

Joint Image-Text Clustering using Deep Neural Networks

Mahyar Khayatkhoei

Aditya Chukka

Chaitanya Mitash

Department of Computer Science, Rutgers University

{ m.khayatkhoei, aditya.chukka, cm1074 }@cs.rutgers.edu

Abstract

Nowadays there is a large amount of image and text data available in several large databases, however, properly aligned image and sentence data, a crucial requirement of any joint learning model, is quite hard to come by. To address this problem, we present a system that can jointly cluster images and text descriptions, and provide aligned data from unaligned uni-modal databases. The contribution of this work is twofold: we first present a system that learns a common embedding space in which image and text description form valid and sensible clusters. Second, we use our system for image captioning and retrieval tasks, and compare to the state of the art approaches. We use Caltech Birds Dataset [29] as reference data set for training and testing.

1. Introduction

Reading a Wikipedia article, containing many sentences and a few images, humans can effectively recognize the alignment between sentences and images, and thus learn in-depth knowledge about the topic of the article. However, finding such alignments is a hard task for visual recognition models.

We hypothesize that this astonishing ability of humans to realize the alignments between sentences and images is mostly rooted in deep understanding of similarities within text and image domains, separately. Understanding these similarities allows humans to extend the alignment knowledge from a handful of seen image text pairs to billions of unseen such pairs.

There has been a wide variety of work on similar tasks, mostly under the captioning and correlation literature. However, majority of these tasks are focused on learning from the text image pairs alone [10, 31], and to the best of our knowledge no previous work has used the separate similarity within each sub domain to extend the information available to the model.

In this work, we present a model that learns associations

between images and text in a multi-modal space, not only using the image sentence pairs, but also the structure of image and text domains separately. This model creates a shared space where a corpus of unaligned images and texts can be clustered jointly. The results can be a solution for obtaining aligned data out of many dense but unaligned data sets.

We will train and evaluate our model on the Caltech Birds Dataset [29]. We also compare our trained model in an image retrieval and captioning task with state of the art approaches. While we might not beat these models we hope that this might be a useful thought to consider.

2. Prior Work

Several attempts have been made in the past to describe the contents of images. Barnard et al. and Socher et al. [1, 26] studied the multimodal correspondences between words and images to annotate segments of images. Some of the works [19, 18, 4, 6] reason about the scene as a whole by inferring about the objects and their spatial distribution. However, these approaches were focused on labeling objects instead of a high-level description of the scene.

Many approaches formulated this task as a ranking problem, where we find the most congruent description for the test image from the sentences in training data [32, 27, 3, 23, 30] or where training sentences are broken down and attached together to describe test data [15, 16, 20]. In general, this formulation is tackled by two family of methods. One is based on learning the latent space of image and text using a canonical correlation objective [31, 7] and the other approach aligns the fragments extracted from image and sentences [9].

Another formulation is generating a new description (i.e not present in the training data) for the input image during testing. Kiros et al. [11, 12] proposed a multimodal log-bilinear model which could generate sentences to describe images. The Recurrent Neural Networks (RNN) in the above approach used a fixed window context where as the most recent approach by Kaparthy et al. [10] used a probability distribution over the next word in a sentence of

