Automatic Judgement of Neural Network-Generated Image Captions

Rajarshi Biswas[1(✉)], Aditya Mogadala[3], Michael Barz[1,2], Daniel Sonntag[1], and Dietrich Klakow[3]

[1] German Research Center for Artificial Intelligence (DFKI), Saarland Informatics Campus D3 2, 66123 Saarbrücken, Germany
{rajarshi.biswas,michael.barz,daniel.sonntag}@dfki.de
[2] Saarbrücken Graduate School of Computer Science, Saarland Informatics Campus D3 2, 66123 Saarbrücken, Germany
[3] Spoken Language Systems (LSV), Saarland Informatics Campus D3 2, 66123 Saarbrücken, Germany
{amogadala,dietrich.klakow}@lsv.uni-saarland.de
http://iml.dfki.de

Abstract. Manual evaluation of individual results of natural language generation tasks is one of the bottlenecks. It is very time consuming and expensive if it is, for example, crowdsourced. In this work, we address this problem for the specific task of automatic image captioning. We automatically generate human-like judgements on grammatical correctness, image relevance and diversity of the captions obtained from a neural image caption generator. For this purpose, we use pool-based active learning with uncertainty sampling and represent the captions using fixed size vectors from Google's Universal Sentence Encoder. In addition, we test common metrics, such as BLEU, ROUGE, METEOR, Levenshtein distance, and n-gram counts and report $F1$ score for the classifiers used under the active learning scheme for this task. To the best of our knowledge, our work is the first in this direction and promises to reduce time, cost, and human effort.

Keywords: Active learning · NLP · NLG · Automated human judgement · Image captioning · Neural networks

1 Introduction

Recently, automatic image caption generation has received a lot of attention in scientific natural language processing (NLP) and applications of natural language generation (NLG) in particular. It has attracted a lot of attention from the machine learning (ML) community as well—because of far reaching NLG-ML-applications ranging from assisting the visually impaired to the development of socially interactive robots [10,16,20,31].

Although significant progress has been made in dealing with the caption generation problem [3,18,24,30], we still need to perform manual human evaluation for assessing the quality of the generated descriptions. This is both expensive

C. Martín-Vide et al. (Eds.): SLSP 2019, LNAI 11816, pp. 261–272, 2019.
https://doi.org/10.1007/978-3-030-31372-2_22

and time consuming. In this work, we have modified [3] to remedy this situation by automating human judgement on the quality of the generated descriptions (see examples in Fig. 1) through an active learning scheme (Fig. 2). Specifically, we infer human judgement on grammatical correctness, image relevance and diversity of the generated captions in an automatic manner.

For this purpose, we employ standard ML classifiers, SVM and logistic regression, under a pool based active learning scheme [29]. First, we generate diverse captions for images in the MSCOCO dataset [22] using the neural architecture in [32] along with beam search [23]. A small number of these captions are randomly selected and binary labels on their grammatical correctness, image relevance and diversity are crowdsourced to train the mentioned classifiers for each task. Using the learned classifiers we predict grammatical correctness, image relevance and diversity labels for the unlabeled captions. Subsequently, a batch of 200 instances which lie close to the decision boundary of the classifiers are selected and annotated using the same crowdsourcing platform. We incorporate them in the training set and re-train the classifiers on the new training set. We repeat this cycle 4 times and report the $F1$ scores for the classifiers on a separate human labeled test set.

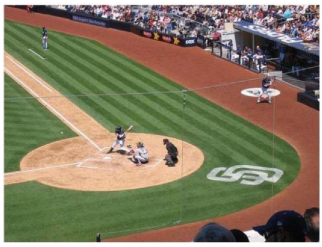
To summarize, our primary contributions are: first, a new approach in the direction of automatic human evaluation of machine generated image captions; and second, a computational model that uses a fixed size vector representation for sentences, obtained from a pre-trained network and standard metrics which produce a good baseline for automating human evaluation. The paper is organized as follows: Sect. 2 describes related work in the field of image caption generation. Section 3 describes in details the method used in our work for automatically inferring human judgement on the three quality aspects discussed above. Section 4 provides experiments and results followed by a short discussion in Sect. 5. Section 6 provides the conclusion.

