Abstract

Quantitative modeling of human brain activity based on language representations has been actively studied in systems neuroscience. However, previous studies examined word-level representation, and little is known about whether we could recover structured sentences from brain activity. This study aims to generate natural language descriptions of semantic contents from human brain activity evoked by visual stimuli. To effectively use a small amount of available brain activity data, our proposed method employs a pre-trained image-captioning network model using a deep learning framework. To apply brain activity to the image-captioning network, we train regression models that learn the relationship between brain activity and deep-layer image features. The results demonstrate that the proposed model can decode brain activity and generate descriptions using natural language sentences. We also conducted several experiments with data from different subsets of brain regions known to process visual stimuli. The results suggest that semantic information for sentence generations is widespread across the entire cortex.

1 Introduction

Quantitative analysis of semantic activities in the human brain is an area of active study. With the development of machine learning methods and the application of such methods to natural language processing, many studies have attempted to interpret and represent brain activity with the semantics categories of words. In this paper, we propose a deep learning method to describe semantic representations evoked by visual stimuli, i.e., higher order perception, with natural language sentences using functional magnetic resonance imaging (fMRI) brain data. This requires a large amount of training data, i.e., brain activity data observed by fMRI. However, assembling a large-scale brain activity dataset is difficult because observing brain activity data with fMRI is expensive and each human brain is different in its size and shape. To handle this problem, we propose a model that associates the image features of the intermediate layer of a pre-trained caption-generation system with brain activity, which makes it possible to generate natural language descriptions of the semantic representation of brain activity. We used three methods to train the corresponding relationships between image features and the brain activity data, i.e., ridge regression, three-layer neural networks, and five-layer deep neural networks (DNN), and we compared the results. We also conducted an experiment using brain activity data from specific brain regions that process visual stimuli rather than data from the whole cortex to reduce the dimensionality of the input data assuming that the dimensionality of brain activity data is too large to train the model.
2 Related Work

2.1 Quantitative analysis of brain activity

Recently, many neuroscience studies have attempted to quantitatively analyze the semantic representation of what a human recalls using the fMRI data of brain activity evoked by visual stimuli, such as natural movies and images (Mitchell et al., 2008; Nishimoto et al., 2011; Pereira et al., 2013; Huth et al., 2012; Stansbury et al., 2013; Horikawa et al., 2013). Stansbury et al. (2013) employed latent Dirichlet allocation (Blei et al., 2003) to assign semantic labels to still pictures using natural language descriptions synchronized with the pictures and discussed the relationships between the visual stimuli evoked by the still pictures and brain activity. Based on these relationships, they constructed a model that classifies brain activity into semantic categories to reveal areas of the brain that deal with particular semantic categories. Cukur et al. (2013) estimated how people semantically change their recognition of objects from brain activity data in cases where the subject pays attention to objects in a video. Huth et al. (2012, 2016a,b) revealed the corresponding relationships between brain activities and visual stimuli using the semantic categories of WordNet (Miller, 1995). They used the categories to construct a map for semantic representation in the cerebral cortex and showed that semantic information is represented in various patterns over broad areas of the cortex. Nishida et al. (2015) showed that, compared to other language models such as Latent Semantic Indexing (Deerwester et al., 1990) and latent Dirichlet allocation, the word2vec (skip-gram) model by Mikolov et al. (2013) gives better accuracy in modeling the semantic representation of human brain activity. Furthermore, they showed that there is a correlation between the distributed semantics obtained based on skip-gram in the word2vec framework with the Japanese Wikipedia corpus and brain activity observed through blood oxygen level dependent (BOLD) contrast obtained using fMRI. As these studies indicate, statistical language models used to analyze semantic representation in human brain activity can explain higher order cognitive representations. In this context, although most previous studies use the semantic categories of words to describe the semantic representation of brain activity, in this study, we aim to take a further step toward quantitative analysis of this relationship using a caption-generation system. With this approach, we can describe brain activity with natural language sentences, which can represent richer information than words. In addition, Güçlü and van Gerven (2016) demonstrated that DNN trained for action recognition can be used to predict how the dorsal stream, a brain area that processes visual stimuli, responds to natural movies. They demonstrated a correlation between the intermediate representations of DNN layers and dorsal stream areas. This suggests the effectiveness of deep learning methods for the analysis and modeling of brain activities evoked by visual stimuli.

