Full-Network Embedding in a Multimodal Embedding Pipeline

Armand Vilalta
Barcelona Supercomputing Center (BSC)
`armand.vilalta@bsc.es`

Dario Garcia-Gasulla
Barcelona Supercomputing Center (BSC)
`dario.garcia@bsc.es`

Ferran Parés
Barcelona Supercomputing Center (BSC)
`ferran.pares@bsc.es`

Jonatan Moreno
Barcelona Supercomputing Center (BSC)
`jonatan.moreno@bsc.es`

Eduard Ayguadé
Barcelona Supercomputing Center (BSC)
Universitat Politècnica de Catalunya
`eduard.ayguade@bsc.es`

Jesus Labarta
Barcelona Supercomputing Center (BSC)
Universitat Politècnica de Catalunya
`jesus.labarta@bsc.es`

Ulises Cortés
Barcelona Supercomputing Center (BSC)
Universitat Politècnica de Catalunya
`ia@cs.upc.edu`

Toyotaro Suzumura
Barcelona Supercomputing Center (BSC)
IBM T.J. Watson
`tsuzumura@us.ibm.com`

Abstract

The current state-of-the-art for image annotation and image retrieval tasks is obtained through deep neural networks, which combine an image representation and a text representation into a shared embedding space. In this paper we evaluate the impact of using the Full-Network embedding in this setting, replacing the original image representation in a competitive multimodal embedding generation scheme. Unlike the one-layer image embeddings typically used by most approaches, the Full-Network embedding provides a multi-scale representation of images, which results in richer characterizations. To measure the influence of the Full-Network embedding, we evaluate its performance on three different datasets, and compare the results with the original multimodal embedding generation scheme when using a one-layer image embedding, and with the rest of the state-of-the-art. Results for image annotation and image retrieval tasks indicate that the Full-Network embedding is consistently superior to the one-layer embedding. These results motivate the integration of the Full-Network embedding on any multimodal embedding generation scheme, something feasible thanks to the flexibility of the approach.

1 Introduction

Image annotation (also known as caption retrieval) is the task of automatically associating an input image with a describing text. Image annotation methods are an emerging technology, enabling semantic image indexing and search applications. The complementary task of associating an input text with a fitting image (known as image retrieval or image search) is also of relevance for the same sort of applications.

State-of-the-art image annotation methods are currently based on deep neural net representations, where an image embedding (*e.g.*, obtained from a convolutional neural network or CNN) and a text embedding (*e.g.*, obtained from a recurrent neural network or RNN) are combined into a unique multimodal embedding space. While several techniques for merging both spaces have been proposed [1, 2, 3, 4, 5, 6], little effort has been made in finding the most appropriate image embeddings to be used in that process. In fact, most approaches simply use a one-layer CNN embedding [7, 8]. In this paper we explore the impact of using a Full-Network embedding (FNE) [9] to generate the required image embedding, replacing the one-layer embedding. We do so by integrating the FNE into the multimodal embedding pipeline

defined by Kiros et al. [1], which is based in the use of a Gated Recurrent Units neural network (GRU) [10] for text encoding and CNN for image encoding. Unlike one-layer embeddings, the FNE represents features of varying specificity in the context of the visual dataset, while also discretizes the features to regularize the space and alleviate the curse of dimensionality. These particularities result in a richer visual embedding space, which may be more reliably mapped to a common visual-textual embedding space.

The generic pipeline defined by Kiros et al. [1] has been recently outperformed in image annotation and image search tasks by methods specifically targeting one of those tasks [4, 11]. We choose to test our contribution on this pipeline for its overall competitive performance, expecting that any conclusion may generalize when applied to other solutions and tasks (*e.g.*, caption generation). This assumption would be dimmer if a more problem-specific methodology was chosen instead.

Our main goal is to establish the competitiveness of the FNE as an image representation to be used in caption related tasks. We test the suitability of this approach by evaluating its performance on both image annotation and image retrieval using three publicly available datasets: Flickr8k [12], Flickr30k [13] and MSCOCO [14]. Results obtained by the pipeline including the FNE are compared with the original pipeline of Kiros et al. [1] using a one-layer embedding, and also with the methods currently obtaining state-of-the-art results on the three datasets.