2 Related Work

Although there has been considerable interest in language grounding in perceptual data [9,25,28], in the recent past there has been an explosion of interest in the problem of image captioning. As a matter of fact, this is part of a broader effort to investigate the boundary between vision and language. The caption generation method in our work uses the neural framework proposed in [6] where, instead of translating text from one language to another, an image is translated into a caption or sentence that describes it. The neural architecture for image caption generation consists of a deep convolutional network [13] and a recurrent neural network [12]. The first approach in this direction is credited to Kiros et al. [18,19] who proposed to construct a joint multimodal embedding space and provide a natural way to perform both ranking and generation. Works [7,30] offer slight contrast as the authors adopt LSTM RNNs instead of stock RNNs. Karpathy et al. [15] proposes to learn a joint embedding space for both ranking and generation. In fact, their model learns to score sentence and image similarity as a function of convnet object detections with outputs of a bidirectional RNN.

The caption generation problem also is a structured learning problem since both the input and output of this problem have a rich structure. That is, the

1. a wooden table topped with different types of food
2. there are many plates of food on the table
3. a variety of different types of food on a table



1. a baseball player taking a swing at a ball
2. a batter catcher an umpire during a baseball game
3. an image of a professional baseball game being played

1. a view of a living room and dining room
2. the living room is clean and ready for us to use
3. a living room filled with lots of furniture and a flat screen tv

1. a man holding a smart phone in his hand
2. a man taking a picture with his cellphone
3. a close up of a person holding a phone

Fig. 1. Diverse image captions generated using beam search

image of a natural scene is made up of multiple random variables, such as, position of objects, their inter-relationship and all of them have a rich joint distribution. Moreover, there needs to be an alignment between the output words of a caption with the spatial regions of the input image. So, to properly address the structured nature of this problem, we make use of attention mechanism in our work. Hence, we have adopted the show, attend and tell architecture by Xu et al. [32] which uses attention to generate the captions for images.

In addition to being an important task in the area of computer vision, image caption generation is also a major problem in the area of Natural Language Generation (NLG) where proper evaluation of such a system is a core issue. The methods for evaluation can be divided into intrinsic and extrinsic methods. Human Judgement falls under the category of intrinsic evaluation methods and one of the most important requirement for new applications such as [2,26]. The common criteria here include readability or fluency, which refer to the linguistic quality of the text, and also accuracy or relevance relative to the input which shows the NLG system's ability to satisfactorily reproduce content. However, none of the image captioning or NLG methods described above have tried to automatically generate human judgement on the quality of their generations and instead relied on conducting time and cost intensive human evaluation through public surveys. It is worth mentioning here that standard metrics, such as, BLEU [27], ROUGE [21] aim to emulate human judgement but often fall short as they suffer from low correlation between them and human judgements, a fact which is