2.2 Caption generation from images

In natural language processing, deep learning has resulted in significant progress in natural language generation and machine translation. In particular, encoder-decoder (enc-dec) models, e.g., sequence-to-sequence models, have been studied (Sutskever et al., 2014; Cho et al., 2014; Kiros et al., 2013; 2014; Bahdanau et al., 2015; Cho et al., 2015). For example, such models have been applied to speech recognition (Chorowski et al., 2015), video captioning (Yao et al., 2015), and text summarization (Nallapati et al., 2016). Typically, two approaches have been used for image-caption generation. The first approach retrieves and ranks existing captions (Kuznetsova et al., 2012, 2014; Vendrov et al., 2016; Yagcioglu et al., 2015), and the second fills sentence templates based on features extracted from a given image (Elliott and Keller, 2013; Elliott and de Vries, 2015; Kulkarni et al., 2013; Mitchell et al., 2012). Recently, methods that employ an enc-dec framework to generate captions for images have been proposed (Donahue et al., 2015; Kiros et al., 2015; Mao et al., 2014; Vinyals et al., 2015). Vinyals et al. (2015) developed an image-caption generation method by building an enc-dec network employing GoogLeNet (ioffe and Szegedy, 2015), which can extract image features effectively, as an encoder and a Long Short-Term Memory Language Model (LSTM-LM) (Hochreiter and Schmidhuber, 1997; Sutskever et al., 2014), which is a deep neural language model, as an decoder. In this study, we build and train an enc-dec network based on those prior studies (Vinyals et al., 2015; Xu et al., 2015; Yao et al., 2016). However, we use brain activity data as input rather than images and attempt to generate natural language descriptions for the brain data.
3 Proposed Method

This study aims to generate natural language sentences that describe what a human being calls to mind using brain activity data observed by fMRI as input information. We combined a image → caption model and a brain activity data → image feature model (Sections 3.1 and 3.2). Figure 1 presents an overview of the proposed method.

3.1 Image → caption model (A)

We employed an image-captioning model (A) based on a DNN framework, i.e., the enc-dec network [Cho et al., 2015; Vinyals et al., 2015], as the main component of the proposed model. In the enc-dec framework, by combining two DNN models functioning as an encoder and a decoder, the model encodes input information as an intermediate expression and decodes the information as an expression in a different modality. Generally, previous studies of image-captioning systems have proposed enc-dec models that combine two DNNs: one DNN extracts image features using a convolutional neural network and the other generates captions using a LSTM with the image features which correspond to an intermediate expression of the model. Similar to such previous models, we constructed an image → image feature → caption model (A) employing VGGNet (Simonyan and Zisserman, 2015) as an encoder and a two-layer LSTM language model (Hochreiter and Schmidhuber, 1997; Sutskever et al., 2014) as a decoder. We used pairs of image and caption data as training data.

3.2 Brain activity data → image feature model (B)

To apply the above image-captioning process to handle brain activity data rather than images, we constructed a model that predicts features extracted by VGGNet from images that evoke visual stimuli in the brain using fMRI brain activity data as input. In other words, the model encodes brain activity data into the intermediate expression in the image → caption model (A). We implemented and compared three models, i.e., a ridge regression model, a three-layer neural network model, and a five-layer DNN model, to determine which machine learning method is suitable for this model. The five-layer DNN model was pre-trained using stacking autoencoders (Bengio et al., 2006) to avoid delay and overfitting in training due to the lack of brain data. We used pairs of fMRI brain activity data and the images a subject observed as training data; however, we did not use natural language descriptions.

4 Process Flow

The process of the proposed method is as follows.

**Step 1. Encode brain activity as an intermediate expression.**

Model (B) predicts the feature of the image a subject watches from the input brain activity data evoked by the visual stimuli. In the following step, the features are provided to model (A) and processed as the intermediate expression.
Table 1: Experimental settings

| Dataset          | Training quantity | Algorithm | Hyper-parameters | Initial parameters | Units per layer | Vocabulary |
|------------------|-------------------|-----------|------------------|--------------------|-----------------|------------|
| image → image feature model | 414,113 samples × 100 epochs | Ridge regression | α = 0.001, β1 = 0.9, β2 = 0.999, eps = 1e-8 | Pre-trained word2vec & VGGNet: pre-trained | 512 x 512 | Frequent 3,469 words |
| brain activity data → image feature model | 4,500 samples × 1,000 epochs | Stochastic gradient descent | gradient clipping threshold : 1, L2-norm : 0.005 | Std normal random numbers | 65,665-4,096 | |
| 1: Ridge Regression | 2: three-layer NN | 3: five-layer DNN | | | |

Step 2-1. Word estimation by the LSTM-LM.

