Multimodal Image News Article Alignment

First Author[1], Second Author[2,3], and Third Author[3]

[1] Princeton University, Princeton NJ 08544, USA
[2] Springer Heidelberg, Tiergartenstr. 17, 69121 Heidelberg, Germany lncs@springer.com
http://www.springer.com/gp/computer-science/lncs
[3] ABC Institute, Rupert-Karls-University Heidelberg, Heidelberg, Germany
{abc,lncs}@uni-heidelberg.de

Abstract. Multimodal learning has received a lot of attention in the recent years. Associating a description to an image in any language is a challenging task as it involves identifying the objects within the image and determining the relationships between them. Often, the documents are multimodal, and hence they may contain text as well as images. Various methodologies have been put forward to match an image to its corresponding description at sentence level. In this work, we are the first to propose a novel joint image-paragraph (i.e. news article) ranking model trained with images and its corresponding paragraphs (i.e. news articles). The image-paragraph ranking model works in such a way that, given an image, the model ranks the best matching news articles and vice-versa. We achieve this correspondence by using a pairwise ranking function and evaluate the model performance on benchmark datasets using Image-Sentence Ranking task evaluation metric. The experimental results show that our model achieves comparable performance to the cutting edge technique.

Keywords: First keyword · Second keyword · Another keyword.

1 Introduction

Generating Image Description is traditionally often considered a demanding undertaking given its complexities in combining perception, learning, and understanding of the language to produce a fitting description of the image and its vice-versa. Aside from recognizing the contents of the image, it is also imperative for the model to understand in what way the contents of the image under study interact with one another. Advancements in Deep Learning techniques and Neural Networks have made object recognition and image translation considerably feasible to achieve. In this paper, we present a model that is trained to align images with sentences and paragraphs (news articles) and its vice-versa. The images are encoded using a deep convolutional neural network (CNN), and the sentences are encoded using long-short-term memory (LSTM) recurrent neural network (RNN) [9]. Using these features, we learn a joint embedding model where we project the image features from the pre-trained VGG-19 [20] CNN into the embedding space of the LSTM. The main objective of this study is to build a multimodal sentence-paragraph ranking model to perform the image-paragraph (news article) alignment. To implement such correspondence, we extended the idea proposed in [13] of visual-semantic embeddings and train our model jointly with images and

paragraphs. In particular, we implement a ranking function that performs the pairwise ranking of images and text that ranks the descriptions for a given input image and vice-versa. By learning to rank, this model can then be used to retrieve the text that are more closely match a new image, or vice-versa. Once the ranks are computed, the number of correct and incorrect predictions can be measured. For training the image-sentence ranking model, we use the Flickr30K [24] and MS COCO [15] datasets. The above two datasets have a large set of images paired with their corresponding descriptions in the English language. The joint alignment task for which a paragraph is required is carried out by Euronews [2] dataset. The news articles come in various languages of which only English is used to train the image-paragraph ranking model.

To the best of our knowledge, we are the first to propose image-paragraph (news article) alignment task. We tackled the problem in two steps. First, we build the image-sentence ranking model using the method proposed in [13] to align images with their corresponding sentences as our baseline. Secondly, the model is then extended and tested with the Euronews dataset [2] to align images to their corresponding news articles (paragraphs). Our ranking model is based on deep neural networks. In the absence of widely accepted evaluation measures, we propose to evaluate our model in the context of image-sentence ranking (ISR) task [5].

The remainder of the paper is organized as follows. In Section 2, the background and related research concerning multimodal learning are discussed. The approach employed for training the model is explained in Section 3; in Section 4, the datasets used for learning the ranking model is explained followed by the experimental setup and results in Section 5. The performance of the models are discussed in Section 6, and finally, in Section 7, the conclusions are presented.

2 Related Work

Multimodal approaches have gained increased attention over the past few years. Multimodal models have been used in various applications: sentiment analysis [17]; visual question answering [3]; multimodal machine translation [22]; combating human trafficking [23]; and detection of cyber-bullying [25]. Below we summarize few studies which are relevant to the current work.

