man and old woman with a walking stick climbing a sand dune arm in arm.

(b) One person helping another person up to the top of a mound of sand underneath a cloudy sky.

Figure 1: Two different images and the text descriptions from the Flickr30K dataset. The two images and the descriptions are semantically similar however occur in different pairs. A cross association by a retrieval system will be considered an error by the currently used $R@K$ measure relying on pairwise loss functions. However, our proposed measure "SemanticMap" will consider the cross association also a semantically correct association.

In classical cross modal retrieval systems, representations of semantically related instances is learned such that

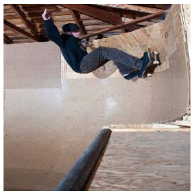
Query Image	Retrieved Text	λ
	– A man riding a skateboard at a skatepark	0.82
	– A man riding a skateboard down a ramp	0.81
	– A person riding a skateboard at a skatepark	0.78
	– A skateboarder trying to get up a ramp	0.78
	– A man riding a skateboard over an obstacle	0.77

Figure 2: Image to text retrieval result (top 5) from the MSCOCO dataset. Query and retrieved text descriptions are semantically related but not in pairs and thus $R@5 = 0$. In the last column the proposed SemanticMap (λ) scores are reported resulting $\lambda@5 = 0.79$ showing high semantic similarity.

some distance measure is minimized between the learned representations in the latent space. Once these representations are obtained, Recall-at-K ($R@K$) metric is often employed to report accuracy of pairwise retrievals. The $R@K$ measure adheres to the manually marked ground truth labels, while ignoring the semantic relationships between the query and the retrieved instances. In this paper, we pose an important question: *do cross modal retrieval systems really leverage on semantic relationships?* We find that dependence on ground truth labels may be an inefficient approach due to its inadequacy to fully exploit semantic relationships between the learned representations. For example, in Figure 2, the query image and the retrieved text descriptions are semantically related, however, $R@K$ does not capture it and reports miss retrieval.

In the existing systems, neural network based mappings have been commonly used to bridge the gap between multiple modalities [41, 56] by building a joint representation of each modality. Typically, separate networks are trained to predict features of each modality and a supervision signal is employed to reduce the distance between image and associated text descriptions [51, 8, 52, 57]. In addition, to capture text context before semantically associating it with the visual data, some techniques employ RNNs [5, 20] along with CNNs stacked in a CRNN fashion [7, 23, 9]. Though by using separate networks, these systems were able to achieve good accuracy, it incurs significant memory overhead. In many modern applications such as mobile devices, memory is a scarce resource therefore less memory demanding systems as the one proposed in this paper are more desirable.

In this paper we propose an image-text cross modal retrieval system that captures semantic similarities between images and their text neighbors without suffering from inaccuracies in the ground truth labels. That is, semantically similar text descriptions may get different labels while our system generates nearby embedding for such text descriptions. We encode text descriptions as images which leads to a single stream network which is jointly trained to bridge the gap between the two modalities. We use a loss function inspired from [53] as a supervision signal to map images and associated text descriptions “closer“ to each other in the shared latent space. We evaluate the proposed system on two publicly available image retrieval datasets including Flickr30k [39, 40] and MSCOCO [30] and obtain comparable results to the existing systems. Furthermore, we reevaluate the commonly used retrieval metric $R@K$, and we identify its inadequacy and introduce a new metric, referred to as **SemanticMap**, leveraging on semantic similarity between the images and text descriptions.

Main contributions of our paper are the following:

- We extensively evaluate the $R@K$ measure and find that it is inadequate to leverage semantic relationships. We propose a new measure SemanticMap exploiting the semantic relationships between the images and the semantically similar text descriptions.
- We propose a cross modal retrieval system using single stream network for image to text and text to image which is equally capable of mapping two different modalities in a joint embedding space without having to use separate networks for each modality.
- We encode text descriptions as images for training the single stream network for cross modal retrieval. To the best of our knowledge, this work is the first approach employing a single stream network for cross modal retrieval tasks.
- We present human baseline results on the MSCOCO test set employing the best performing network. The sole purpose of such an experiment is to correlate the limitations of $R@K$ captured by **SemanticMap** with evaluations of human experts.

The rest of the paper is structured as follows: we explore the related literature in Section 2. We explore $R@K$ and identify its inadequacy and propose a new measure in Section 3, a new cross modal retrieval system is proposed in Section 4, followed by experimental evaluations in Section 5. Human baseline results are presented in Section 6, followed by ablation study in Section 7. Finally conclusions and future directions are discussed in Section 8.