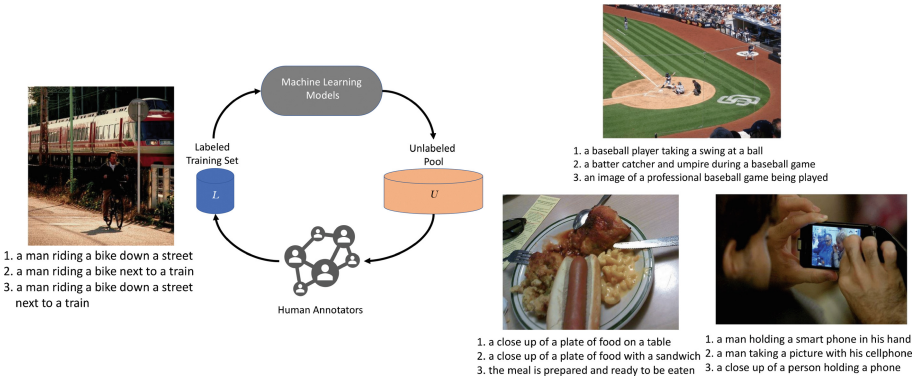


Fig. 2. Pool based active learning scheme

widely reported in the NLG community. In addition, these standard metrics are dependent on groundtruth information since they measure the overlap between a generation and its groundtruth for quality assessment. This in our view is a severe limitation and prevents true automatic evaluation of NLG tasks.

To the best of our knowledge, we believe our attempt which uses fixed size vectors from pretrained sentence encoders, is the first one in the direction of automated human judgement for quality assessment which does not require groundtruth information and thus reduces cost, boosts productivity.

3 Method

We aim at automating human judgements on neural network generated image captions using active learning. In the following, we describe the caption generator, the features that we consider for modeling human judgements and our active learning approach.

3.1 Image Caption Generation

For generating the image captions we use the Show, Attend and Tell [32] approach on the MSCOCO dataset [22] as depicted in Fig. 4. In this approach instead of using a single fixed dimensional vector to represent the image, a set of fixed dimensional vectors from a lower convolution layer of the CNN architecture is used. This helps to maintain a fine-grained correspondence between the different portions of a $2D$ image represented through the corresponding vectors. With this the decoder becomes more powerful as it can focus selectively on different parts of an image during the generation process by selecting a subset of the feature vectors.

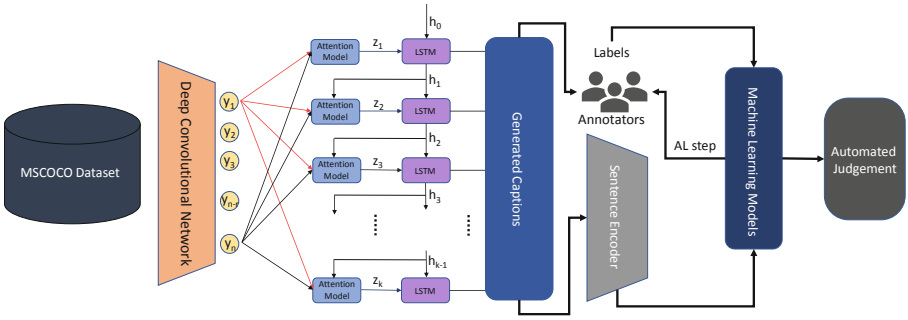


Fig. 3. Full schema for generating automated human judgement

The detailed operations of the LSTM based decoder, used in [32] for generating the captions, are described through the following equations,

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ g_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ tanh \end{bmatrix} T_{D+m+n,n} \begin{bmatrix} E_{y_{(t-1)}} \\ h_{t-1} \\ \hat{z}_t \end{bmatrix} \tag{1}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{2}$$

$$h_t = o_t \odot tanh(c_t) \tag{3}$$

where, i_t, f_t, c_t, o_t, h_t denote input, forget, memory, output gates and the hidden state respectively. It is to be noted that T represents a mapping of the form $f_{s,t} : \mathbb{R}^s \rightarrow \mathbb{R}^t$. Thus, $T_{D+m+n,n}$ is a mapping from $\mathbb{R}^{(D+m+n)}$ to $\mathbb{R}^n$. $\hat{z} \in \mathbb{R}^D$ denotes the context vector responsible for capturing the visual information related to a specific location in the input image. E denotes the embedding matrix and has the dimension $m \times k$. The dimension of the embedding vector is given by m while the dimension of the LSTM hidden state is denoted by n. Furthermore, σ and $\odot$ represent the logistic sigmoid and element-wise multiplication respectively.