In [13], the authors have used an encoder-decoder model which learns the multimodal joint embedding with text and images along with a language model that decodes the distributed representations from embedding space. The encoder ranks the images and the captions/sentences. The sentences are encoded using LSTM [9] recurrent neural networks whereas the images are projected onto the embedding space of the LSTM from a deep convolution neural network. On the other hand, the decoder is used to produce a new caption for an image from scratch. They trained and performed tests on the model using the Flickr8K [10] and Flickr30K [24] datasets. To evaluate the performance of the LSTM sentence encoder for ranking images and descriptions, a pairwise ranking loss function was introduced which was minimized to learn to rank the images with captions. By comparing the results with other relevant models, they observed that their model outperformed the others in some metrics as well as they were also outperformed by other models (multi-modal RNN) [16] in certain metrics. While their paper

focusses on generating the description of an image and the joint alignment of sentences and images, they don't consider paragraphs for the ranking task.

In a similar manner, [21] introduced the dependency tree RNN model (DT-RNN). The RNN model uses dependency trees to encode sentences into the vector space. They used a pre-trained CNN to extract the image features. Using the image features from the pre-trained CNN and vectors from the DT-RNN, they implement a pairwise ranking function to learn their multimodal embeddings. In [11], authors focus on the multimodal embedding of visual and image features. They trained their model to rank images and sentences such that given an image query, then can rank a fixed set of sentences and vice-versa. The authors trained their model on the Flickr8K, Flickr30K and Pascal1K [18] datasets. Their model bought about the latent, inter-modal correspondence between images and text by breaking down the fragments of images and sentences in a joint common embedding space. In another study [5] which is very similar to our work a ranking model that learns embeddings from multimodal and multilingual data has been proposed. Their model takes the images and descriptions of multiple languages into consideration. Like in the other models discussed above, they also introduce a pairwise ranking function that is adapted to rank the images and sentences having more than two or three input sources images and their corresponding multilingual sentences. Their objective function maps the images and multiple sentences in some languages for which they make use of the Multi30k [8] dataset. To evaluate the embeddings, they compare their model with three baselines namely image-sentence ranking (ISR), semantic textual similarity (STS), and neural machine translation (NMT). In ISR, they observed that their multilingual model showed consistent improvements in ranking images with given sentences, but it was not equally effective while ranking the sentences given images. For the STS baseline, they use their model to calculate the distance between a pair of sentences. In the NMT method, they use the multilingual multi-modal embedding (MLMME) model to re-rank n-best lists generated with different baselines.

We propose our model based on the previous works done in this directions and extend the idea of training multi-modal image and sentence level joint embeddings to the scenario where the training data is not only for sentences but also paragraphs (i.e. news articles).

3 Methodology

The overall methodology is depicted in Fig. 1. It consists of two main components: Image-text encoder and pair-wise ranking function. The details of the each components are presented below.

3.1 Image and paragraph encoding

The first step towards training the model is to project the multimodal data to a joint embedding space. The joint embedding space is learned using visual (images) and textual (sentences or paragraphs) modalities from the training data. For our visual modality, we make use of a pre-trained model for extracting the image features. It consists of a 16-layer and a 19-layer network of which we use the 19-layer VGG model (VGG19)

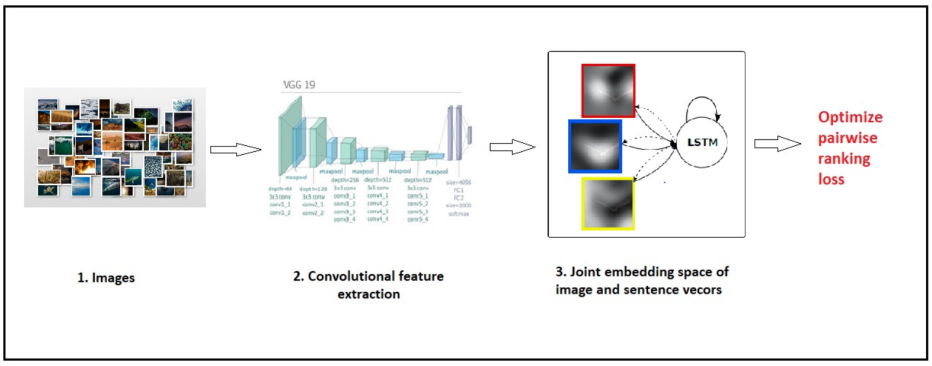

Fig. 1: Overall methodology.