2. Related Work

Several works in the field of multimodal representation learning have been proposed in the recent years. Although each task is different from the other, the underlying principle has remained the same: to achieve semantic image-text multimodal representation. In this section we explore the related literature under four subsections.

2.1. Classical Approaches

One of the classical approaches towards image-text embedding is Canonical Correlation Analysis (CCA) [18]. This method finds linear projections that maximize the correlation between modalities. Works such as [15, 28] incorporate CCA to map representations of image and text to a common space. Although being a rather classical approach, the method is efficient enough. Recently, deep CCA has also been employed to the problem of obtaining a joint embedding for multimodal data [27]. However, the major drawback is that using CCA it is computationally expensive because it requires to load all data into memory to compute the covariance score.

2.2. Deep Metric Learning Approaches

Deep metric learning approaches have shown promising results on various computer vision tasks. Employing metric learning to multimodal tasks requires within-view neighborhood preservation constraints which are explored in several works [22, 34, 44]. Triplet networks [21, 50] along with siamese networks [17, 43, 3] have been used to learn a similarity function between two modalities. However, most of these techniques [51] require separate networks for each modality which increases the computational complexity of the whole process. Furthermore, these networks suffer from dramatic data expansion while creating sample pairs and triplets from the training set.

2.3. Ranking Supervision Signals

Many different multimodal approaches employ some kind of ranking loss function as a supervision signal. Works presented in [54, 12] employ a ranking loss which penalizes when incorrect description is ranked higher than the correct one. Similarly, the ranking loss can be employed in bi-directional fashion where the penalty is based on retrieval of both modalities.

2.4. Classification Methods

Jointly representing multiple modalities in a common space can also be employed for classification purposes. Work in [29] employs classification loss along with two neural networks for both modalities (text and image) for zero-shot learning. Work in [42] employs attention-based mechanism to estimate the probability of a phrase over different region proposals in the image. In nearly every visual question answering (VQA) method, separate networks are trained for image and text; however, [24] treats the problem as a binary classification problem by using text as input and predicting whether or not an image-question-answer triplet is correct using softmax.

In contrast to all these existing approaches, we propose a single stream network to extract representation from multiple modalities without pairwise or triplet recombination at the input.

3. Re-evaluating $R@K$ Measure

Consider a cross modal retrieval scenario with a pairwise loss function $\mathcal{L}_c(x_i, x_i^{\pm}) = d(x_i, x_i^{\pm})$; where $d(x_i, x_i^{\pm})$ is a distance metric. A common objective of the existing methods is to reduce the distance between the query image and its associated text descriptions. These systems leverage on manually marked image/text pairs only. Once the network is trained employing separate networks for each modality and jointly embedding the image/text pairs onto a latent space, $R@K$ is employed at the inference stage. However, $R@K$ is an inefficient metric since it functions on the manually imposed restriction on training and test datasets. We observe that $R@K$ is based on whether query's pair appeared or not in retrieval result. So, even if retrieval result is semantically reasonable and if query's pair did not appear in the retrieval result, the R@K score can be considerably low. For example, the word beach is semantically related to every image which contains beach environment, however, if the word does not occur along with the image pair in Top K retrieved results, $R@K$ comparatively gives low scores. The work presented in [38] argued regarding R@K on similar grounds. We believe that since it does not compute a function that can accurately map input samples, the metric ($R@K$) suffers from inaccuracy in capturing semantically similar data points [16]. In the current work, we emphasize that retrieval systems should be evaluated on their ability to bridge the gap between semantically correlated images and text descriptions. For this purpose, we introduce a new metric referred to as SemanticMap.

3.1. Proposed SemanticMap ($\lambda@K$)

One of our contribution in this work is the introduction of the metric SemanticMap which leverages on semantic relationships between different modalities. Consider an image representation in the latent space $X_l = (x_1^l, x_2^l, x_3^l, ..., x_n^l) \in \mathcal{R}^n$ belonging to l^{th} class of images and a latent text representation $Y_m = (y_1^m, y_2^m, y_3^m, ..., y_n^m) \in \mathcal{R}^n$ belonging to m^{th} class of text description. If the image and the text description is semantically similar, the latent representations should also be similar as defined below.

$$\lambda(X_l, Y_m) = \frac{\sum_{i=1}^{n} x_i^l y_i^m}{\sqrt{\sum_{i=1}^{n}(x_i^l)^2}\sqrt{\sum_{i=1}^{n}(y_i^m)^2}}; \forall X_l, Y_m \quad (1)$$

Query: A man is playing tennis on the tennis court.