2   Related work

In the last few years, several solutions have been proposed to the problem of building common representations for images and text with the goal of enabling cross-domain search [1, 2, 3, 4, 5]. This paper builds upon the methodology described by Kiros et al. [1], which is in turn based on previous works in the area of Neural Machine Translation [15]. In their work, Kiros et al. [1] define a vectorized representation of an input text by using GRU RNNs. In this setting, each word in the text is codified into a vector using a word dictionary, vectors which are then fed one by one into the GRUs. Once the last word vector has been processed, the activations of the GRUs at the last time step conveys the representation of the whole input text in the multimodal embedding space. In parallel, images are processed through a Convolutional Neural Network (CNN) pre-trained on ImageNet [16], extracting the activations of the last fully connected layer to be used as a representation of the images. To solve the dimensionality matching between both representations (the output of the GRUs and the last fully-connected of the CNN) an affine transformation is applied on the image representation.

Similarly to the approach of Kiros et al. [1], most image annotation and image retrieval approaches rely on the use of CNN features for image representation. The current best overall performing model (considering both image annotation and image retrieval tasks) is the Fisher Vector (FV) [4], although its performance is most competitive on the image retrieval task. FV are computed with respect to the parameters of a Gaussian Mixture Model (GMM) and an Hybrid Gaussian-Laplacian Mixture Model (HGLMM). For both images and text, FV are build using deep neural network features; a VGG [17] CNN for images features, and a word2vec [18] for text features. For the specific problem of image annotation, the current state-of-art is obtained with the Word2VisualVec (W2VV) model [11]. This approach uses as a multimodal embedding space the same visual space where images are represented, involving a deeper text processing. Finally for the largest dataset we consider (MSCOCO), the best results in certain metrics are obtained by MatchCNN (m-CNN) [5], which is based on the use of CNNs to encode both image and text.

3   Methods

The multimodal embedding generator pipeline of Kiros et al. [1] represents images and textual captions in the same space. It is composed by two main elements, one which generates image embeddings and another one which generates text embeddings. We replace the original image embedding generator by
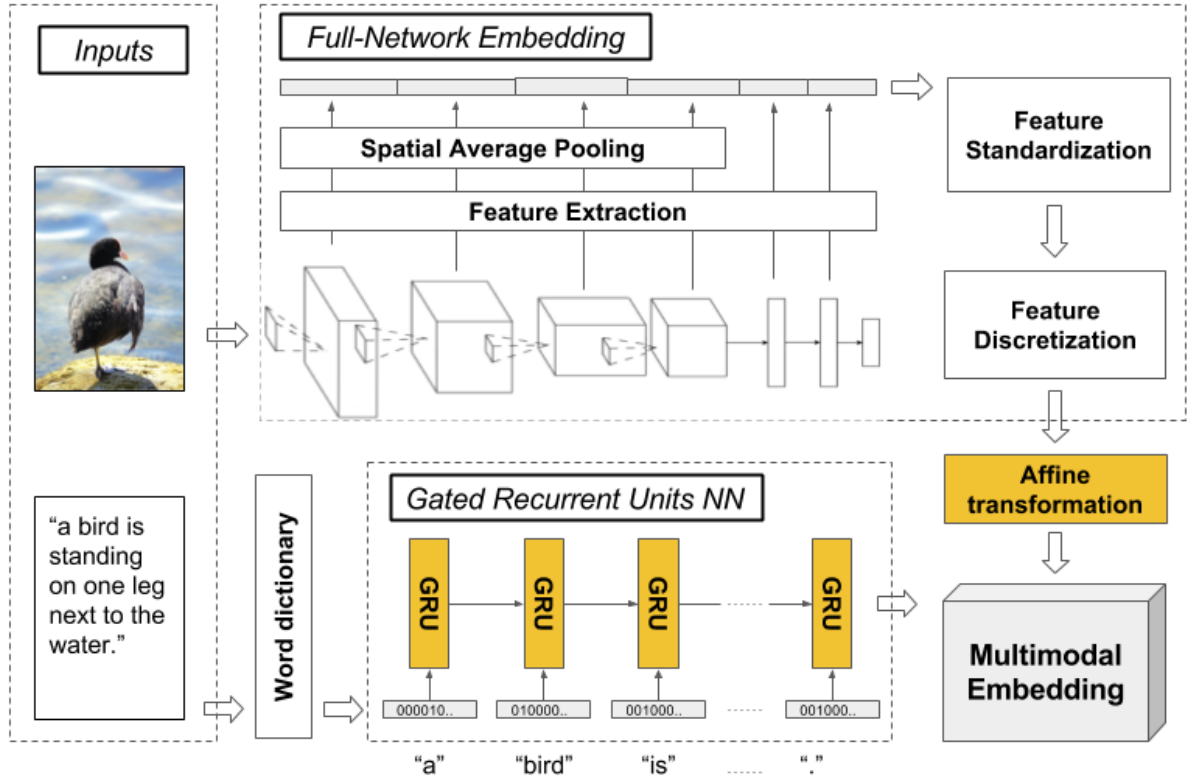
Figure 1: Overview of the proposed multimodal embedding generation pipeline with the integrated full-network embedding. Elements colored in orange are components modified during the neural network training phase. During testing, only one of the inputs is provided.

