Generating Natural Language Descriptions for Semantic Representations of Human Brain Activity

Eri Matsuo
Ichiro Kobayashi
Ochanomizu University
g1220535@is.ocha.ac.jp
koba@is.ocha.ac.jp

Shinji Nishimoto
Satoshi Nishida
Center for Information and Neural Networks,
National Institute of Information and Communications Technology
nishimoto@nict.go.jp
s-nishida@nict.go.jp

Hideki Asoh
Artificial Intelligence Research Center,
National Institute of Advanced Industrial Science and Technology
h.asoh@aist.go.jp

Abstract

Quantitative analysis of human brain activity based on language representations, such as the semantic categories of words, have been actively studied in the field of brain and neuroscience. Our study aims to generate natural language descriptions for human brain activation phenomena caused by visual stimulus by employing deep learning methods, which have gained interest as an effective approach to automatically describe natural language expressions for various type of multi-modal information, such as images. We employed an image-captioning system based on a deep learning framework as the basis for our method by learning the relationship between the brain activity data and the features of an intermediate expression of the deep neural network owing to lack of training brain data. We conducted three experiments and were able to generate natural language sentences which enabled us to quantitatively interpret brain activity.

1 Introduction

In the field of brain and neuroscience, analyzing semantic activities occurring in the human brain is an area of active study. Meanwhile, in the field of computational linguistics, the recent evolution of deep learning methods has allowed methods of generating captions for images to be actively studied. Combining these backgrounds, we propose a method to quantitatively interpret the states of the human brain with natural language descriptions, referring to prior methods developed in the fields of both brain and neuroscience and computational linguistics. Because it is difficult to prepare a large-scale brain activity dataset to train a deep neural model of generating captions for brain activity from scratch, therefore, to handle this problem, we instead reuse a model trained to generate captions for images as the basis for our method. We apply brain activity data, instead of images, to the image caption-generation frameworks proposed by Vinyals et al.(2015) and Xu et al.(2015) to generate natural language descriptions expressing the contents of the brain activity. In this way, we aim to achieve a quantitative analysis of brain activities through language representation.

2 Related Studies

2.1 Language representation estimated from brain activity

In recent years, in the field of brain and neuroscience, the quantitative analysis of what a human recalls using brain activity data observed via functional magnetic resonance imaging (fMRI) while he or she watches motion pictures has been actively studied (Mitchell et al., 2008; Nishimoto et al., 2011; Pereira et al., 2013; Huth et al., 2012; Stansbury et al., 2013; Horikawa et al., 2013). Huth et al. (2012) created a map for semantic representation at the cerebral cortex by revealing the corresponding relationships between brain activities and the words of WordNet (Miller, 1995), thus representing objects and actions in motion pictures. Stansbury et al. (2013) employed Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to assign semantic labels to still pictures using natural language descriptions synchronized with the pictures and discussed the resulting relationship between the visual stimulus evoked by the still pictures and brain activity. Based on these relationships, they have built a model that classifies brain activity into semantic categories, revealing the areas of the brain that deal with particular categories. Cukur et al. (2013) estimated how a human being

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics – Student Research Workshop, pages 22–29,
Berlin, Germany, August 7-12, 2016.

semantically changes his or her recognition of objects from the brain activity data in cases where he or she pays attention to objects in a motion picture. As mentioned above, Statistical models analyzing semantic representation in human brain activity have attracted considerable attention as appropriate models to explain higher order cognitive representations based on human sensory or contextual information.

Furthermore, Nishida et al.(2015) demonstrated that skip-gram, employed in the framework of word2vec proposed by Mikolov (2013), is a more appropriate model than the conventional statistical models used for the quantitative analysis of semantic representation in human brain activity under the same experimental settings as the prior studies. Moreover, they showed that there is a correlation between the distributed semantics, obtained by employing skip-gram to build distributed semantic vectors in the framework of word2vec with the Japanese Wikipedia corpus, and brain activity observed through blood oxygen level dependent (BOLD) contrast imaging via fMRI.

Prior studies have attempted to quantitatively analyze the relationship between semantic categories and human brain activity from the perspective of language representation, especially, the semantic categories of words. In this study, we aim to take a step further toward quantitatively analyzing this relationship by expressing brain activity with natural language descriptions.