[20]. The VGG-19 network is trained on more than a million images and can classify images into 1000 object categories. Hence, the model has learned plentiful image representations over an exhaustive range of images. The VGG-19 model takes the input of the fixed size 224x224 RGB image. The input image then passes through a stack of convolutional layers where each filter is of the size 3x3 and each max pooling layer is of the filter size 2x2. Three fully connected layers follow the pile of convolutional layers, first two with 4096 channels and the final one with 1000 channels. Softmax is the activation function for the final layer, and all the hidden layers are equipped with the rectification (ReLU) [14] non-linearity. Using this network architecture, we extract the image features for each of the three datasets. To be precise, we use the features from the penultimate layer of the VGG-19 network which outputs a 4096-dimension feature vector for our image-sentence ranking baseline model, and features from the softmax layer for the image-paragraph ranking model.

In the textual modality comprising sentences and paragraphs (news articles), we use a recurrent neural network to encode the text. An encoder-decoder framework [6] that consists of two RNNs which acts as an encoder and decoder pair has been proposed to improve the performance of a machine translation system. We train the two networks jointly in such a way that, given a source sequence, the model is trained to learn the conditional probability of the target sequence. We use the encoder part which maps a sequence of sentences to a fixed length vector. Specifically, we use a RNN with gated recurrent units that read through the text word by word and generates a sequence of vectors representing the text for each embedding. Hence, the sentence representation of a given input is obtained from its corresponding encoder RNN's vector sequence. As the final step, the textual embeddings and the image embeddings are normalized to share the same dimensionality between them.

3.2 Image-paragraph ranking

To implement the image-paragraph ranking model, the pairwise ranking loss needs to be calculated i.e., the ranking scores for the images and text has to be computed. Our model not only computes the ranking score for images and sentences, but it also performs

the pairwise ranking of images and paragraphs (news articles). The image-document multimodal pairwise ranking loss function is given by:

$$\min_{\theta} \sum_{x} \sum_{k} \max\{0, \alpha - s(x, v) + s(s, v_k)\} + \sum_{v}^{x} \sum_{k}^{k} \max\{0, \alpha - s(v, x) + s(v, x_k)\}$$

In the above function, "x" stands for image embedding and "v" stands for text embedding. "s" is the scoring function which is given as $s(x, v)$ where x and v are scaled using cosine similarity to make a unit norm. "α" denotes a model parameter i.e., the margin and "θ" denotes all the parameters to be learned. "V_k" is a non-descriptive text for the image embedding x, and "x_k" is a non-descriptive image for the text embedding v and $k \in K$ where "K" is the dimensionality of the embedding space. In simpler terms, here, the first portion of the function is for ranking images and the second half of the function is for ranking the sentences. We train the model with the image and text features with the goal of optimizing the pairwise ranking function.

4 Datasets

To train the image-sentence ranking model, we make use of the Flickr30K and the MS COCO datasets. The Flickr30K dataset is an extension of the Flickr8K [10] dataset which consists of 31,783 photos and 158,915 captions capturing everyday activities and events from the flickr.com website. Each image has 5 different descriptions in English described independently by 5 individual annotators. The images in the dataset don't contain celebrities and famous locations, instead, it covers a diverse set of scenes and situations. The train, validation and test split sets contain 29k, 1k, and 1k images respectively where each image corresponds to 5 captions describing the image.

The Microsoft Common Objects in Context (MS COCO) is an extensive dataset which contains a total of 328k images and 2.5 million captions of 91 object types that could be recognized by a 4-year-old. Just like the Flickr30K dataset, the MS COCO dataset also consists of 5 descriptions for each image that was gathered from crowd workers. The dataset particularly focusses on capturing non-iconic images capturing objects in a natural context. For training our image-sentence ranking model, we make use of the first half of the dataset splits that was released in 2014, in which the train set comprises of 82,783 images with 5 captions per image making a total of 413,915 captions in English. Thereby, the train, validation and test split sets contain 82,783, 1k and 1k images respectively with their corresponding descriptions.