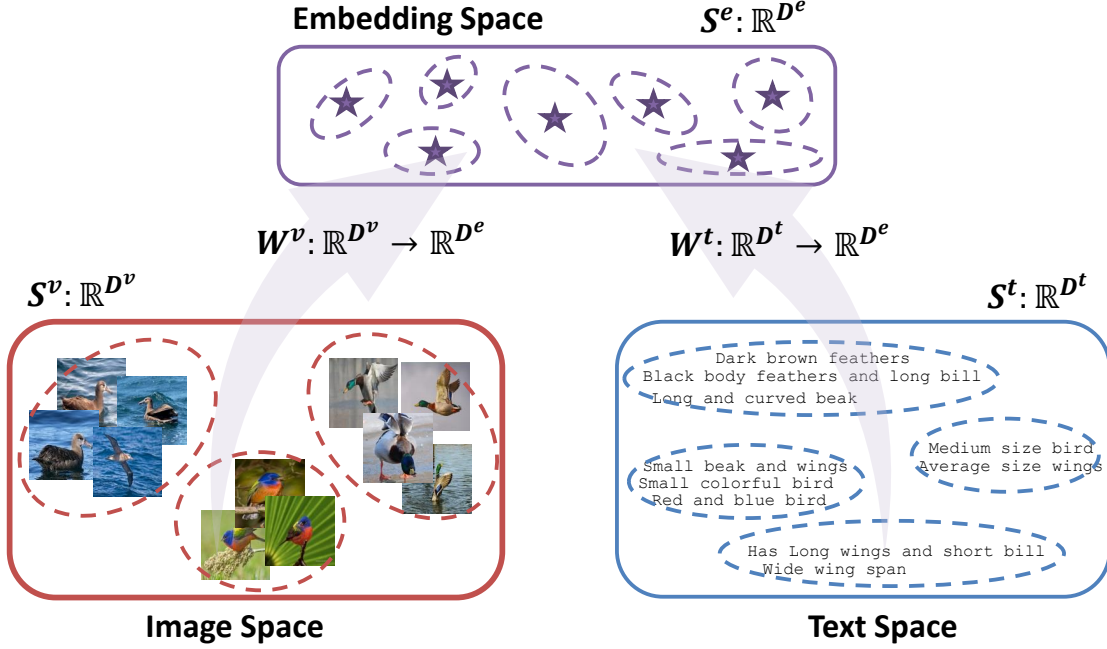


Figure 1: Overview of our model: image and text description spaces are projected into a shared embedding space, such that clusters in this shared space keep uni modal clusters intact, and image text pairs in training set fall within same cluster, i.e. relevant images and text descriptions are clustered together. Note that sentences in text space may not appear in training pairs with any of the images in image space, and vice versa.

all previously generated words. This approach achieves the current state of the art performance in both the ranking and generative evaluation frameworks.

Convolutional Neural Networks (CNNs) [17, 14] have been widely used for representing images in higher-level feature space since it has emerged as a powerful tool for image classification and object detection [25]. Pretrained words vectors [22, 8, 2] are used to obtain lower dimension feature spaces in texts. Recurrent Neural Networks (RNNs) [21, 28] are extensively used in language modeling.

Finally, Gong et al.[5] used weakly annotated photo collections to improve image-sentence embeddings. However, our approach is unique in the way that it uses associations in individual image and text space, alongside the paired training examples to learn the alignment.

3. Our Model

3.1. Overview

The goal of our model is to find an embedding space where images and text descriptions (at sentence level) can be clustered jointly. More formally, assume we have images $v_i \in S^v$ and sentences $t_i \in S^t$, where $S^v: \mathbb{R}^{D^v}$ is the D^v dimensional image space, and $S^t: \mathbb{R}^{D^t}$ is the D^t dimensional text space. We first form clusters C^v and C^t

in image and text spaces separately. Next, we seek W^v and W^t transformations which can project the image and text spaces to a common embedding space $S^e: \mathbb{R}^{D^e}$, that is $W^v \cdot v_i \in S^e$ and $W^t \cdot t_i \in S^e$, subject to these constraints:

- Clusters in image space, C^v , are not violated by clusters in S^e .
- Clusters in text space, C^t , are not violated by clusters in S^e .
- Aligned text-image pairs that appear in training set, i.e. (v_i, t_i) , fall within the same cluster in S^e .

Note that W^v and W^t are not simple matrices, they represent deep neural networks, this formulation is used only for brevity. A variant of CNN model and RNN model will be used for W^v and W^t respectively. See Fig. 1 for an overview of our model.

3.2. Problem Statement

Given a set of sentences/captions T and a set of images V , we want to find clusters C over the joint image-sentence space such that:

$$\forall x \in c_i : E_{y \in c_i} [f(x, y)] \geq E_{y \notin c_i} [f(x, y)] + \delta \quad (1)$$

where $f(x, y)$ is a compatibility function with domain in $\mathbb{R}$, and $\delta > 0$ is some margin. Here, x and y can represent both image and text.

The dataset we used (Caltech Birds Dataset CUB200-2011 [29]) is similar to the one used in the work of Reed et al. 2016 [24]. Here, each image has a set of 10 corresponding sentences.

The encoders are same as the ones used in the work of Kiros et al. 2014 [13] depicted in Fig. 2. Image encoder is a **Deep Convolutional Network** (CNN) with an output dimension as 4096 and the Text encoder is a **Gated Recurrent Unit** (GRU) with an output dimension of 1024. These two networks are used to extract feature vectors $v \in \mathbb{R}^{D^V}$ and $t \in \mathbb{R}^{D^T}$ from image and text input respectively. As reported earlier, the model tries to learn the weights (W^V , W^T) such that

$$\begin{aligned} W^V : \mathbb{R}^{D^V} &\rightarrow \mathbb{R}^{D^E} \\ W^T : \mathbb{R}^{D^T} &\rightarrow \mathbb{R}^{D^E} \end{aligned}$$

and use them to embed image and text feature vectors into a joint embedding space: $\hat{v} = W^V.v$ and $\hat{t} = W^T.t$

We define a contrastive loss function such that when minimized, it captures the desired effects mentioned in previous section. In the following we discuss each part of this objective, targeting image and text projection objectives.