2.2 Caption generation from images

Many previous studies on image caption generation have been based on two principal approaches. The first approach is to retrieve existing captions from a large database for a given image by ranking the captions (Kuznetsova et al., 2012; Kuznetsova et al., 2014; Vendrov et al., 2016; Yagcioglu et al., 2015). The second approach is to fill sentence templates based on the features extracted from a given image, such as objects and spatial relationships (Elliott and Keller, 2013; Elliott and Vries, 2015; Kulkarni et al., 2013; Mitchell et al., 2012). Although these approaches can produce accurate descriptions, they are neither flexible nor natural descriptions such as the ones written by humans. Recently, multiple methods proposed for generating captions for images have been developed based on the encoder-decoder (enc-dec) framework (Cho et al., 2014; Cho et al., 2015), which is typically used for media transformation (Chorowski, 2015), e.g., machine translation (Sutskever et al., 2014; Cho et al., 2014; Kiros et al., 2014; Bahdanau et al., 2015), to generate captions for images (Donahue et al., 2015; Kiros et al., 2015; Mao et al., 2014; Vinyals et al., 2015).

In the enc-dec framework, by combining two deep neural network models functioning as an encoder and a decoder, the enc-dec model first encodes input information into an intermediate expression and then decodes it into an expression in a different modality than that of the input information. Vinyals et al. (2015) achieved caption generation for images by building a enc-dec network employing GoogLeNet (Ioffe and Szegedy, 2015), which works effectively to extract the features of images, as the encoder, and Long Short-Term Memory Language Model (LSTM-LM) (Hochreiter and Schmidhuber, 1997; Sutskever et al., 2014), which is a deep neural language model, as the decoder. Xu et al. (2015) proposed a model using the Attention Mechanism (Cho et al., 2015) and demonstrated that the model achieved high precision when generating captions. Attention Mechanism is a system that automatically learns to pay attention to different parts of the input for each element of the output (Bahdanau et al., 2015; Cho et al., 2015; Yao et al., 2015).

In our study, we provide an enc-dec network with brain activity data as input, instead of an image, and attempt to generate natural language descriptions for this data.

3 Proposed Method

First, the process to generate captions for images using deep neural networks, employed in Vinyals et al.(2015) and Xu et al.(2015), works as follows:

Step 1. *Encoder: Extraction of features using VGGNet*
The encoder VGGNet (Simonyan and Zisserman, 2015), a pre-trained deep convolutional network, extracts the features from the input image. In the case with Attention Mechanism, the output of the encoder is $512 \times 14 \times 14$ dimensional data from the intermediate convolutional layer of VGGNet. In the case without Attention Mechanism, the output is 4,096 dimensional data from the last fully-connected layer of VGGNet.

Step 2. *Process for intermediate expression*
In the case with Attention Mechanism, the weighted sum of the set of the intermediate expressions calculated in Step 1 is computed as the input for the decoder. The weighted coefficients are learned by means of a multi-layered perceptron based on the hidden states of the decoder at the previous time step and 512 intermediate expressions. In the case without Attention Mechanism, the output of the encoder from Step 1 is just the input of the decoder in Step 3.

Step 3. *Decoder: Word estimation by LSTM-LM*
The LSTM-LM decoder predicts the next word from the intermediate expression produced in Step 2 and the hidden states of LSTM at the previous time step.

Step 4. *Caption generation by iterative word estimation*
A caption is generated by estimating the words one-by-one repeating Steps 2 and 3 until either the length of the sentence exceeds the predefined maximum or the terminal symbol of a sentence is output.

This study aims to generate natural language sentences that describe the events a human being calls to mind from the human brain activity input data observed by fMRI via the above caption-generation process. Figures 1 and 2 show overviews of our methods with and without Attention Mechanism, respectively. In essence, we train a simple model, a 3-layered perceptron (multi-layered perceptron; MLP) or ridge regression model, to learn the corresponding relationships between the cerebral nerve activity data stimulated by the input images and the features of the same image extracted by VGGNet, namely, the intermediate expression as for the image caption generator. The model replaces VGGNet as the encoder when brain activity data are used as input information instead of images. Then, the rest of the process to generate captions is the same as that of the above image caption generator. The process of the proposed method is as follows:

Step 1. *Encode brain activity to an intermediate expression*
The model, which is pre-trained to learn the mapping from the brain activity data stimulated by an image to the features extracted from the same image by VGGNet, maps the input brain data to an intermediate expression.

Step 2 $\sim$ **4.** The rest of the process is the same as above.