For handling the MSCOCO data, we adopt the data splits proposed in [14] in which the training set contains 113,287 images with each having 5 corresponding captions while the validation and test sets contain 5,000 images with each having 5 corresponding groundtruth captions. For our work, we build a vocabulary by dropping a word which has a frequency below 5 leading to a vocabulary size of 10,000 words. We use image features obtained from the RESNET-101 architecture with 101 layers [11]. The dimensions for the LSTM hidden state, image, word and attention embeddings are set to 512 for our model. We train our model under the cross entropy objective, using beam search for the decoder and ADAM [17] as the preferred optimizer. We use beam search with a beam width of 200 from which we select the top three captions for each image. We use this setup to generate captions for all images in the MSCOCO test set and use them for evaluating our proposed approach for automatically inferring human judgement on them.

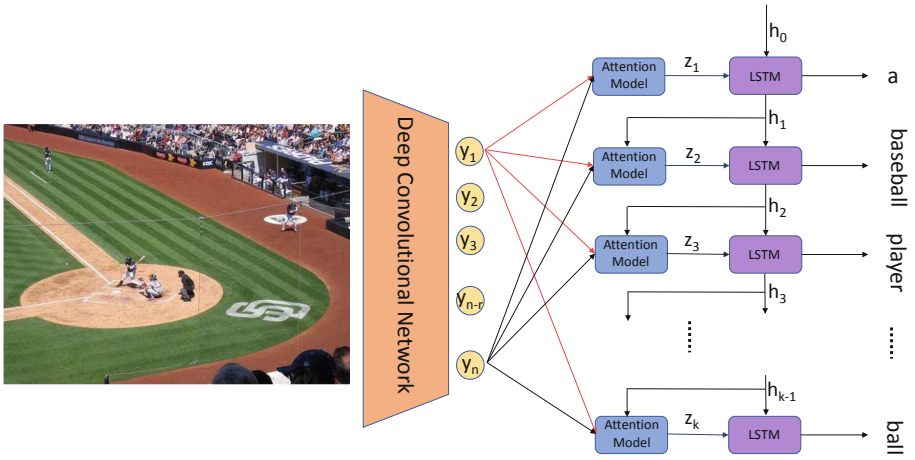


Fig. 4. Neural caption generation mechanism [32]

3.2 Features

We consider two different representations of the generated image captions for the purpose of training different classifiers for automating human judgement using active learning. First, we generate a dense vector representation of the captions using the pre-trained Universal Sentence Encoder [4]. It is a 512-dimensional vector, representing each caption, which promises to capture the context and semantic meaning of the sentence. We consider this representation to be useful for identifying syntactic or grammatical accuracy, image relevance and for identifying diverse captions, i.e., the ones which are more informative compared to the other describing the same image. The second representation for captions that we test is a 10 dimensional feature vector formed from different metrics which are popular in the caption generation community. These include overlap scores, such as, BLEU [27], ROUGE [21], METEOR [1] between the model generated captions and their corresponding groundtruths. Also Levenshtein distance, Levenshtein ratio and the ratio of number of unique unigrams, unique bigrams in the set of generated captions compared to the total number of words in the set of the generated captions.

3.3 Active Learning

We use pool-based active learning with uncertainty sampling for automating judgement on the quality of generated captions. We model the tasks of automatic human judgement on grammatical correctness, image relevance and diversity as binary classification problems. We initially select a random batch of generated captions and obtain human judgement labels for them using the crowdsourcing platform (Figure Eight https://www.figure-eight.com/). For each task, we train

different classifiers with this initial labeled data and then apply them to the unlabeled pool of captions to predict their labels. For every active learning iteration we select 200 instances on which prediction probabilities for the binary labels, for each task, are between the threshold 0.45 to 0.55. These instances are annotated by crowdworkers and incorporated into the training set for re-training the respective classifiers. This cycle is repeated 4 times. For each iteration, we report the performances of the classifiers on a completely separate human labeled test set. Figure 3 provides a schematic diagram for the entire process.