The LSTM-LM decoder of model (A) predicts the next word from the image feature produced in Step 1 and the hidden states of LSTM at the previous time step.

Step 2-2. Caption generation by iterative word estimation.

A caption is generated by sequentially estimating words by repeating Step 2-1 until either the length of the sentence exceeds a predefined maximum or the terminal symbol of a sentence is output.

As mentioned above, we construct a brain activity data → image feature → caption model (C) by training the brain activity data → image feature model (B) and the image → image feature → caption model (A) individually and execute them sequentially in the prediction phase. Note that model (C) uses only fMRI brain activity data as input, i.e., without images.

5 Experiments

Chainer\[^1\] was used as the deep learning framework to construct the neural networks.

5.1 Experiment (A): image → caption model

5.1.1 Experimental settings

Microsoft COCO\[^2\], which includes 414,113 pairs of images and their captions, was used as the dataset for the experiments. The hyper-parameters of the models used in the experiments were set based on previous studies (Vinyals et al. 2015; Mitchell et al. 2016; Cho et al. 2014). The parameters to be learned were initialized by random values obtained based on standard normal distribution. The initial parameters of the word embedding layer were those used by word2vec learned using a skip-gram with window size = 5. The pre-trained synaptic weights and hyper-parameters were used for VGGNet. Furthermore, 3,469 words that appear more than 50 times in the training data were used to generate natural language descriptions. The settings for learning are shown in the leftmost column of Table 1.

5.1.2 Results & Discussion

We confirmed the process of learning by the convergence of the perplexity of output sentences recorded for each epoch. Figure 2 shows the natural language descriptions for two images randomly selected from test images. In the first example, a considerably reasonable natural language description was generated. In the second example, appropriate expressions for the subject of the generated sentence, i.e., a cat, and its color where selected. Furthermore, the prepositions, e.g., in and on, and articles, e.g., a and an, were also used correctly. Therefore, the output sentences correctly describe the content of the images. As shown above, reasonable captions were generated for the test images and the perplexity converged near 2.5; therefore, an appropriate model was built to generate natural language descriptions from the images. Note that errors that frequently appear in the generated

---

1. http://chainer.org/
2. http://mscoco.org/

Figure 2: Captions for randomly selected images

A man is surfing in the ocean on his surfboard. A black and white cat is sitting on the toilet.
sentences seem to depend on image processing rather than language processing. For instance, in the second example, standing is expressed as sitting and a washbasin is expressed as a toilet.

5.2 Experiment (B-1): brain → image feature model

5.2.1 Experimental settings

For the learning dataset for the corresponding relationships between brain activity and image features, we employed the brain activity data of a subject stimulated by natural movies (Nishimoto et al., 2011), i.e., the BOLD signal observed by fMRI, and still pictures taken from the movies provided as visual stimuli, which were synchronized with the brain activity data. In the natural movies, there are various kinds of movies about natural phenomenon, artifacts, humans, films, 3D animations, etc., whose length of time are a few tens of seconds. As the input data, we employed 65,665 voxels corresponding to the cerebral cortex part among 96 × 96 × 72 voxels observed by fMRI, (see, Figure 4), then learned the corresponding relationships between the brain data and the image features, whose dimensionality is 4,096, extracted from the image using VGGNet. We used 4,500 samples as training data (recorded every two seconds for 9,000 seconds), which is a small number for learning a DNN. In terms of ridge regression and the three-layer neural network model, the parameters to be learned were initialized with random values obtained from the normal standard distribution. For the five-layer DNN model, to increase the speed of learning and avoid overfitting, we used 7,540 unlabeled brain data samples to pre-train the networks using autoencoders for 200 epochs per layer, and we used the obtained synaptic weights as the initial network parameters. The learning settings are shown in the three columns on the right side of Table 1.

5.2.2 Results & Discussion

We recorded the mean squared error (MSE) for each epoch and confirmed that the MSEs of the three models converged. For evaluation, we conducted an experiment to retrieve the images, which have similar image features to those estimated from brain activity data, from 82,783 images of the Microsoft COCO training dataset with MSE metric. Figure 3 shows the result of retrieving similar images. As for ridge regression and three-layer neural network, those models retrieved proper images from most training data, so we confirmed that the models could extract proper image features from brain activity data. Furthermore, it also could be said that the images similar to the stimulation images which evoked brain activities were retrieved from the image database. However, as for five-layer DNN, the same unrelated images were retrieved for all input brain activity data, and the results for the test samples were worse than those for training samples even when employing ridge regression or the three-layer neural network although the MSEs for the test dataset converged. The reason for this
Figure 5: Exp. (C-1): Stimulation images and generated descriptions

is that the input dimension, i.e., 65,665, was much larger than that of the parameters to be learned, and the number of training data samples, i.e., 4,500, was small. As a result, overfitting occurred due to the lack of adjustment of hyper-parameters.