4 Experiments

In this study, we conducted three experiments, under the conditions shown in Table 2, using the model of caption generation for images and the model to learn the corresponding relationships between the brain activity data and the features obtained from VGGNet. The model for Exp.1 is illustrated in Figure 1, and the models for both Exps.2 and 3 are illustrated in Figure 2.

4.1 Experimental settings

We employed Chainer[1] as the deep-learning framework. We used Microsoft COCO[2], which contains 414,113 pairs of data with still pictures and natural language descriptions of their contents, as the training data for the building caption-generation model. In this study, we have so far been able to train the network with 168,000 pairs of the total dataset in the below experiments.

Table 2: The experimental setting.

Exp.	Image to Caption	Brain to Intermediate
Exp.1	Attention Mechanism	3-layered MLP
Exp.2	Without	(Neural Network)
Exp.3	Attention Mechanism	Ridge regression

We employed the brain activity data of a subject being stimulated by motion pictures (Nishimoto et al., 2011) as the data for training and evaluation. In the experiments, we used BOLD signals observed every 2s via fMRI while the subject was watching motion pictures as the brain activity data, and the still pictures extracted from the motion pictures were synchronized with the brain data. The brain activity data were observed throughout the brain and were recorded in $100(x)\times100(y)\times32(z)$ voxels. We employed 30,662 voxels corresponding to only the cerebral cortex region, which is the area of the whole brain, in the above observed voxels as input brain data (see, Figure 3). In the Exp.1, the multi-layered perceptron learns the corresponding relationships between the input 30,662 dimensional brain data and the $14\times14\times512$=100,352 dimensional data of the intermediate layer of VGGNet. In Exps.2 and 3, the 4,096 dimensional feature vector output by VGGNet is the target that needs to be correlated with the brain activity data. We have only 3,600 training brain activity data which are too small to train deep neural networks, so we have applied a pre-trained deep neural image caption generator to the task for describing brain activity caused by visual stimulation.

4.2 Exp.1: With Attention Mechanism and 3-layered MLP

First, we confirmed that our caption-generation model with the Attention Mechanism was well trained and had learned "attention" by generating captions and visualizing the attention for two pictures randomly selected from the COCO test dataset, as shown in Figure 4. Figure 5 shows two sets of still pictures and the generated descriptions for the brain activity data selected from the test dataset. Table 3 shows the perplexity of the generated captions for the COCO images and the

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

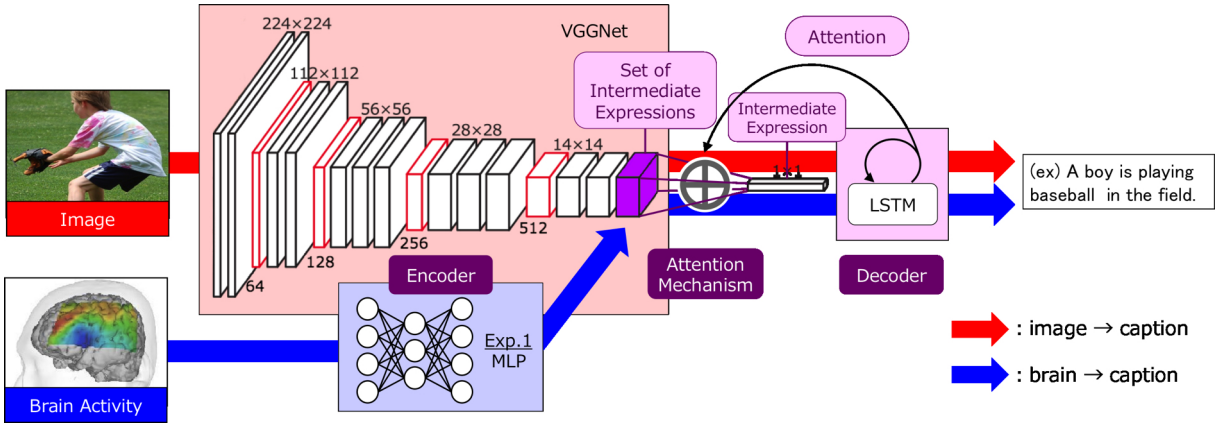

Figure 1: Overview of our approach (Exp.1: With Attention Mechanism and 3-Layered MLP).