For training the image-paragraph ranking model, we make use of data that has images and their corresponding news articles which were collected from the news website Euronews. The Euronews website comes with news articles in 9 languages: French, Arabic, English, German, Spanish, Italian, Turkish, Ukrainian and Portuguese in several genres like Sports, Politics, etc. The Euronews corpus created by [2] comprises image sources of political items along with their identical news articles in English and German. As the intended model focusses only with monolingual multimodal learning, the data is cleaned further to extract only the English text for its image source. The corpus is created with news article data from 2013 to 2015 of which we consider only the

Dataset	Flickr30k	MS COCO	Euronews
Training Set	29,000	82,783	13,096
Validation Set	1,000	1,000	1,000
Test Set	1,000	1,000	1,000
Total	31,000	84,783	15,096

Table 1: Training, validation and test splits of images in the datasets.

data from 2013 and some of 2014 as it involves extensive manual data cleaning. Each image corresponds to only one paragraph (news article). The train, validation and test split sets contain 13,096, 1k and 1k images respectively along with their corresponding news article. We present the details of the three datasets in Table 1.

5 Experimental Setup and Results

The experiments are conducted in two stages. First, we start with the image-sentence ranking model on the Flickr30K and MS COCO datasets and then followed by the image-paragraph ranking model using the Euronews dataset. The performance of the models is evaluated with the image-sentence ranking task [5].

5.1 Image-sentence ranking model

We start our experiment towards training an image-sentence ranking model using the Flickr30K dataset. As explained in Section 3 (methodology), the image features were extracted through the pre-trained VGG-19 deep network while an RNN encoder was employed to vectorize the sentences. Each image in the dataset was trained with 5 different captions. Apparently, in the model parameters, the word embeddings were initialized to 300-dimensional vectors. The maximum length of the sentence vectors was set to 100. The RNN encoder is used with GRU which is 1024 dimensional. The image embeddings are 4096 dimensional which we extract from the penultimate layer of the VGG-19 network. The initial learning rate for training the model is set to 0.0002. We use the very popular adam [12] optimizer, similar to the stochastic gradient descent, in which the model computes individual learning rates for different parameters in an adaptive fashion as the training unfolds. The model is trained in mini batches of 128 through the whole dataset and the number of epochs was set to 15. The same experimental setup is followed for the MS COCO dataset as well which is larger than the Flickr30K dataset comparatively. The above experiments were implemented using the Theano [4] library.

5.2 Image-paragraph ranking model

The image-paragraph ranking model is an extension of the image-sentence ranking model where the model is trained to align images with their corresponding news articles on the Euronews dataset. The model architecture is similar to the model trained with sentences and images with some modifications. The images features are extracted

Flickr30k	**Image to Text**		**Text to Image**	
Method	**Baseline**	**Our model**	**Baseline**	**Our model**
$Recall@K = 1$	23.0	28.7	16.8	21.9
$Recall@K = 5$	50.7	55.4	42.0	47.6
$Recall@K = 10$	62.9	68.6	56.5	59.6
$Medr.$	5.0	4.0	8.0	6.0

Table 2: Recall@K (K=1, 5, 10) and $Medr.$ on Flickr30k dataset

from the final softmax layer of the fully connected VGG-19 model unlike the image-sentence ranking model, and the weights are assigned based on the pre-trained Imagenet [19]. As a result, we increased the dimensions of the image embeddings for training the model. The word embeddings were changed to 600D vectors from 300D vectors as the size of the text is larger comparatively. At the same time, the maximum length of the sentence vectors was changed from 100 to 300 to match the bigger textual sequences of the news-articles text. Other model parameters were retained from the image-sentence ranking model, and the experiment was carried out for 15 epochs. The model was also modified to adapt to images with only one description per image. The open source Keras [7] neural network library that runs on top of Tensorflow [1] backend was used to extract features from the pre-trained VGG-19 model.