the FNE, which results in the architecture shown in Figure 1. Next we describe these components in further detail.

3.1 Full-network Embedding

The FNE generates a representation of an input image by processing it through a pre-trained CNN, and extracting the neural activations of all convolutional and fully-connected layers. After the initial feature extraction process, the FNE performs a dimensionality reduction step for convolutional activations, by applying a spatial average pooling on each convolutional filter. After the spatial pooling, every feature (from both convolutional and fully-connected layers) is standardized through the z-values, which are computed over the whole image train set. This standardization process puts the value of the each feature in the context of the dataset. At this point, the meaning of a single feature value is the degree with which the feature value is atypically high (if positive) or atypically low (if negative) in the context of the dataset. Zero marks the typical behavior.

The last step in the FNE pipeline is a feature discretitzation process. The previously standardized embedding is usually of large dimensionality (*e.g.*, 12,416 features for VGG16) which entails problems related with the curse of dimensionality. The usual approach to address this issue is to apply some dimensionality reduction methods (*e.g.*, PCA) [19, 20]. FNE uses a different approach, reducing expressiveness through the discretization of features, while keeping the dimensionality. Specifically, the FNE discretization maps the feature values to the $\{-1, 0, 1\}$ domain, where -1 indicates an unusually low value (*i.e.*, the feature is significant by its absence for an image in the context of the dataset), 0 indicates that the feature has an average value (*i.e.*, the feature is not significant) and 1 indicates an uncommonly high activation (*i.e.*, the feature is significant by its presence for an image in the context of the dataset). The mapping of standardized values into these three categories is done through the definition of two constant thresholds.

The optimal values of these thresholds can be found empirically for a labeled dataset [21]. Instead, we use threshold values shown to perform consistently across several domains [9].

3.2 Multimodal embedding

In our approach, we integrate the FNE with the multimodal embedding pipeline of Kiros et al. [1]. To do so we use the FNE to obtain an image representation instead of the output of the last layer of a CNN, as the original model does. The encoder architecture processing the text is used as it is, using a GRUs recurrent neural network to encode the sentences. To combine both embeddings, Kiros et al. [1] use an affine transformation on the image representation (in our case, the FNE) identical to a fully connected neural network layer. This extra layer is trained simultaneously with the GRUs. The elements of the multimodal pipeline that are tuned during the training phase of the model are shown in orange in Figure 1.

In simple terms, the training procedure consist on the optimization of the pairwise ranking loss between the correct image-caption pair and a random pair. Assuming that a correct pair of elements should be closer in the multimodal space than a random pair. The loss $\mathcal{L}$ can be formally defined as follows:

$$\mathcal{L} = \sum_{\mathbf{I}} \sum_{k} max\left\{0, \alpha - s(\mathbf{i}, \mathbf{c}) + s(\mathbf{i}, \mathbf{c}_k)\right\} + \sum_{\mathbf{C}} \sum_{k} max\left\{0, \alpha - s(\mathbf{i}, \mathbf{c}) + s(\mathbf{c}, \mathbf{i}_k)\right\} \tag{1}$$

Where $\mathbf{i}$ is an image vector, $\mathbf{c}$ is its correct caption vector, and $\mathbf{i}_k$ and $\mathbf{c}_k$ are sets of random images and captions respectively. The operator $s(\bullet, \bullet)$ defines the cosine similarity. This formulation includes a margin term α to avoid pulling the image and caption closer once their distance is smaller than the margin. This makes the optimization focus on distant pairs instead of improving the ones that are already close.