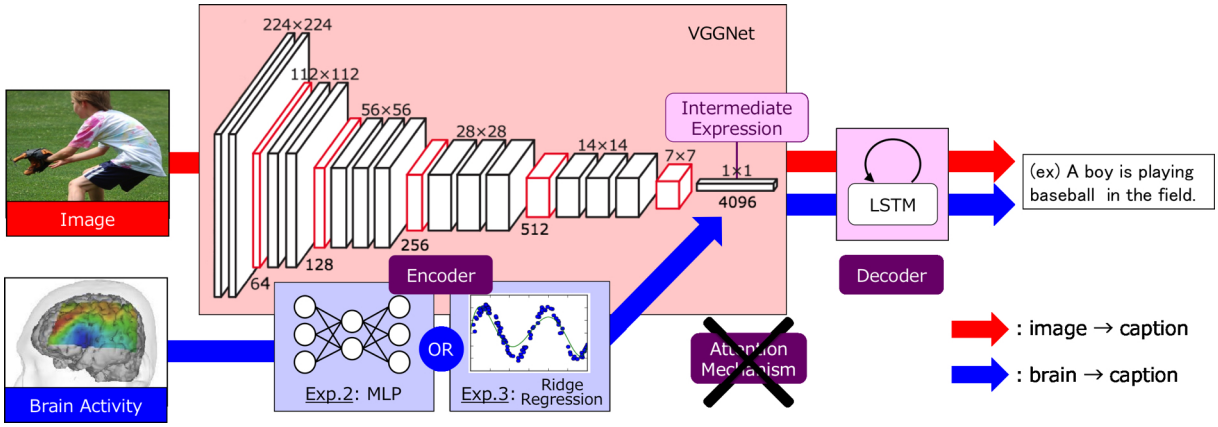

Figure 2: Overview of our approach (Exps.2 and 3: Without Attention Mechanism and with 3-Layered MLP or Ridge regression).

Table 2: Details of the experimental settings.

	Exp.1: Image to Caption with Attention	Exps.2 and 3: Image to Caption without Attention	Exp.1: Brain to Intermediates 3-Layered MLP	Exp.2: Brain to Intermediate 3-Layered MLP	Exp.3: Brain to Intermediate 3 Ridge regression
Dataset	Microsoft COCO		brain activity data		
Hyper-parameters	learning rate : 1.0 (× 0.999) gradient norm threshold : 5 L2-norm : 0.005		learning rate : 0.01 gradient norm threshold : 5 L2-norm : 0.005		L2-norm : 0.5
Learned parameters	Attention & LSTM initialized in [-0.1,0.1]	LSTM initialized in [-0.1,0.1]	weight in 3-Layered MLP initialized in [-0.2,0.2]		parameters in Ridge reg. initialized to 0
Units per layer	196 units	1,000 units	30,662 - 1,000 - 100,352	30,662 - 1,000 - 4,096	-
Vocabulary	Frequent 3,469 words (512-dim vector)		-		
Algorithm	stochastic gradient descent		stochastic gradient descent		ridge regression
Loss function	cross entropy		mean squared error		

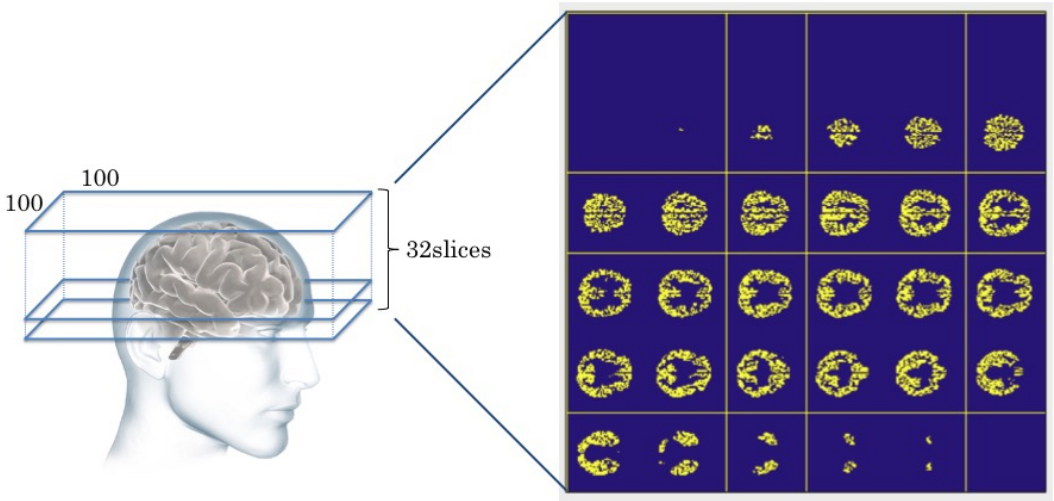

Figure 3: 30,662 voxels observed as the cerebral cortex region.

mean square error in the training process of the 3-layered MLP; the decreasing values of both quantities indicates the training progress.