5.3 Evaluation Metrics

We evaluate our models using the Image-Sentence ranking (ISR) task [5] which is a form of an extrinsic evaluation where we compare the performance of our image-sentence ranking model performed on the Flickr30K and MS COCO datasets. The ISR task is evaluated for the metrics: Recall@K (high value is good) and Median Rank (low value is good). Recall@K measures the mean number of times an image is correctly mapped to its corresponding sentence within the top K-retrieved entries and vice-versa for the sentence to image task. We also evaluate our model with the Median rank (Medr.) metric that describes the closest truth result from the ranked list. The image-paragraph ranking model is also evaluated in the same manner.

5.4 Baseline [13]

To the best of our knowledge, no prior study has addressed the problem of image-paragraph (news article) ranking. A seminal work towards this direction is proposed by Kiros et al. [13]. They proposed a methodology based on encoder-decoder framework which enables to learn joint embedding space from text and images along with a language model that decodes the distributed representations from space. To evaluate the performance of their model, a pairwise ranking loss function was introduced which was minimized to learn to rank the images with captions. For comparative evaluation, we consider their model as baseline for the current study.

MS COCO	**Image to Text**		**Text to Image**	
Method	**Baseline**	**Our model**	**Baseline**	**Our model**
$Recall@K = 1$	43.4	43.6	31.0	31.3
$Recall@K = 5$	75.7	73.7	66.7	66.0
$Recall@K = 10$	85.8	84.9	79.9	79.5
$Medr.$	2.0	2.0	3.0	3.0

Table 3: Recall@K (K=1, 5, 10) and $Medr.$ on MS COCO dataset

5.5 Results

In Table 2 and Table 3, we present the results obtained by the image-sentence ranking model on Flickr30K and MS COCO datasets, respectively. In Table 4, we present the results obtained by our image-paragraph ranking model based on Recall@K and Medr. metrics with the Euronews dataset.

Euronews	**Image to Text**	**Text to Image**
Method	**Our model**	**Our model**
$Recall@K = 1$	0.5	1.0
$Recall@K = 5$	3.5	2.8
$Recall@K = 10$	5.5	5.7
$Medr.$	98	96

Table 4: Recall@K (K=1, 5, 10) and $Medr.$ on Euronews dataset

In Table 5, we present the results of the image-paragraph ranking model evaluated on the Recall@k metric at k = 15, 25, 50. The experiment is conducted to test the efficiency of the model at a higher "K" value as the model trained on the Euronews dataset.

6 Discussions

It can be seen from the results Section that our proposed model achieves comparable results for the image-sentence ranking model as baseline [13] on the Flickr30K and MS COCO datasets (Table 2 and Table 3). Each of the metric (Recall@1, Recall@5, Recall@10) produced good ranking scores and were comparable to the results produced by the baseline model. The performance of the model with MS COCO dataset is marginally better than that of the model with the Flickr30K dataset. The model is trained with 82k images 413k captions on the MS COCO dataset, while the Flickr30K dataset is trained with 29k images 145k captions, hence, from this behavior, it is evident that the model performs well with larger datasets.

It can be seen from the results presented in Table 2 that our model has performed better than Baseline model. One possible reason could probably be due to the learning rate which was .0002 as opposed to 0.2 in [13]. Another possible reason could be due

the selection of better optimization technique (adam optimizer over stochastic gradient descent).

Euronews	**Image to Text**	**Text to Image**
Method	**Our model**	**Our model**
$Recall@K = 15$	13.0	10.1
$Recall@K = 25$	17.0	14.2
$Recall@K = 50$	26.0	27.0

Table 5: Recall@K (K=15, 25, 50) on Euronews dataset

On the other side, the results produced by the image-paragraph ranking model is not on the higher side. One major downside could be the size of the training data as there were only 13k images. Other reasons for the low performance could be attributed to the number of news articles per image. An article or a document can be written in a number of ways using different words, or different forms of speech. Hence, it could be very rare to have two exactly similar article for the same image. Looking at one possible remedy to improve the model, it would be more relevant to extract keywords from the news articles (paragraphs) of a given image so that the model could be trained with those keywords and hence the model could map an image with a matching caption having the most number of keywords to produce better results (ranks). Having fixed length vectors for word embeddings could also be a bottleneck in improving the model's performance. As the Euronews dataset has images only related to politics, there is a potential chance of domain mismatch with the dataset used to train VGG-19 network which results in reduced performance. One potential solution for the domain mismatch problem would be to use transfer learning approach. Despite all these limitations, the novel area of research for ranking the images with paragraphs is wide open for further exploration.