5.3 Experiment (C-1): brain → caption model

5.3.1 Experimental settings

We built a model that generates a natural language description from brain activity data by combining the model in Experiment (A) and the three models in Experiment (B-1). We then generated descriptions based on the three methods, i.e., ridge regression, the three-layer neural network, and the five-layer DNN. In addition, we generated captions from the same images used in the above experiments with model (A).

5.3.2 Results & Discussion

The natural language descriptions generated from the four brain activity data samples (i.e., two training data and two test data), and their images are shown in Figure 5.3.1. To compare the results, we also show the captions generated using the model in Experiment (A). The models were evaluated with BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014), which are the most commonly automatic evaluation metrics in the caption generation literature (see, Figure 8). However, there are not available groundtruth sentences, i.e., human generated descriptions for the images that evoked brain activities, which are necessary for computing these metrics. Thus, we generated captions directly from 300 image samples of the test brain dataset using the image → caption model in Experiment (A) and then selected 60 samples among those whose generated sentences were evaluated as appropriate by a human. For each image sample, we regarded 10 sentences generated by 10-best beam search as groundtruth data. Note that there are several criticisms made of the BLEU and METEOR evaluation metrics, although they are considered standard metrics for translation and captioning tasks.

Human understandable natural language descriptions were generated using only brain activity data. As well as the model in Experiment (A), noun phrases or sentences with correct grammar, including prepositions and articles, were generated stably. The descriptions generated using the brain activity data and those using the images were nearly the same for the models employing ridge regression and the three-layer neural network. Thus, we confirm that learning the corresponding relationships between brain activity data and image features was successful, and the proposed method functioned well. Taking into account of the results of Experiment (B-1), it was considered natural to find the same sentences were generated for all input information in the case of using the five-layer DNN model, and the quality of the generated sentences was low even when employing ridge regression or the three-layer neural network to learn the relationship. In addition, overfitting also occurred using the five-layer DNN without pre-training with autoencoders. Furthermore, as discussed relative to Experiment (A), the caption-generation model was learned properly, however the descriptions generated directly from the images were somewhat improper for the second test example because the quality of the images differed, i.e., images in Microsoft COCO dataset were prepared for image recognition. Therefore, the content of the images was considerably understandable and describable using natural language. On the other hand, the natural movies were single-shot pictures of various types of movies that included blurring, darkening, letters, animation, etc. The content of some of these
A young man is doing tricks on his skateboard.
A man is playing tennis on the court. A young man is playing tennis on the court.
A man is playing tennis on the court with his racket.
A man is playing tennis on the court with his racket.
A man sitting on the ground with an umbrella.
A polar bear is standing in the water.
A dog laying on the ground next to an orange frisbee.
A dog laying on the ground next to an orange frisbee.
A black and white dog laying on the ground.
A dog is sitting on the floor in front of an open door.

Figure 7: Exp. (C-2): Stimulation images and generated descriptions

images was difficult to describe using natural language. Interestingly, a proper caption was generated using brain activity data with the three-layer neural network model compared to the image-captioning model for the second training example. It is unlikely that a human would confuse a clock with a pair of scissors. However, the image-captioning model made this mistake due to image processing errors in VGGNet. Thus, in this case, we assume that the image features obtained using brain activity data worked better than the features obtained directly from the images.

5.4 Experiment (B-2): brain activity data → image feature model

In Experiment (C-1), the model trained in Experiment (B-1) was overfitted because there was an insufficient amount of brain activity data and the dimension of the data was very large. Therefore, we conducted an experiment using parts of the brain that react to visual stimuli for dimension reduction, and then trained brain activity data → image feature models (B) for higher quality.