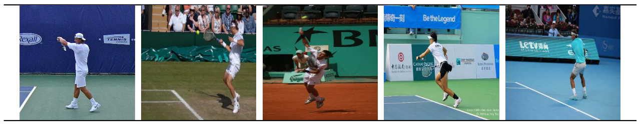


Figure 3: An example of text to image retrieval result (top 5) from the MSCOCO test set. Query and retrieved images are semantically related but not in pairs and thus $R@5 = 0$. However, $\lambda@5$ = 0.71 which shows significant semantic similarity between the query and retrieved results. (Best viewed in color)

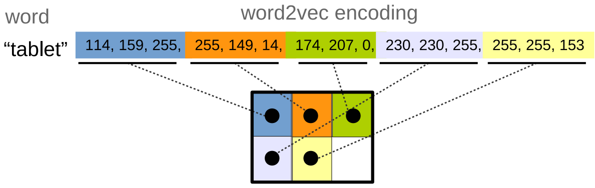


Figure 4: The word "*tablet*" is encoded into an image using Word2Vec encoding with vector length 15. Consecutive words in the text descriptions are encoded as image preserving relative position of each word. Note that words that occur in similar context will have similar embedding, thus the encoding will be similar in color space. (Best viewed in color)

We define for the top K retrieved results, where $K = 1, 5, 10$, SemanticMap ($\lambda@K$) using Equation 1:

$$\lambda@K = \frac{1}{N \cdot K} \sum_{l=1}^{c} \sum_{m=1}^{K} \lambda(X_l, Y_m) \tag{2}$$

where X_l, Y_m can be of any class and N is the total number of instances in the test dataset. Note that Equation 2 takes into consideration first K similarity index(s) $\lambda(X_l, Y_m)$ of a ranked list. To the best of our knowledge, this is the first time similarity index is computed between two different modalities.

Consider a retrieval result shown in Figure 3 from MSCOCO dataset. In the image, there are three major objects of semantic significance: man, playing and tennis court. The retrieved images and the query text description do not exist in the same ground truth pair, therefore $R@1 = 0$ and $R@5 = 0$. The retrieved images belong to 5 different ground truth classes and are listed in the order of decreasing λ score: $0.82, 0.75, 0.69, 0.68, 0.64$, resulting in $\lambda@1 = 0.82$ and $\lambda@5 = 0.71$ which shows that the retrieved images has significant semantic similarity with the query text description.

4. Proposed Cross Modal Retrieval System

One of the core ideas of this paper is to bridge the gap between the image and encoded text description. Our proposed approach eliminates the need for multiple networks for either modality, since similar results can be achieved with a single stream network. The Figure 5 visually explains the architecture of the network. The detailed proposed approach is presented in the following subsections:

4.1. Encoding Text Descriptions as Images

Semantics plays a crucial role to understand the meaning of a text description. Though humans can understand semantics easily, automatic semantic understanding is still a challenging task. Word2Vec word embedding [35] takes one step towards mathematically representing the semantic relationships between words. Its objective function causes words that occur in a similar context to have similar word embedding. Gallo et al. [13] has recently presented an encoding scheme exploiting Word2Vec to reconstruct the semantics associated with a text description as an image. Though Gallo et al. has used this encoding for only text classification, we extend it for cross modal retrieval. We employ this encoding scheme to encode text descriptions as images and use these encoded text images as input to the neural networks originally developed for image input. This encoding scheme has enabled us to use a single stream network for both text to image and image to text retrieval. The single stream network based cross modal retrieval has potential for memory efficient and computationally-inexpensive applications on low powered devices. We explain the encoding scheme to transform a text description into an image in Figure 4.

4.2. The Single Stream Network for Image to Text and Text to Image Retrieval

Our proposed method is generic and any appropriate deep network can be employed. In our implementation, we use InceptionResNet-V1 as a single stream network for text-to-image and image-to-text retrievals (Figure 5). The network is trained using both the images and the encoded text descriptions. In a cross modal dataset, suppose there are n_t text descriptions associated with n_i images in class c. Each image and the text description is input to the network and $n_t + n_i$ feature vectors f_c are obtained at the output of the network. During training, geometric center of $n_t + n_i$ feature vectors is computed and distance of each feature vector from the center is minimized.

$$d(f_c) = \sum_{i=1}^{n_t+n_i} \parallel f_c^i - \frac{1}{n_t+n_i} \sum_{j=1}^{n_t+n_i} f_c^j \parallel_2^2 \tag{3}$$