The generated sentences for Exp.1 are not grammatically correct – they primarily comprise unsuitable meaningless prepositions and do not explain the contexts of the images. As for the evaluation of learning the model, the mean square error did not decrease sufficiently. This is probably caused by the fact that the output dimensions (100,352) are too large compared with the input dimensions (30,662) to learn the corresponding relationships between the brain activity data and the set of intermediate expressions.

Figure 4: Caption generation with Attention Mechanism: target picture (left), generated caption (center), and attention (right).

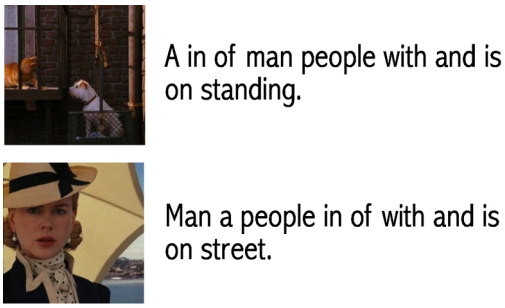

Figure 5: Exp.1:Presented stimuli and the descriptions generated from the evoked brain activity.

Table 3: Exp.1: Training evaluation.

Num. of data	Perplexity	Iteration	MSE
14000	88.67	1	118.32
42000	66.24	5	116.44
84000	60.40	10	114.31
126000	60.10	15	112.36
168000	60.32	16	112.01

4.3 Exp.2: Without Attention Mechanism and with 3-Layered MLP

Using the same procedure as Exp.1, we confirmed that our caption-generation model without Attention Mechanism was well trained by generating captions for two pictures.

Figure 6 shows two sets of still pictures and the generated descriptions for the brain activity data. Table 4 shows the perplexity of the generated image captions and the mean square error of MLP.

Relative to the result of Exp.1, the meaningless prepositions disappear and the generated words seem to depict the image, that is, our model acquires more appropriate expressions, both syntactically and semantically. This is probably because MLP could learn the relationship better by reducing the output dimension from 100,352 to 4,096; we can confirm this by looking at the decrease in the mean square error.

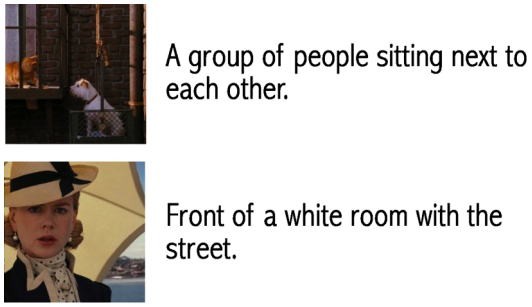

Figure 6: Exp.2:Presented stimuli and the descriptions generated from the evoked brain activity.

Table 4: Exp.2: Training evaluation.

Num. of data	Perplexity	Iteration	MSE
14000	96.50	1	28.95
42000	47.87	5	22.70
84000	47.22	10	17.19
126000	47.37	15	13.37
168000	46.30	20	10.76

4.4 Exp.3: Without Attention Mechanism and with Ridge regression

Figure 7 shows two sets of pictures and descriptions for the brain activity data selected from the test data. The perplexity of the generated captions for the images is the same as in Exp.2, and the mean square error using ridge regression is 8.675.

The generated sentences are syntactically established, that is, prepositions, e.g., “in” and “on,” and articles, e.g., “an,” are precisely used. Compared with the results of Exps.1 and 2, we can see that the grammar of the descriptions has considerably improved. Except for the subjects in the descriptions, the contents of the images are correctly described. In particular, in the descriptions of the second image, an appropriate description of the image is generated, as we see that the person and umbrella are both recognized and their relationship is correctly described. In addition, Exp.3 had the lowest mean square error.