We use a SVM classifier with three different oversampling techniques for handling data imbalance in grammatical correctness and relevance estimation: Random Oversampling (ROS), Synthetic Minority Oversampling [5] technique (SMOTE) and Adaptive Synthetic [8] oversampling (ADASYN). We use the SVM and logistic regression without any oversampling for inferring human judgments on diversity, because the labels are balanced. The data imbalance for grammatical correctness and image relevance stems from the fact that most of the model generated captions are grammatically correct and relevant to their corresponding images compared to the few which are incorrect. Whereas, for diversity the data is balanced as for each image there is only one caption which is diverse and another which is not diverse.

In brief, ROS employs oversampling randomly to handle the issue of class imbalance whereas SMOTE is an oversampling approach where the minority class is oversampled by creating synthetic samples instead of oversampling with replacement. Oversampling for the minority class is done by considering each observation in the minority class and then generating synthetic examples along the line segments joining any or all of the k minority class nearest neighbors. The k nearest neighbors are chosen randomly depending upon the amount of oversampling needed. ADASYN on the other hand, aims to reduce the learning bias introduced by the original imbalance in the data distribution and at the same time, it adaptively shifts the decision boundary to focus on those samples which are difficult to learn.

4 Experiments and Results

For automatically determining human judgement on the three quality aspects of the generated captions, we first conduct surveys on a crowdsourcing platform to obtain the labels for an initial batch of randomly selected captions. We train different classifiers using these labels under an active learning scheme and report their performances on a separate test set. The labels for the test set are obtained separately using the same crowdsourcing platform.

We show that the performance of the classifiers, under active learning, using the 512-dimensional feature vector representation obtained from the sentence encoder [4] is much better compared to the representation using standard metrics based vector for all the tasks. This also establishes a new baseline for generating automatic human judgements without groundtruth information.

4.1 Results of Active Learning for Grammatical Accuracy

It is important to note that the dataset for grammatical accuracy is highly imbalanced since most of the model generated captions are grammatically correct. So, we combine different oversampling techniques (ROS, SMOTE and ADASYN) with a SVM and report the $F1$ score on the test set for initial (*Base*) and subsequent active learning iterations (*Iter 1–4*) for which the classifier is retrained. Table 1 shows the scores for models trained with the vector representations from the Universal Sentence Encoder and Table 2 for models based on the 10-dimensional metric vector. $F1$ scores from the two tables establish that standard metrics perform poorly in comparison to the features obtained from the universal sentence encoder for automating judgement on grammatical accuracy.

Table 1. Grammatical accuracy: $F1$ score of SVM using vector representation from Universal Sentence Encoder.

Classifier	Base	Iter1	Iter2	Iter3	Iter4
ROS + SVM	0.6650	0.6925	0.6922	0.6911	0.6821
SMOTE + SVM	0.6440	0.6711	0.6711	0.6794	0.6828
ADASYN + SVM	0.6651	0.6446	0.6505	0.6757	0.6559

Table 2. Grammatical accuracy: $F1$ score of SVM with sentence representation using metric scores.

Classifier	Base	Iter1	Iter2	Iter3	Iter4
ROS + SVM	0.4473	0.4445	0.4373	0.3998	0.3233
SMOTE + SVM	0.4722	0.4401	0.4202	0.3880	0.4115
ADASYN + SVM	0.3746	0.3839	0.2886	0.4444	0.4444

4.2 Results of Active Learning for Image Relevance

The dataset for image relevance also suffers from data imbalance, which is why we use SVMs in combination with oversampling, as well. We report the $F1$ score obtained with each combination for initial and subsequent active learning iterations on the test set for caption representations using Google's Universal Sentence Encoder [4] (see Table 3) and the one using a vector of overlap metrics discussed above (see Table 4). For automatic human judgment on image relevance of the generated captions, we see that the features from the sentence encoder produce superior results compared to the standard metric based features.