1

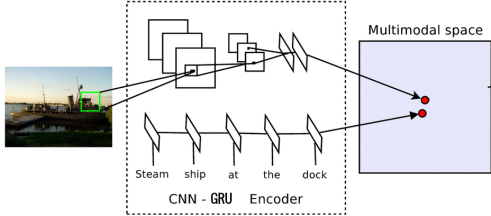


Figure 2: Encoder: A **deep convolutional network** (CNN) and a **Gated Recurrent Unit** (GRU) for learning a joint image-sentence embedding 1

3.3. Image Penalty Scores

- Alignment: this score tries to encourage image feature vectors to be mapped closer to true text than false text.

$$S_1(z) = \sum_{p \notin C_z^T} \max(0, \alpha + f(z, p) - E_{y \in C_p^V} [f(y, p)]) \quad (2)$$

¹ Image Courtesy : Ryan Kiros, Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models

- Mixing: this score tries to encourage image feature vector to be mapped closer to true text than false image.

$$S_2(z) = \sum_{y \notin C_z^V} \max(0, \alpha + f(z, y) - E_{p \in C_y^T} [f(y, p)]) \quad (3)$$

- Clustering: this score tries to encourage image feature vector to be mapped closer to true image than false image.

$$S_3(z) = \sum_{y \notin C_z^V} \max(0, \alpha + f(z, y) - E_{u \in C_y^V} [f(y, u)]) \quad (4)$$

- z, y, u denote images, and p denotes text. C_z^T is all clusters in Text space that contain at least one text in pair with image z . C_p^V is all clusters in Image space that contain at least one image in pair with text p . C_z^V is the cluster of image z in Image space. $f(a, b) = \hat{a}^T \hat{b}$ where $\hat{a}, \hat{b}$ denote normalized projections of a and b into embedding space respectively.

3.4. Text Penalty Scores

- Alignment: this score tries to encourage text feature vector to be mapped closer to true image than false image.

$$S_4(p) = \sum_{z \notin C_p^V} \max(0, \alpha + f(z, p) - E_{q \in C_z^T} [f(z, q)]) \quad (5)$$

- Mixing: this score tries to encourage text feature vector to be mapped closer to true image than false text.

$$S_5(p) = \sum_{q \notin C_p^T} \max(0, \alpha + f(p, q) - E_{z \in C_q^V} [f(z, q)]) \quad (6)$$

- Clustering: this score tries to encourage text feature vector to be mapped closer to true text than false text.

$$S_6(p) = \sum_{q \notin C_p^T} \max(0, \alpha + f(p, q) - E_{r \in C_q^T} [f(q, r)]) \quad (7)$$

- z denotes images, and p, q denote text. C_p^V is all clusters in Image space that contain at least one image in pair with text p . C_z^T is all clusters in Text space that contain at least one text in pair with image z . C_p^T is the cluster of text p in Text space. $f(a, b) = \hat{a}^T \hat{b}$ where $\hat{a}, \hat{b}$ denote normalized projections of a and b into embedding space respectively.

3.5. Final objective

We take a sum over all previous penalty scores for images and text, over the whole dataset:

$$Loss = \sum_{z \in S_{paired}^V} S_1(z) + S_2(z) + S_3(z) + \sum_{p \in S_{paired}^T} S_4(p) + S_5(p) + S_6(p) \quad (8)$$

$S_{paired}^V, S_{paired}^T$ are set of images and text that come in pairs in training data respectively.

4. Evaluation

4.1. Implementation Details

We used **Lasagne**(a light weight python implementation for **Theano**) for implementing the deep neural networks. Our code is based on the work of Kiros et al. 2014 [13]. The image features are extracted from the fully connected layer "fc7" of VGG-19 model trained on ImageNet. Sentence features are extracted from a GRU model with a fixed window size of 100. **ADAM** algorithm is used for optimization.

From the 200 classes in CUB200-2011 dataset, we use 100 for training, 50 for validation, and 50 for test. The split is the same as used by Reed et al. [24]. Classes are assumed as ground truth image and text clusters, however our model works with any form of initial clustering in separate text and image space. The dimension of the embedding space is 1024 where as the image features space is of dimension 4096 (due to the final fully connected layer of VGG), and the dimension of text features is 1024. Training our model on a **GeForce Titan** with 8GiB memory for 15-20 epochs with a batch size of 256 takes about 1 hour.