Thus during the training phase, image and the encoded text are treated in similar fashion. Thus a single stream net-

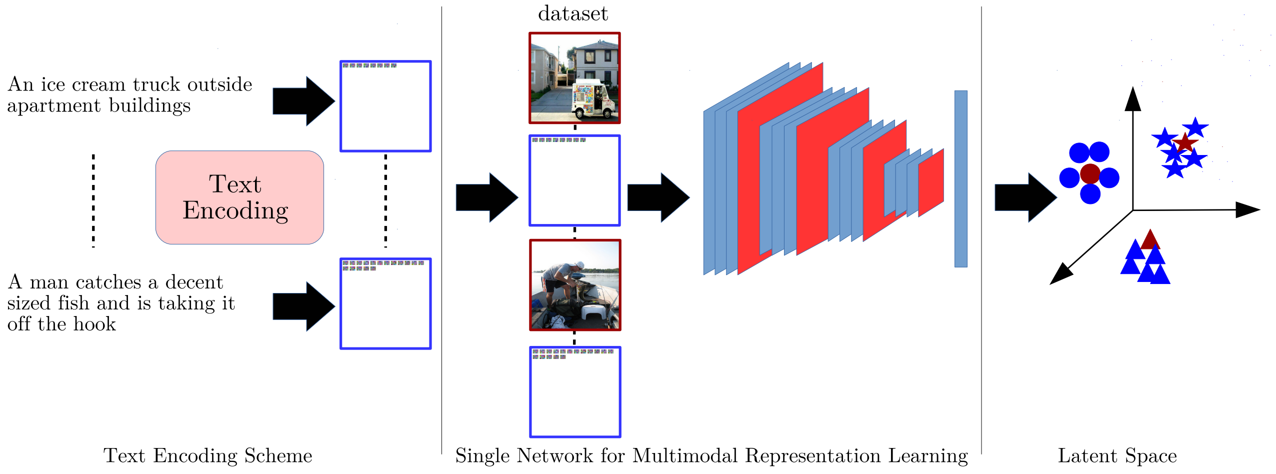

An ice cream truck outside apartment buildings
Text Encoding
A man catches a decent sized fish and is taking it off the hook
dataset
Text Encoding Scheme
Single Network for Multimodal Representation Learning
Latent Space

Figure 5: In the first phase, the proposed cross modal retrieval system transforms word embedding into encoded text which is fed to the single stream network along the images in the second phase. Finally, the supervision signal embed both images and text descriptions in the latent space. (Best viewed in color)

work can effectively bridge the gap between image and text descriptions eliminating the need for multiple networks for each modality.

In our implementation, instead of using the traditional loss function, we extend center loss for cross-modal distance learning jointly trained with softmax loss [53]. This loss function simultaneously learns all centers for features of images and encoded text descriptions in a mini-batch and penalizes the distances between each center and images along with the associated text descriptions. It thus imposes neighborhood preserving constraint within each modality as well as across the modalities. If there are m classes in a mini batch, the loss function is given by

$$\mathcal{L}(\text{mini batch}) = \frac{1}{2}\sum_{c=1}^{m} d(f_c) \tag{4}$$

This loss function minimizes the variation between image and encoded text within a class and effectively preserves the neighborhood structure. In this way, encoded text and images which are not semantically related do not occur in the same neighborhood.

5. Experiments

We evaluate the proposed retrieval system on two publicly available image-text retrieval datasets including MSCOCO [30] and Flickr30K [39, 40]. MSCOCO contains $123,287$ images, and each image is annotated with five captions. We use 1000 images for testing and the rest for training as proposed by the original authors and used by Wang et al. [51] and referred it as COCO-1k. Flickr30K contains $31,783$ images that are collected from the Flickr website along with five captions for each image. We use 1000 images for testing and rest for training as described in [25].

5.1. Implementation Details

The number of layers and the overall architecture for the network is kept the same as used by the original authors [19]. The size of the input images and text encoding is 256×256 and the size of the output feature vectors is 128-D. For optimization, we employ Adam optimizer [26] because of its ability to adjust the learning rate during training. We use Adam's initial learning rate of 0.05 and employ weight decay strategy by decaying with a factor of $5\text{e}{-5}$. The network is trained for 100 epoches. The mini batch size was fixed to randomly selected 45 images and text descriptions.