In these three experiments, we confirmed that the methods without Attention Mechanism perform better than that with Attention Mechanism and that ridge regression produces better results than 3-layered perceptron. Therefore, we can conclude that a simple method that can avoid overfitting the data is more appropriate for noisy and small data, such as brain activity data. However, in Exp.2, if we trained the network with more datasets, this result might be changed because we have observed that the mean square error of MLP has been decreasing.


A man is sitting on top of the table.


A man is in the back of an umbrella.

Figure 7: Exp.3:Presented stimuli and the descriptions generated from the evoked brain activity.

5 Conclusions

We proposed a method to generate descriptions of brain activity data by employing a framework to generate captions for still pictures using deep neural networks and by learning the corresponding relationships between the brain activity data and the features of the images extracted by VGGNet. We conducted three experiments to confirm our proposed method. We found that the model without Attention Mechanism using ridge regression performed the best in our experimental settings. In the future, we aim to increase the accuracy of our method to generate captions by revising the parameter settings, using additional training data and introducing evaluation measures, such as BLEU and METEOR. Moreover, we will further consider ways to learn the relationship between brain activity data and the intermediate expression and will introduce Bayesian optimization to optimize the parameter settings.

References

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

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

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*. In EMNLP'14.

K. Cho, A. Courville, Y. Bengio. 2015. *Describing Multimedia Content using Attention based Encoder Decoder Networks*. Multimedia, IEEE Transactions on, 17(11): 1875-1886.

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

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*. In CVPR'15.

D. Elliott and F. Keller. 2013. *Image description using visual dependency representations*. In EMNLP'13.

D. Elliott and A. P. de Vries. 2015. *Describing Images using Inferred Visual Dependency Representations*. In ACL'15.

S. Hochreiter and J. Schmidhuber. 1997 *Long Short-Term Memory*. Neural Computation 9(8).

T. Horikawa, M. Tamaki, Y. Miyawaki, Y. Kamitani. 2013. *Neural Decoding of Visual Imagery During Sleep*. SCIENCE VOL 340.

A. G. Huth, S. Nishimoto, A. T. Vu, 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*. In arXiv preprint arXiv:1502.03167.

R. Kiros, R. Salakhutdinov, and R. Zemel. 2014. *Multimodal neural language models*. In ICML'14.

R. Kiros, R. Salakhutdinov, and R. Zemel. 2015. *Unifying visual-semantic embeddings with multimodal neural language models*. In NIPS'15 Deep Learning Workshop.

G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, T. L. Berg. 2013. *Babytalk: Understanding and generating simple image descriptions*. Pattern Analysis and Machine Intelligence, 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*. In ACL'12.

P. Kuznetsova, V. Ordonez, T. Berg, and Y. Choi. 2014. *TREETALK: Composition and compression of trees for image descriptions*. In ACL'14.

J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, A. Yuille. 2014. *Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)*. In ICLR'14.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean. 2013. *Distributed Representations of Words and Phrases and their Compositionality*. In NIPS'13.

G. A. Miller. 1995. *WordNet: a lexical database for English*. Communications of the ACM, Volume 138, Pages 39-41.

T. M. Mitchell, S. V. Shinkareva, A. Carlson, K. M. Chang, V. L. Malave, R. A. Mason, M. A. Just. 2008. *Predicting Human Brain Activity Associated with the Meanings of Nouns*. Science 320, 1191.

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*. In ACL'12.

S. Nishida, A. G. Huth, J. L. Gallant, 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, J. L. Gallant. 2011. *Reconstructing visual experiences from brain activity evoked by natural movies*. Current Biology, 21(19):1641-1646.

F. Pereira, G. Detre, and M. Botvinick, 2011. *Generating text from functional brain images*. Frontiers in Human Neuroscience, Volume 5, Article 72.

F. Pereiraa, M. Botvinicka, G. Detre, 2013. *Using Wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments*. Artificial Intelligence, Volume 194, January 2013, Pages 240-252.

K. Simonyan, and A. Zisserman. 2015. *Very deep convolutional networks for large-scale image recognition*. In 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, Pages 1025-1034, September 4, 2013, Elsevier Inc.

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

I. Vendrov, R. Kiros, S. Fidler, R. Urtasun. 2016. *Order-Embeddings of Images and Language*. In ICLR'16.

O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. *Show and tell: A neural image caption generator*. In 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*. In ICML'15.

S. Yagcioglu, E. Erdem, A. Erdem, R. Cakici. 2015. *A Distributed Representation Based Query Expansion Approach for Image Captioning*. In ACL'15.

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