4 Experiments

In this section we evaluate the impact of using the FNE in a multimodal pipeline (FN-MME) for both image annotation and image retrieval tasks. To properly measure the relevance of the FNE, we compare the results of the FN-MME with those of the original multimodal pipeline reported by Kiros et al. [1] (CNN-MME). Additionally, we define a second baseline by using the original multimodal pipeline with a training configuration closer to the one used for the FNE experiments (*i.e.*, same source CNN, same MME dimensionality, *etc.*). We refer to this second baseline as CNN-MME*.

4.1 Datasets

In our experiments we use three different publicly available datasets:

The **Flickr8K** dataset [12] contains 8,000 hand-selected images from Flickr, depicting actions and events. Five correct captions are provided for each image. Following the provided splits, 6,000 images are used for train, 1,000 are used in validation and 1,000 more are kept for testing.

The **Flickr30K** dataset [13] is an extension of Flickr8K. It contains 31,783 photographs of everyday activities, events and scenes. Five correct captions are provided for each image. In our experiments 29,000 images are used for training, 1,014 conform the validation set and 1,000 are kept for test. These splits are the same ones used by Kiros et al. [1] and by Karpathy and Fei-Fei [22].

The **MSCOCO** dataset [14] includes images of everyday scenes containing common objects in their natural context. For captioning, 82,783 images and 413,915 captions are available for training, while 40,504 images and 202,520 captions are available for validation. Captions from the test set are not publicly available. Previous contributions consider using a subset of the validation set for validation and a different subset for test. In most cases, such subsets are composed by either 1,000 or 5,000 images for each set, with their corresponding 5 captions per image. In our experiments we consider both settings.

Table 1: Results obtained for the Flickr8 dataset. R@K is Recall@K (high is good). Med r is Median rank (low is good). Best results are shown in **bold**.

Model		Image Annotation				Image Retrieval			
		R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
FV	[4]	21.2	50.0	64.8	5	**31.0**	**59.3**	**73.7**	**4**
m-CNN	[5]	24.8	53.7	67.1	5	20.3	47.6	61.7	5
Bi-LSTM	[24]	29.3	58.2	69.6	**3**	19.7	47.0	60.6	5
W2VV	[11]	**33.6**	**62.0**	**75.3**	**3**	-	-	-	-
CNN-MME	[1]	18.0	40.9	55.0	8	12.5	37.0	51.5	10
CNN-MME*		21.0	45.7	60.4	7	14.0	35.8	48.6	11
FN-MME		23.3	50.8	66.8	5	15.0	38.2	51.6	10

4.2 Implementation and Evaluation Details

The caption sentences are word-tokenized using the Natural Language Toolkit (NLTK) for Python [23]. The choice of the word embedding size and the number of GRUs has been analyzed to obtain a range of suitable parameters to test in the validation set. The total number of different words is 8,919 for Flickr8k, 22,962 for Flickr30k and 32,775 for MSCOCO. Using all the words present in the dataset is likely to produce overfitting problems when training on examples containing words that only occur a few times. This overfitting problem may not have a huge impact on performance, but it may add undesired noise in the multimodal representation. The original setup [1] limited the word embedding to the 300 most frequent words, while using 300 GRUs. The Bi-LSTM model [24] in contrast defines the vocabulary size to include words appearing more than 5 time in the dataset, leading to dictionaries of size 2,018 for Flickr8k, 7,400 for Flickr30k and 8,801 for MSCOCO. Our own preliminary experiments on the validation set showed that increasing multimodal space dimensionality and dictionary length slightly improved the performance of image retrieval, in detriment of image annotation. However, the combined performance difference remains rather small when using non-extreme parameter values (*e.g.*, a model with 10,000 words vocabulary on MSCOCO dataset show a 0.4% average recall reduction when compared with a 2,000 words model). Since we are building a model for solving both tasks, we kept the parameters obtaining the highest combined score in the validation set. For the Flickr datasets, the word embedding is limited to the 1,000 most frequent words. For the MSCOCO dataset, we use a larger dictionary, considering the 2,000 most frequent words. In both cases we use 2,048 GRUs, which is also the dimensionality of the resultant multimodal embedding space.