7 Conclusions

In this study, we have proposed a novel ranking model that is trained to align images with its corresponding news articles and vice versa. As the first step, we implemented an image-sentence ranking model that was trained on the benchmark Flickr30k and MS COCO datasets. The results produced by the model was evaluated with Recall@K and Median Rank metrics. Our model attained the state-of-the-art performance. Further we extend the image-sentence ranking model to image-paragraph ranking model. The image-paragraph ranking model was trained on the Euronews dataset. The idea of ranking images with paragraphs could be used in a variety of real life applications. For example, in motor insurance claims, when an image of a damaged vehicle is given as input, the best matching descriptions of the image could be retrieved to make the business easier and faster. In future, we would extend the proposed monolingual multimodal ranking model to a multilingual multimodal image-paragraph ranking model that supports multiple languages. All the resources developed in this current study including dataset used and codes are available at `www.anonymous.com`.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: a system for large-scale machine learning. In: OSDI. vol. 16, pp. 265–283 (2016)
2. Afli, H., Lohar, P., Way, A.: Multinews: A web collection of an aligned multimodal and multilingual corpus. In: Proceedings of the First Workshop on Curation and Applications of Parallel and Comparable Corpora. pp. 11–15 (2017)
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2425–2433 (2015)
4. Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., Bengio, Y.: Theano: A cpu and gpu math compiler in python. In: Proc. 9th Python in Science Conf. vol. 1 (2010)
5. Calixto, I., Liu, Q.: Sentence-level multilingual multi-modal embedding for natural language processing. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017. pp. 139–148 (2017)
6. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
7. Chollet, F., et al.: Keras: The python deep learning library. Astrophysics Source Code Library (2018)
8. Elliott, D., Frank, S., Sima'an, K., Specia, L.: Multi30k: Multilingual english-german image descriptions. arXiv preprint arXiv:1605.00459 (2016)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
10. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research **47**, 853–899 (2013)
11. Karpathy, A., Joulin, A., Fei-Fei, L.F.: Deep fragment embeddings for bidirectional image sentence mapping. In: Advances in neural information processing systems. pp. 1889–1897 (2014)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. corr abs/1411.2539 (2014) (2014)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
15. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
16. Mao, J., Xu, W., Yang, Y., Wang, J., Yuille, A.L.: Explain images with multimodal recurrent neural networks. arXiv preprint arXiv:1410.1090 (2014)
17. Poria, S., Chaturvedi, I., Cambria, E., Hussain, A.: Convolutional mkl based multimodal emotion recognition and sentiment analysis. In: Data Mining (ICDM), 2016 IEEE 16th International Conference on. pp. 439–448. IEEE (2016)
18. Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J.: 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. pp. 139–147. Association for Computational Linguistics (2010)

19. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
21. Socher, R., Karpathy, A., Le, Q.V., Manning, C.D., Ng, A.Y.: Grounded compositional semantics for finding and describing images with sentences. Transactions of the Association of Computational Linguistics **2**(1), 207–218 (2014)
22. Specia, L., Frank, S., Sima'an, K., Elliott, D.: A shared task on multimodal machine translation and crosslingual image description. In: Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. vol. 2, pp. 543–553 (2016)
23. Tong, E., Zadeh, A., Jones, C., Morency, L.P.: Combating human trafficking with multimodal deep models. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 1547–1556 (2017)
24. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: 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)
25. Zhong, H., Li, H., Squicciarini, A.C., Rajtmajer, S.M., Griffin, C., Miller, D.J., Caragea, C.: Content-driven detection of cyberbullying on the instagram social network. In: IJCAI. pp. 3952–3958 (2016)