For the sake of comparison with other techniques, we use $R@K$ metric as described in [48]. We employ the $R@1$, $R@5$ and $R@10$ which means that the percentage of queries in which the first 1, 5 and 10 items are found in the ground truth. We also compute scores with the proposed SemanticMap metric in a similar fashion $\lambda@1$, $\lambda@5$ and $\lambda@10$.

5.2. Data Augmentation

We perform multiple experiments with different data augmentation schemes. A standard setting consists of one image with five associated encoded text descriptions, which we refer as *config-standard*. With further experiments on different augmentations of data, we observe that increasing number of training images do not help with the R@K scores. This is due to high semantic similarity between images of different classes due to frequent appearance of common objects for example persons, vehicles etc. Since a person or a vehicle might appear in a large number of classes,

Table 1: Comparison of the proposed system with current state-of-the-art methods using $R@K$ measure on MSCOCO and Flickr30K dataset.

	MSCOCO						Flickr30K					
Model	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 [25]	38.4	69.9	80.5	27.4	60.2	74.8	–	–	–	–	–	–
HM-LSTM [37]	43.9	-	87.8	36.1	-	86.7	–	–	–	–	–	–
m-RNN-vgg [33]	41.0	73.0	83.5	29.0	42.2	77.0	35.4	63.8	73.7	22.8	50.7	63.1
Order-embedding [48]	46.7	-	88.9	37.9	-	85.9	–	–	–	–	–	–
m-CNN(ensemble) [32]	42.8	73.1	84.1	32.6	68.6	82.8	33.6	64.1	74.9	26.2	56.3	69.6
Str. Pres. [51]	**50.1**	**79.7**	**89.2**	**39.6**	**75.2**	**86.9**	40.3	**68.9**	**79.9**	29.7	**60.1**	**72.1**
Two-Way [10]	–	–	–	–	–	–	**49.8**	67.5	-	**36.0**	55.6	-
TextCNN [38]	13.6	39.6	54.6	10.3	35.5	55.5	–	–	–	–	–	–
FV-HGLMM [38]	14.3	40.5	55.8	12.7	39.0	57.2	–	–	–	–	–	–
Our Work (*cfg-std*)	13.2	30.4	41.9	12.2	33.0	46.7	10.5	26.2	36.8	8.2	22.82	32.0
Our Work (*cfg-2*)	13.0	32.9	46.0	12.94	36.62	49.94	27.0	43.1	52.1	23.7	43.6	53.4
Our Work (*cfg-3*)	40.0	64.4	76.7	30.9	62.7	73.7	–	–	–	–	–	–

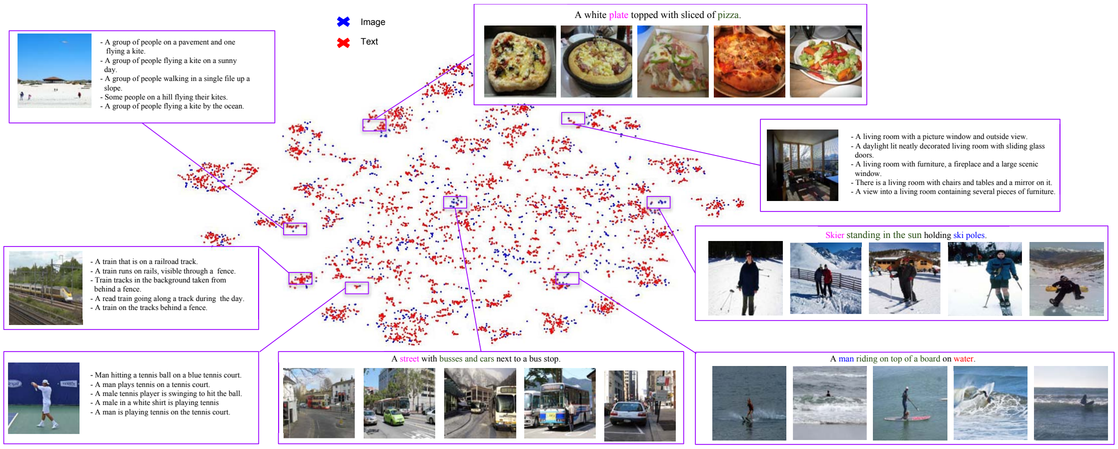


Figure 6: Embedding of MSCOCO test set in latent space visualized using t-SNE [47]. The semantically similar text feature vectors (shown in red) and the image feature vectors (shown in blue) are close in the embedding space. Few bidirectional retrieval results are also shown which are semantically similar. (Best viewed in color)