For generating the image embedding we use the classical VGG16 CNN architecture [17] as source model pretrained for ImageNet [16]. This architecture is composed of 16 convolutional layers combined with pooling layers followed by two fully connected layers and the final softmax output layer. When using the FNE, this results in a image embedding space of 12,416 dimensions.

On all our experiments (both the CNN-MME* and the FN-MME) the margin parameter α is set to 0.2, and the batch size to 128 image-caption pairs. Within the same batch, every possible alternative image-caption pair is used as contrasting example. The models are trained up to 25 epochs, and the best performing model on the validation set is chosen (*i.e.*, early stopping). We use gradient clipping for the GRUs with a threshold of 2. We use ADAM [25] as optimization algorithm, with a learning rate of 0.0002 for the Flickr datasets, and 0.00025 for MSCOCO.

To evaluate both image annotation and image retrieval we use the following metrics:

- **Recall@K** (R@K) is the fraction of images for which a correct caption is ranked within the top-K retrieved results (and vice-versa for sentences). Results are provided for R@1, R@5 and R@10.

- **Median rank** (Med r) of the highest ranked ground truth result.

Table 2: Results obtained for the Flickr30 dataset. R@K is Recall@K (high is good). Med r is Median rank (low is good). Best results are shown in **bold**.

Model		Image Annotation				Image Retrieval			
		R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
FV	[4]	25.0	52.7	66.0	5	**35.0**	**62.0**	**73.8**	**3**
m-CNN	[5]	33.6	64.1	74.9	3	26.2	56.3	69.6	4
Bi-LSTM	[24]	28.1	53.1	64.2	4	19.6	43.8	55.8	7
W2VV	[11]	**39.7**	**67.0**	**76.7**	**2**	-	-	-	-
CNN-MME	[1]	23.0	50.7	62.9	5	16.8	42.0	56.5	8
CNN-MME*		30.4	58.0	69.5	4	18.9	44.6	57.0	7
FN-MME		30.4	61.8	73.2	3	22.1	47.6	59.8	6

Table 3: Results obtained for the MSCOCO dataset. Top part shows results when using 1,000 images as test set, while the bottom shows results when using 5,000 images for test. R@K is Recall@K (high is good). Med r is Median rank (low is good). Best results are shown in **bold**. † results not in original paper.

	Model		Image Annotation				Image Retrieval			
			R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
test 1K	FV	[4]	25.1	59.8	76.6	4	**39.4**	67.9	80.9	**2**
	m-CNN	[5]	42.8	73.1	84.1	**2**	32.6	**68.6**	**82.8**	3
	CNN-MME †		43.4	75.7	**85.8**	**2**	31.0	66.7	79.9	3
	CNN-MME*		41.2	72.8	85.1	2	26.2	58.6	73.9	4
	FN-MME		**47.3**	**76.8**	**85.8**	2	31.4	65.4	78.7	3
test 5K	FV	[4]	10.8	28.3	40.1	17	**17.3**	**39.0**	**50.2**	**10**
	Bi-LSTM	[24]	16.6	39.4	52.4	9	11.6	30.9	43.4	13
	CNN-MME*		18.7	42.8	56.7	8	10.4	28.7	40.6	17
	FN-MME		**21.1**	**46.6**	**60.2**	**6**	13.4	34.6	47.4	12

4.3 Results

For both image annotation and image retrieval tasks on the Flickr8k dataset, Table 1 shows the results of the proposed FN-MME, the reported results of the original model CNN-MME, the results of the original model when using our configuration CNN-MME*, and the current state-of-the-art (SotA). Tables 2 and 3 are analogous for the Flickr30k and MSCOCO datasets. Additional results of the CNN-MME model were made publicly available later on by the original authors [26]. We include these for the MSCOCO dataset, which was not evaluated in the original paper [1].