5. Results

5.1. Experiment I: Captioning

Our first evaluation uses Recall@K, which is the percentage of images for which any of the K-retrieved caption is a correct caption (one of the 10 ground truth captions for that image). We also compute the median rank of the closest correct caption. Table 1 and Table 2 show the results for image to text and text to image respectively.

Even though this evaluation is not natural for the cluster alignment objective we proposed, with our experiments we prove that we can still match the accuracy provided by Kiros et al. on the CUB Birds dataset.

The overall low recall($\approx 10\%$) for all the methods can be justified by the fact that the dataset contains several instances of the same bird with minor differences in the images/captions. However, relatively low value of median

rank suggests, we were still able to get actual images and captions close in the embedded space.

method	R1	R5	R10	Medr
Kiros et al.	1.4	6.0	11.1	101
Alignment loss	1.0	5.3	9.4	137
Combined loss	1.5	5.3	9.4	133

Table 1: Recall@K and median rank for image to text retrieval on the CUB Bird dataset

method	R1	R5	R10	Medr
Kiros et al.	1.1	4.4	7.6	163
Alignment loss	0.7	3.1	5.7	188
Combined loss	0.7	3.1	5.9	185

Table 2: Recall@K and median rank for text to image retrieval on the CUB Bird dataset

5.2. Experiment II: Clustering

For this experiment, we project the test images and captions to the embedded space and cluster these vectors using an off the shelf clustering algorithm like KMeans. We use KMeans because in this experiment we know the exact number of ground truth clusters (50 classes). This embedded space clustering is then compared to the ground truth clusters to evaluate the precision and recall, i.e. for every pair of elements (image-image, image-text, text-text) in the embedding space we validate whether they conform with the ground truth clusters.

The results in Table 3 and Table 4 demonstrate that our method was able to bring together and cluster effectively all images and text describing a common category. The difference is quite significant with respect to image-text pairs, for which the previous approaches only reasoned about the relative distances between an image and text pair compared to other text or vice versa.

method	Precision	Recall
Kiros et al.	17.2	20.9
Alignment loss	17.9	19.7
Combined loss	23.9	25.8

Table 3: precision and recall for all pairs in the embedded space clusters

method	Precision	Recall
Kiros et al.	0.000	0.000
Alignment loss	0.000	0.000
Combined loss	20.3	19.2

Table 4: precision and recall for image-text pairs in the embedded space clusters

5.3. Illustrative Results

In this section we show images and text randomly selected from 3 clusters on the embedding space generated by our model for the test set. These examples are not cherry picked and have been selected by a fair random process. You can see that the clusters are reasonable: relevant images and text appear in same clusters. While some images appear to be incorrectly clustered, their misclustering can be justified by some similarities either to texts or images of their clusters, e.g. the white bird with red leg is clustered with colorful birds (2nd row, 3rd column) or more background 4th row, 3rd column

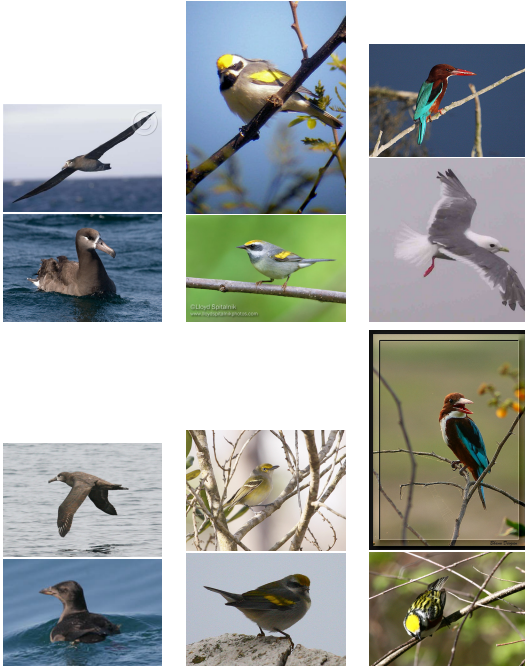


Table 5: Random Images from the Test Set that are in the same cluster. Each column shows one cluster.

In the figure 5 images in the same column belong to similar cluster. In each column you can find 4 images and there are 3 columns. So for each clusters (i.e each column) we have picked three captions each.

- Cluster/Column - 1

- a large bird with large black wings, a gray body, and large hooked gray beak.
- this large bird is almost all grey with a long hooked bill.
- this water bird has white around its eyes, grey feathers and a long hooked bill.