Table 3. Image relevance: $F1$ score of SVM with sentence representation from Universal Sentence Encoder.

Classifier	Base	Iter1	Iter2	Iter3	Iter4
ROS + SVM	0.5863	0.5982	0.6028	0.5807	0.6005
SMOTE + SVM	0.5940	0.6098	0.5886	0.5757	0.6110
ADASYN + SVM	0.5901	0.6024	0.6214	0.6075	0.6254

Table 4. Image relevance: $F1$ score of SVM with sentence representation using metric scores.

Classifier	Base	Iter1	Iter2	Iter3	Iter4
ROS + SVM	0.5709	0.5389	0.5389	0.5399	0.5306
SMOTE + SVM	0.5706	0.5446	0.5315	0.5306	0.5122
ADASYN + SVM	0.4002	0.4138	0.5709	0.5306	0.5211

4.3 Results of Active Learning for Diversity

Finally, we report the $F1$ scores for predicting human judgement on the diversity of the generated captions using logistic regression and SVM models. Since the dataset for diversity is balanced, we do not use any of the oversampling techniques. Table 5 shows the scores for models using the Universal Sentence Encoder, Table 6 the scores for models using the metric vector.

From the tables below, we see that for automatically determining human judgement on diversity of the generated captions, we see that feature vectors obtained from the sentence encoder do not provide significant advantage over the metric based vectors.

Table 5. Diversity: $F1$ score of classifiers with sentence representation from Universal Sentence Encoder.

Classifier	Base	Iter1	Iter2	Iter3	Iter4
Log. Reg.	0.5294	0.5175	0.5411	0.5400	0.5288
SVM	0.5288	0.5116	0.4642	0.4630	0.4658

Table 6. Diversity: $F1$ score of classifiers with sentence representation using metric scores.

Classifier	Base	Iter1	Iter2	Iter3	Iter4
Log. Reg.	0.529	0.558	0.482	0.490	0.57
SVM	0.523	0.530	0.52	0.50	0.58

5 Discussion

The results from our experiments show that feature vectors obtained from the pretrained sentence encoder [4] produce much higher $F1$ scores compared to standard overlap metrics when employed for the task of automatically inferring human judgement on neural network generated image captions. We believe the reason behind this performance increase is that the vectors from the sentence encoder capture the semantic and syntactic information present in the captions more than the standard overlap metrics such as BLEU, ROUGE, METEOR etc. Moreover, representing the generated captions with fixed size feature vectors, obtained from the pretrained sentence encoder [4], do not require corresponding groundtruth information for the captions. In our opinion, this is a major advantage over standard metrics which are completely dependent on groundtruth information.

The results further indicate that we can automate human judgement on grammatical accuracy and image relevance more successfully compared to automatically determining human judgement on diversity. However, we believe our approach, which combines feature vectors and standard ML classifiers under the active learning scheme, can significantly reduce annotation cost. In addition, the requirement for groundtruth information for automating human judgement on different quality aspects of neural network generated captions and NLG evaluation in general is reduced.

6 Conclusion

We implemented a technical architecture and conducted experiments to demonstrate that active learning can be used for automatically generating human judgement on the quality of the captions generated by a neural image caption generator. For this purpose, we tested sentence representations obtained from Google's Universal Sentence Encoder and another one obtained using standard metrics computed between the generated captions and their corresponding groundtruths. Subsequently, we trained SVM and logistic regression classifiers under an active learning framework and reported the $F1$ scores for a separate test set.

The $F1$ scores of the used classifiers show that under active learning better results are obtained using the 512 dimensional vectors from Universal Sentence Encoder across all three tasks. Also, we found that under active learning better results are obtained for the task of automating judgement on grammatical correctness and image relevance compared to the performance of automating judgement on diversity. Note that automatic human judgement on quality assessment is novel and an important step towards automated quality assessments in the evaluation of image captions and natural language generation in general. Our approach will be tested in future experiments as we believe it can reduce manual evaluation costs thereby simplifying NLG evaluation significantly.