5.4.1 Experimental settings

For the training dataset, among the 65,665 voxels of the brain activity data evoked by the visual stimuli used in Experiment (B-1), the particular areas of the cerebral cortex used for image processing were selected as input information. Nishida et al. (2015) built a model that learns the corresponding relationships between brain activity data evoked by visual stimuli, i.e., images, and the distributed semantics based on a skip-gram that represents the content of images. They predicted the accuracy of how much each voxel contributes to the prediction of the corresponding relationships by means of the correlation co-efficiency obtained using the learned model. In this study, with prediction accuracy thresholds of c.c. = 0.05, 0.1, 0.15, and 0.2, useful voxels for image processing were selected to predict semantic representations, and the number of voxels for each threshold setting was 21,437, 9,923, 5,961, and 3,539, respectively. In contrast, as higher dimensional input information, we also conducted an experiment with the brain data whose dimensionality is 89,206 by adding the subcortex area, which primarily governs memory and space perception, in addition to the cerebral cortex. Here, the three-layer neural network was used as the learning model and the experimental settings were the same as those of Experiment (B-1).

5.4.2 Results & Discussion

Figure 6 shows the changes of the MSEs on the test dataset for each epoch.

In the experiments with 89,206 and 21,437 dimensional data, the MSEs (approximately 1.33 and 1.16, respectively) were nearly the same as that obtained with 65,665 dimensional data, i.e., Experiment (B-1) with the three-layer neural network (approximately 1.11). On the other hand, for the experiments with 9,923, 5,961, and 21,437 dimensional data, we confirmed that overfitting occurred. The minimum errors were 1.17 (24 epochs), 1.13 (30 epochs), and 1.10 (30 epochs), respectively. We obtained the best MSE result using 3,539 dimensional data. We compared and discussed six types of models in the following section.

5.5 Experiment (C-2): brain → caption model

5.5.1 Experimental settings

Six natural language sentences were generated from brain activity data by combining the model in Experiment (A) and the six models in Experiment (B-2). For the three models with 9923, 5961, and
3539 dimensions, in which overfitting occurred, we employed the parameters at the epoch when MSE was the lowest score.

Figure 8: Exps. (C-1) & (C-2): BLEU-4 & METEOR scores

5.5.2 Results & Discussion

Six descriptions obtained for the two brain activity data samples selected from the training data. Those descriptions and their images are shown in Figure 7. Figure 8 shows BLEU-4 and METEOR scores computed in the same way as Experiment (C-1).

Against expectation, it was observed that the model trained with complete cerebral cortex data tended to produce better descriptions than the other models. In contrast, improper descriptions were generated with the data from the particular region of the cerebral cortex. This was also observed when using the model obtained at 100 epochs. The quality of the descriptions with specific brain regions was worse than that of the model trained with the whole cortex data, nevertheless learning the corresponding relationships between brain activity and visual stimuli should become easier when handling low dimensional data. Thus, we assume that the information required to predict semantic information to understand images is included in brain areas that barely react to visual stimuli. This result agrees with the result obtained by Cukur et al. (2013), i.e., most brain regions are modulated by image processing rather than particular regions of the cortex. Moreover, this also demonstrates that neural networks are effective for extracting important features from high dimensional data.

As for the evaluation metrics, we confirmed that the BLUE-4 and METEOR scores for five-layer DNNs, where overfitting clearly occurred, were the lowest for most of the models. In addition, especially as for the METEOR score, the models with three-layer neural networks got slightly better results than those with ridge regression. This corresponds to the subjective evaluation based on generated descriptions mentioned above.

6 Conclusions

We have proposed a method to generate descriptions using brain activity data by employing a framework to generate captions for images using DNNs and by learning the corresponding relationships between brain activity data and the image features extracted using VGGNet. We constructed models based on three experimental settings for training methods, and we were successful in generating natural language descriptions from brain activity data evoked by visual stimuli. The quality of the descriptions was higher when using a three-layer neural network. Moreover, by carefully examining the descriptions generated using models trained with data from different regions of the cortex, our results suggest that most brain regions are modulated by visual processing in the human brain. In future, we plan to increase the amount of brain activity data, apply additional machine learning methods, i.e., Bayesian optimization, whitening, etc., and revise the hyper-parameters to increase prediction accuracy. Furthermore, we would like to investigate proper objective methods to evaluate the generated natural language descriptions.
References

D. Bahdanau, K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. ICLR’15.

Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. 2006. Greedy layer-wise training of deep networks. NIPS’06 19:153–160.

D. Blei, A. Ng, and M. Jordan. 2003. Latent dirichlet allocation. Journal of Machine Learning Research 3:993–1022.

K. Cho, A. Courville, and Y. Bengio. 2015. Describing multimedia content using attention based encoder decoder networks. Multimedia, IEEE Transactions 17(11):1875–1886.