First, let us consider the impact of using the FNE. On all cases, the multimodal pipeline proposed by Kiros et al. [1] obtains equal or better results when using the FNE. This is the case for the originally reported results (CNN-MME), for the results made available later on by the original authors (CNN-MME†), and for the experiments we do using same configuration as the FN-MME (CNN-MME*). The comparison we consider to be the most relevant is the FN-MME against the CNN-MME*, as these contain the least differences besides the image embedding being used. In this particular case, the FN-MME outperforms the CNN-MME* by 3 percentual points on average for the Flickr datasets, and roughly by 4 points for the MSCOCO dataset.

To measure the relevance of the improvement provided by using the FNE, we compare the FN-MME model with the current state-of-the-art for image annotation and image retrieval. For the Flickr datasets,

particularly for image annotation tasks, the performance of the FN-MME is significantly closer to the state of the art than the other variants of the same model (CNN-MME, CNN-MME†, CNN-MME*). Remarkably, the FN-MME provides the best reported results on image annotation for the MSCOCO dataset. However, let us remark that the competitive W2VV method [11] has no reported results for MSCOCO. The results of the FN-MME for image retrieval tasks are significantly further from the state-of-the-art. Overall, the competitiveness of FN-MME increases with dataset size.

5   Conclusions

For the multimodal pipeline of Kiros et al. [1], using the Full-Network image embedding results in consistently higher performances than using a one-layer image embedding. These results suggest that the visual representation provided by the FNE is superior to the current standard for the construction of most multimodal embeddings.

When compared to the current state-of-the-art, the results obtained by the FN-MME are significantly less competitive than problem-specific methods. Since this happens for all models using the same pipeline (CNN-MME, CNN-MME†, CNN-MME*), these results indicate that the original architecture of Kiros et al. [1] is itself outperformed in general by more problem-specific techniques.

Since the FNE is compatible with most multimodal pipelines based on CNN embeddings, as future work of this paper we intend to evaluate the performance of the FNE when integrated into the current state-of-the-art on image annotation (W2VV [11]) and image retrieval (FV [4]). If the boost in performance obtained by the FNE on the Kiros et al. [1] pipeline translates to these other methods, such combination would be likely to define new state-of-the-art results on both tasks.

Acknowledgements

References

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

[2] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. Multimodal neural language models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 595–603, 2014.

[3] Andrej Karpathy, Armand Joulin, and Fei Fei F Li. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897, 2014.

[4] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior 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.

[5] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2623–2631, 2015.

[6] Qing Sun, Stefan Lee, and Dhruv Batra. Bidirectional beam search: Forward-backward inference in neural sequence models for fill-in-the-blank image captioning. *arXiv preprint arXiv:1705.08759*, 2017.

[7] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *Icml*, volume 32, pages 647–655, 2014.

[8] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813, 2014.

[9] Dario Garcia-Gasulla, Armand Vilalta, Ferran Parés, Jonatan Moreno, Eduard Ayguadé, Jesus Labarta, Ulises Cortés, and Toyotaro Suzumura. An out-of-the-box full-network embedding for convolutional neural networks. *arXiv preprint arXiv:1705.07706*, 2017.

[10] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[11] Jianfeng Dong, Xirong Li, and Cees GM Snoek. Word2visualvec: Image and video to sentence matching by visual feature prediction. *CoRR, abs/1604.06838*, 2016.

[12] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147. Association for Computational Linguistics, 2010.

[13] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C Lawrence Zitnick, and Piotr Dollar. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014.

[15] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[16] 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. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[19] Arsalan Mousavian and Jana Kosecka. Deep convolutional features for image based retrieval and scene categorization. *arXiv preprint arXiv:1509.06033*, 2015.

[20] Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. Factors of transferability for a generic convnet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9):1790–1802, 2016.

[21] Dario Garcia-Gasulla, Ferran Parés, Armand Vilalta, Jonatan Moreno, Eduard Ayguadé, Jesús Labarta, Ulises Cortés, and Toyotaro Suzumura. On the behavior of convolutional nets for feature extraction. *arXiv preprint arXiv:1703.01127*, 2017.

[22] Andrej Karpathy and Li 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.

[23] Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.

[24] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. Image captioning with deep bidirectional lstms. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 988–997. ACM, 2016.

[25] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[26] Ryan Kiros. visual-semantic-embedding. URL `https://github.com/ryankiros/visual-semantic-embedding`.