thus these may get wrongly associated. However, with distinct combinations of data, we observe that increasing encoded text descriptions improve the R@K score. In addition to image and its associated encoded text descriptions, we crop encoded text to a size of 227 × 227 and feed its enlarged version to the network. With this configuration, we have one image along with its horizontally flipped version and ten different encoded text descriptions, referred to as *config-2*. However, since the method does not require loading all the data in memory at once, bottlenecks are avoided. We report results on both configurations in Table 1. Furthermore, we modify *config-2* where we down sample the images to 128 × 128 for MSCOCO dataset, referred to as *config-3*, and report the results in Table 1. Note that *config-3* did not yield comparative results on Flickr30k dataset so we ignore it while reporting results.

5.3. Comparison with $R@K$ Measure

For comparison with other approaches, we computed $R@K$ for text-to-image and image-to-text retrieval. Compared to the current state-of-the-art, our method performance is comparatively low. The main reason is due to the fact that $R@K$ is based on whether query's pair appeared or not in the retrieval result. So, even if retrieval result is semantically similar (Table 1) and if query's pair did not appear in the retrieval result, the $R@K$ score is considerably low. Furthermore, we recommend that retrieval systems should be evaluated on the efficiency of bridging the

Table 2: Comparison of the proposed system (*cfg-3*) with current state-of-the-art method Structure Preserving [51] using $\lambda@K$ measure. Note that the Structure Preserving method is the best existing method in Table 1.

	MSCOCO						Flickr30K					
Model	Image-to-Text			Text-to-Image			Image-to-Text			Text-to-Image		
	$\lambda@1$	$\lambda@5$	$\lambda@10$	$\lambda@1$	$\lambda@5$	$\lambda@10$	$\lambda@1$	$\lambda@5$	$\lambda@10$	$\lambda@1$	$\lambda@5$	$\lambda@10$
Str. Pres.	67.24	64.63	62.74	64.07	59.29	**56.30**	**62.30**	**59.59**	**57.79**	**59.05**	**54.60**	**51.97**
Our Work	**68.67**	**65.25**	**62.86**	**66.70**	**59.42**	54.46	49.13	45.52	43.33	43.97	39.22	36.43

Table 3: Comparison of the proposed system (*cfg-3*) with current best method using $\lambda@K$ on MSCOCO-1k test set. Best K retrievals out side the pairs are selected demonstrating generalization performance in the wild.

Model	Image-to-Text			Text-to-Image		
	$\lambda@1$	$\lambda@5$	$\lambda@10$	$\lambda@1$	$\lambda@5$	$\lambda@10$
Str. Pres.	64.94	62.61	61.02	61.88	58.01	**55.34**
Our Work	**67.57**	**64.17**	**61.81**	**65.42**	**58.36**	53.55

Table 4: Quantitative results, $R@10$, of ablation model on MSCOCO test set. Note that Ab. Net refers to ablation model with *cfg-3*.

Model	Image-to-Text	Text-to-Image
Ab. Net.	21.30	24.92

gap between multiple modalities and not on $R@K$ which does not leverage on semantic relationships between modalities and rather matches Top-K retrieved results with ground truth class boundaries. As an example, the word beach is semantically related to every image which contains beach environment, however, if the word does not occur along with the image pair in Top K retrieved results, $R@K$ will yield low score. Park et al. [38] have also criticized $R@K$ measure on similar grounds.

The Figure 6 is t-SNE [47] visualization of features from MSCOCO test set i.e. 1k images with five text descriptions for each image. Once the network is trained on the dataset, features of the test set are extracted from the model and are fed to t-SNE, Visualization verify that the proposed single stream network is capable of bridging gap between image and encoded text descriptions in a latent space.The image and text encoding description are overlapped and distributed enough for being discriminated in retrieval. Some bidirectional retrieval results from MSCOCO test set are also present in Figure 6. Query and retrieved objects are semantically related in these examples.

5.4. Comparison With $\lambda@K$ Measure

In this section, we explore $\lambda@K$ results on the MSCOCO and Flickr30K datasets with our approach and compare it with the best performing approach on $R@K$ which is named as "Structure Preserving" [51] as shown in Table 2. With this comparison, we show that our model achieves better performance than the current best approach using $\lambda@K$ measure. In other words, our model effectively bridges the gap between image and text descriptions. It is evident from Table 2 that results on Flickr30K show some variation from that of MSCOCO. The reason is Flickr30K test set has less semantic similarity compared to the MSCOCO dataset.