Acknowledgement. This research was funded in part by the German Federal Ministry of Education and Research (BMBF) under grant number 01IS17043 (project SciBot). Aditya Mogadala was supported by the German Research Foundation (DFG) as part of SFB1102.

References

1. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72 (2005)
2. Barz, M., Polzehl, T., Sonntag, D.: Towards hybrid human-machine translation services. EasyChair Preprint (2018)
3. Biswas, R.: Diverse Image Caption Generation And Automated Human Judgement through Active Learning. Master's thesis, Saarland University (2019)
4. Cer, D., et al.: Universal sentence encoder. arXiv:1803.11175 (2018)
5. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W.: Smote: synthetic minority oversampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
6. Cho, K., Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: EMNLP (2014)
7. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. In: CVPR (2015)
8. Haibo, H., Bai, Y., Garcia, E., Li, S.: ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: IEEE International Joint Conference on Neural Networks, pp. 1322–1328 (2008)
9. Harnad, S.: The symbol grounding problem. Physica **42**, 335–346 (1990)
10. Harzig, P., Brehm, S., Lienhart, R., Kaiser, C., Schallner, R.: Multimodal image captioning for marketing analysis, February 2018
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
12. Hochreiter, S., Schmidhuber, J.: Long short term memory. Neural Comput. **9**, 1735–1780 (1997)
13. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167 (2015)
14. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)
15. Karpathy, A., Joulin, A., Fei-Fei, L.: Deep fragment embeddings for bidirectional image sentence mapping. In: NIPS (2014)
16. Kim, J., Rohrbach, A., Darrell, T., Canny, J., Akata, Z.: Textual explanations for self-driving vehicles. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11206, pp. 577–593. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01216-8_35
17. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
18. Kiros, R., Salahutdinov, R., Zemel, R.: Multimodal neural language models. In: ICLR, pp. 595–603 (2014)
19. Kiros, R., Salahutdinov, R., Zemel, R.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv:1411.2539 (2014)
20. Kisilev, P., Sason, E., Barkan, E., Hashoul, S.Y.: Medical image captioning : learning to describe medical image findings using multitask-loss CNN (2016)

21. Lin, C.: Rouge: a package for automatic evaluation of summaries. In: Text Summarization Branches Out (2004)
22. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
23. Lowerre, B., Reddy, R.: The harpy speech understanding system. In: Readings in Speech Recognition, pp. 576–586 (1990)
24. Mao, J., Xu, W., Yang, Y., Wang, J., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-RNN). arXiv:1412.6632 (2014)
25. Oviatt, S., Schuller, B., Cohen, P., Sonntag, D., Potamianos, G.: The Handbook Of Multimodal-Multisensor Interfaces, Volume 1: Foundations, User Modeling, and Common Modality Combinations. ACM, New York (2017)
26. Oviatt, S., Schuller, B., Cohen, P., Sonntag, D., Potamianos, G., Kruger, A.: Introduction: scope, trends, and paradigm shift in the field of computer interfaces, pp. 1–15. ACM, New York (2017)
27. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation. In: Association for Computational Linguistics, pp. 311–318 (2002)
28. Roy, D., Reiter, E.: Connecting language to the world. Artif. Intell. **167**, 1–12 (2005)
29. Settles, B.: Active Learning Literature Survey, vol. 52, no. 55-66, p. 11. University of Wisconsin, Madison (2010)
30. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: CVPR (2015)
31. Xu, A., Liu, Z., Guo, Y., Sinha, V., Akkiraju, R.: A new chatbot for customer service on social media. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 3506–3510 (2017)
32. Xu, K., er al.: Show, attend and tell: neural image caption generation with visual attention. In: ICML (2015)