K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. EMNLP’14.

J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio. 2015. Attention-based models for speech recognition. arXiv preprint: 1506.07503.

T. Cukur, S. Nishimoto, A. G. Hut, and J. L. Gallant. 2013. Attention during natural vision warps semantic representation across the human brain. Nature Neuroscience 16:763–770.

S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. 1990. Indexing by latent semantic analysis. the American Society of Information Science 41(6):391–407.

M. Denkowski and A. Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. EACL’14 Workshop on Statistical Machine Translation.

J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. CVPR’15.

D. Elliott and A. P. de Vries. 2015. Describing images using inferred visual dependency representations. ACL’15.

D. Elliott and F. Keller. 2013. Image description using visual dependency representations. EMNLP’13.

U. Güçlü and M. A.J. van Gerven. 2016. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. NeuroImage 145:329–336.

S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. Neural Computation 9(8).

T. Horikawa, M. Tamaki, Y. Miyawaki, and Y. Kamitani. 2013. Neural decoding of visual imagery during sleep. Science 340.

A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant. 2016a. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532:453–458.

A. G. Huth, T. Lee, S. Nishimoto, N. Y. Bilenko, A. T. Vu, and J. L. Gallant. 2016b. Decoding the semantic content of natural movies from human brain activity. Frontiers in Systems Neuroscience 10(81).

A. G. Huth, S. Nishimoto, A. T. Vu, and J. L. Gallant. 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76(6):1210–1224.

S. Ioffe and C. Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint:1502.03167.

R. Kiros, R. Salakhudtinov, and R. Zemel. 2013. Recurrent continuous translation models. EMNLP’13.

R. Kiros, R. Salakhudtinov, and R. Zemel. 2014. Multimodal neural language models. ICML’14.
R. Kiros, R. Salakhutdinov, and R. Zemel. 2015. Unifying visual-semantic embeddings with multimodal neural language models. *NIPS’15 Deep Learning Workshop*.

G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *Pattern Analysis, Machine Intelligence and IEEE Transactions on 35*(12):2891–2903.

P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. 2012. Collective generation of natural image descriptions. *ACL’12*.

P. Kuznetsova, V. Ordonez, T. L. Berg, and Y. Choi. 2014. Treetalk: Composition and compression of trees for image descriptions. *ACL’14*.

J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. 2014. Deep captioning with multimodal recurrent neural networks (m-rnn). *ICLR’14*.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. *NIPS’13*.

G. A. Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 138:39–41.

M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daume III. 2012. Midge: Generating image descriptions from computer vision detections. in european chapter of the association for computational linguistics. *ACL’12*.

M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daume III. 2016. Generating natural questions about an image. *ACL’16*.

T. M. Mitchell, S. V. Shinkareva, A. Carlson, K. M. Chang, V. L. Malave, R. A. Mason, and M. A. Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science* 320(1191).

R. Nallapati, B. Zhou, C. Nogueira dos santos, C. Gulcehre, and B. Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *CoNLL’16*.

S. Nishida, A. G. Huth, J. L. Gallant, and S. Nishimoto. 2015. Word statistics in large-scale texts explain the human cortical semantic representation of objects, actions, and impressions. *Society for Neuroscience Annual Meeting 2015* 333(13).

S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant. 2011. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology* 21(19):1641–1646.

K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. *ACL’02*.

F. Pereira, M. Botvinicka, and G. Detre. 2013. Using wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. *Artificial Intelligence* 194:240–252.

K. Simonyan and A. Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *ICLR’15*.

D. E. Stansbury, T. Naselaris, and J. L. Gallant. 2013. Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron* 79:1025–1034.

I. Sutskever, O. Vinyals, and Q. V. Le. 2014. Sequence to sequence learning with neural networks. *NIPS’14*.

I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun. 2016. Order-embeddings of images and language. *ICLR’16*.

O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. Show and tell: A neural image caption generator. *CVPR’15*. 
K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *ICML’15*.

S. Yagcioglu, E. Erdem, A. Erdem, and R. Cakici. 2015. A distributed representation based query expansion approach for image captioning. *ACL’15*.

L. Yao, N. Ballas, K. Cho, J. R. Smith, and Y. Bengio. 2016. Empirical upper bounds for image and video captioning. *ICLR’16 workshop*.

L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. 2015. Describing videos by exploiting temporal structure. *ICCV’15*.