In this work, we leverage on capturing relationships between multiple modalities without the pairwise restrictions. We point out that the performance of retrieval systems should be based on their ability to effectively bridge gap between the information from multiple sources. In Table 3 we restrict $\lambda@K$ to discard pairwise retrieval results i.e. if a text description is retrieved using an image as query belongs to the same class as the query image, it is ignored. Rather we enforce $\lambda@K$ to consider image and text descriptions not occurring in pairs but are semantically similar. This helps to quantitatively evaluate how our method generalizes compared to the state-of-art methods, in the wild when no ground truth is available. We use implementation provided by Structure Preserving [51] to extract features and compute the scores. The results indicate the limitations of networks trained using pairwise strategies.

6. Human Baseline

In this section we account for the semantically reasonable queries with the help of 3 human annotators. We use the best performing network with our proposed approach on MSCOCO and extract embedding for MSCOCO-1k test set. We compute $R@10$ on the test set and manually annotate $R@K$ missed queries as semantically related or not. If the human evaluator considers the retrieved result to be semantically aligned with the query, we consider it a hit. The human annotators are domain experts and final results are reported by averaging the scores of 3 different people in Table 6. To the best of our knowledge, this is the first attempt a human baseline has been reported on cross-modal

Table 5: Ablation Study. Qualitative analysis of Image-to-Text retrieval result on ablation model and *cfg-3* which is best configuration on MSCOCO test set. Groundtruth matched text are marked as red and bold, while some text sharing semantics as ground truth, but in different class pairs, are marked as blue and bold. (Best viewed in color)

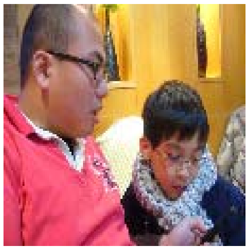
Query	Best Configuration Model cfg-3	Ablation Model (SqueezeNet)
	– **A man and boy are looking at a cellphone.** – **A father and his son is looking at a cell phone.** – **A man is showing a boy with a scarf his cell phone.** – Two college graduates pack up after a long day. – **A young boy with his father spending time.**	– A woman and a girl sitting at a dinner table talking. – Two men and a woman holding wine glasses. – A girl reclining on her bed and reading a book. – A woman sitting at a long table with an unhappy look on her face. – Two women sitting across a table from each other.

Table 6: Human baseline results. 3 human experts evaluated the queries on Image-to-Text retrieval of MSCOCO test set. The last row indicates the average of all 3 results.

Human Experts	Image-to-Text
Expert # 1	85.40
Expert # 2	85.00
Expert # 3	83.40
R@10	84.60

retrieval systems for statistical analysis. The sole purpose of this experiment is to examine the statistical gain in terms of retrieval where the missed $R@K$ results are re-evaluated by human evaluators. The objective of this study is to present a cross-evaluation to indicate the limitations.

7. Ablation Study

In this section we 'take-apart' the individual components to identify the inter-dependence. This is to identify the most important parts of the pipeline. As a part of the ablation study, different configurations and their results on $R@K$ are reported in Table 1. To identify the important of single stream network, we replace it with another network, SqueezeNet, and report quantitative results of the network, see Table 4. We call the pipeline with SqueezeNet as ablation model. Furthermore, we also report the qualitative results of ablation model and best model on MSCOCO to establish difference between learning ability of both networks, see Table 5.

8. Conclusion and Future Work

In this work a new performance measure is suggested for the cross modal retrieval systems. Currently used measure $R@K$ do not capture semantic relationship between the query and the retrievals, rather adhere to the pairs in the dataset. The new proposed measure captures semantic similarity between the query and the retrievals. Moreover, a novel cross modal retrieval system is also proposed which uses a single stream network for text to image and image to text retrieval. The proposed network is trained using center loss as the supervision signal. The text descriptions are also encoded as images which enabled the system to use existing deep neural networks for both text and the images. The proposed system was able to bridge the gap between two different modalities by projecting semantically similar text and images close by in the latent space. The proposed system is evaluated on two publicly available datasets including MSCOCO and Flickr30k. The proposed model have shown comparable results with the current state-of-the-art methods. Finally, we present human baseline to account for the semantically reasonable queries with the help of 3 human annotators and report averaged results.

This work can be extended to other cross modal retrieval tasks such as audio to image and video to text etc. Furthermore, we would like to explore different data augmentation schemes to find more effective configurations which help semantically aligning images and the text descriptions in a neighborhood. Research towards different encoding schemes to exploit the semantic dependencies within text is another important research direction.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. pages 2425–2433, 2015.