- Cluster/Column - 2

- the captions are - a mostly grey bird with a small head and large body, a black band across the eyes, yellow on the top of the head, and yellow markings on the wings
- this bird is white, yellow, and brown in color, with a black beak
- a petite bird with mostly grey wing feathers, yellow wing bars, and a yellow nape.

- Cluster/Column - 3

- bird has a very large red beak and blue wing feather and dark red breast and head
- the bird has teal wing fringes with a black wing center, a white neck and small necklace, and a long, nearly flat beak with rust colored head
- this bird has wings that are grey and has red feet.

5.4. Conclusion and Future Work

We noticed that the problem of finding a joint clustering shared space is much more challenging than simply finding nearest captions, and our approach, while very limited, has the potential to address this problem to some extent. Taking advantage of good ground truth clusters is indeed an important part of this process, enabling us to use the information from separate clusters together with joint data.

In the future, we aim to compare with more approaches including Deep CCA, improve on the structure of the encoders (for example: we could use a probabilistic approach like Karpathy [10] for training GRU instead of a fixed size window). We also would try different data sets to study how the characteristics of data can affect our approach.

References

- [1] K. Barnard, P. Duygulu, D. Forsyth, N. D. Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *JMLR*, 2003. 1
- [2] Y. Bengio, H. Schwenk, J.-S. Senecal, F. Morin, and J.-L. Gauvain. Neural probabilistic language models. *Innovations in Machine Learning*, 2006. Springer. 2
- [3] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. *ECCV*, 2010. 1
- [4] S. Fidler, A. Sharma, and R. Urtasun. A sentence is worth a thousand pixels. *CVPR*. 1
- [5] Y. Gong, et al. Improving image-sentence embeddings using large weakly annotated photo collections. *European Conference on Computer Vision*, 2014. 2
- [6] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1–8, IEEE, 2009. 1
- [7] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: data, models and evaluation metrics. *Journal of Artificial Intelligence Research*. 1
- [8] R. Jeffrey Pennington and C. Manning. Glove: Global vectors for word representation. 2014. 2
- [9] A. Karpathy, A. Joulin, and F.-F. Li. Deep fragment embeddings for bidirectional image sentence mapping. *arXiv:1406.5679*. 1
- [10] A. Karpathy and F.-F. Li. Deep visual-semantic alignments for generating image descriptions. *CVPR*. 1, 5
- [11] R. Kiros, R. Salakhutdinov, , and R. Zemel. Multi-modal neural language models. *International Conference on Machine Learning*, page 595–603, 2014. 1
- [12] R. Kiros, R. Salakhutdinov, , and R. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv:1411.2539*, page 595–603, November 2014. 1
- [13] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 3, 4
- [14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012. 2
- [15] P. Kuznetsova, V. Ordonez, A. C. Berg, and T. L. Berg. Collective generation of natural image descriptions. *ACL*, 2012. 1
- [16] P. Kuznetsova, V. Ordonez, T. L. Berg, U. C. Hill, and Y. Choi. Treetalk: Composition and compression of trees for image descriptions. *Transactions of the Association for Computational Linguistics*, 2(10):351–362, 2014. 1
- [17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 1998. 2
- [18] L.-J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. *ICCV*. 1
- [19] L.-J. Li, R. Socher, and L. Fei-Fei. Classification, annotation and segmentation in an automatic framework. *IEEE Conference*. 1
- [20] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing simple image descriptions using web-scale n-grams. *CoNLL*, 2011. 1
- [21] T. Mikolov, M. Karafi, L. Burget, J. Cernock, , and S. Khudanpur. Recurrent neural network based language model. *INTERSPEECH*, 2010. 2
- [22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *NIPS*, 2013. 2
- [23] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. *NIPS*, 2011. 1
- [24] S. Reed, Z. Akata, B. Schiele, and H. Lee. Learning deep representations of fine-grained visual descriptions. *arXiv preprint arXiv:1605.05395*, 2016. 3, 4
- [25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. 2014. 2
- [26] R. Socher and L. Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. *CVPR*, 2010. 1
- [27] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *TACL*, 2014. 1
- [28] I. Sutskever, J. Martens, and G. E. Hinton. Generating text with recurrent neural networks. *INTERSPEECH, ICML* 2011. 2
- [29] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011. 1, 3
- [30] M. S. Y. Ji and T. Darrell. Learning cross-modality similarity for multinomial data. *ICCV*, 2011. 1
- [31] F. Yan and K. Mikolajczyk. Deep correlation for matching images and text. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 1
- [32] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2014. 1