[2] R. Arandjelovic and A. Zisserman. Look, listen and learn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 609–617, 2017.

[3] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pages 737–744, 1994.

[4] G. Chrupała, L. Gelderloos, and A. Alishahi. Representations of language in a model of visually grounded speech signal. *Association for Computational Linguistics*, 2017.

[5] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Gated feedback recurrent neural networks. In *International Conference on Machine Learning*, pages 2067–2075, 2015.

[6] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. pages 1086–1090, 2018.

[7] Y.-A. Chung and W.-H. Weng. Learning deep representations of medical images using siamese cnns with application to content-based image retrieval. *arXiv preprint arXiv:1711.08490*, 2017.

[8] S. Dey, A. Dutta, S. K. Ghosh, E. Valveny, J. Lladós, and U. Pal. Learning cross-modal deep embeddings for multi-object image retrieval using text and sketch. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 916–921, 2018.

[9] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[10] A. Eisenschtat and L. Wolf. Linking image and text with 2-way nets. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4601–4611, 2017.

[11] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015.

[12] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

[13] I. Gallo, S. Nawaz, and A. Calefati. Semantic text encoding for text classification using convolutional neural networks. In *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, volume 5, pages 16–21, 2017.

[14] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in neural information processing systems*, pages 2296–2304, 2015.

[15] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *European Conference on Computer Vision*, pages 529–545, 2014.

[16] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1735–1742, 2006.

[17] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg. Matchnet: Unifying feature and metric learning for patch-based matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3279–3286, 2015.

[18] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[20] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[21] E. Hoffer and N. Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92, 2015.

[22] J. Hu, J. Lu, and Y.-P. Tan. Discriminative deep metric learning for face verification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1875–1882, 2014.

[23] Y. Huang, W. Wang, and L. Wang. Instance-aware image and sentence matching with selective multimodal lstm. 2017.

[24] A. Jabri, A. Joulin, and L. van der Maaten. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739, 2016.

[25] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[26] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

[27] B. Klein, G. Lev, G. Sadeh, and L. Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4437–4446, 2015.

[28] B. Klein, G. Lev, G. Sadeh, and L. Wolf. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

[29] J. Lei Ba, K. Swersky, S. Fidler, et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4247–4255, 2015.

[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755, 2014.

[31] Y. Liu, L. Liu, Y. Guo, and M. S. Lew. Learning visual and textual representations for multimodal matching and classification. *Pattern Recognition*, 84:51–67, 2018.

[32] L. Ma, Z. Lu, L. Shang, and H. Li. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE international conference on computer vision*, pages 2623–2631, 2015.

[33] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *International Conference on Learning Representations*, 2015.

[34] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *Computer Vision–ECCV 2012*, pages 488–501. 2012.

[35] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[36] A. Nagrani, S. Albanie, and A. Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[37] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Hierarchical multimodal lstm for dense visual-semantic embedding. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 1899–1907, 2017.

[38] G. Park and W. Im. Image-text multi-modal representation learning by adversarial backpropagation. *arXiv preprint arXiv:1612.08354*, 2016.

[39] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[40] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *IJCV*, pages 123(1):74–93, 2017.

[41] R. Qiao, L. Liu, C. Shen, and A. v. d. Hengel. Visually aligned word embeddings for improving zero-shot learning. *arXiv preprint arXiv:1707.05427*, 2017.

[42] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834, 2016.

[43] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[44] B. Shaw, B. Huang, and T. Jebara. Learning a distance metric from a network. In *Advances in Neural Information Processing Systems*, pages 1899–1907, 2011.

[45] K. Sohn, W. Shang, and H. Lee. Improved multimodal deep learning with variation of information. In *Advances in Neural Information Processing Systems*, pages 2141–2149, 2014.

[46] N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012.

[47] L. Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245, 2014.

[48] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun. Order-embeddings of images and language. 2016.

[49] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[50] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393, 2014.

[51] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5005–5013, 2016.

[52] L. Wang, Y. Li, and S. Lazebnik. Learning two-branch neural networks for image-text matching tasks. *arXiv preprint arXiv:1704.03470*, 2017.

[53] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515, 2016.

[54] J. Weston, S. Bengio, and N. Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *International Joint Conference on Artificial Intelligence*, 2011.

[55] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[56] Y. Zhang, B. Gong, and M. Shah. Fast zero-shot image tagging. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[57] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, and Y.-D. Shen. Dual-path convolutional image-text embedding. *arXiv preprint arXiv:1711.05535*, 2017.