

GENERATION OF VIEWED IMAGE CAPTIONS FROM HUMAN BRAIN ACTIVITY VIA UNSUPERVISED TEXT LATENT SPACE

Saya Takada[†], Ren Togo[‡], Takahiro Ogawa^{††} and Miki Haseyama^{†††}

[†] School of Engineering, Hokkaido University, Japan

[‡] Education and Research Center for Mathematical and Data Science, Hokkaido University, Japan

^{†††} Faculty of Information Science and Technology, Hokkaido University, Japan

E-mail: {takada, togo, ogawa}@lmd.ist.hokudai.ac.jp, miki@ist.hokudai.ac.jp

ABSTRACT

Generation of human cognitive contents based on the analysis of functional magnetic resonance imaging (fMRI) data has been actively researched. Cognitive contents such as viewed images can be estimated by analyzing the relationship between fMRI data and semantic information of viewed images. In this paper, we propose a new method generating captions for viewed images from human brain activity via a novel robust regression scheme. Unlike conventional generation methods using image feature representations, the proposed method makes use of more semantic text feature representations, which are more suitable for the caption generation. We construct a text latent space with unlabeled images not used for the training, and the fMRI data are regressed to the text latent space. Besides, we newly make use of unlabeled images not used for the training phase to improve caption generation performance. Finally, the proposed method can generate captions from the fMRI data measured while subjects are viewing images. Experimental results show that the proposed method enables accurate caption generation for viewed images.

Index Terms— Image captioning, deep neural network (DNN), neuroscience, functional magnetic resonance imaging (fMRI).

1. INTRODUCTION

The real world is composed of various objects, and the extraction of their semantic information from the human brain is one of the major themes of neuroscience [1, 2]. Therefore, many researchers have tried to estimate semantic information, *e.g.*, what people see [3–5] and imagine [6, 7]. Several studies have attempted to estimate semantic information using brain activities measured by implantable microelectrode array (MEA) [8], electroencephalography (EEG) [9] and functional magnetic resonance imaging (fMRI) [6, 10]. However, implantable MEA is one of the invasive measurement methods and imposes heavy burdens on people for measuring brain activity. On the other hand, EEG and fMRI belong to non-invasive measurement methods. Particularly, since fMRI data have higher spatial resolution compared to EEG, this becomes one of standard brain signal measurement instruments for the task of estimating semantic information from brain activity [11–15].

It has been reported that different fMRI data are obtained from viewed images whose contents are also different [3]. This suggests that fMRI data include individual visual content information. In fact, since we can express everything around us by language representations, it is considered that the semantic representation space in the

brain is closely related to the semantic information of the language. Hence, recent researches have tried to estimate the semantic information by associating brain activity data with words or sentences representing viewed images or videos [16, 17].

Recent works [16, 17] have achieved an estimation of semantic information of viewed images. For example, Nishida *et al.* [16] tried to acquire the semantic representation space of the language by using a linguistic statistic model called word2vec [18] and showed that word2vec archives higher performance than other linguistic statistic models in modeling brain responses to natural stimuli. Matsuo *et al.* [17] have proposed a method for generating captions for viewed images from fMRI data using an image captioning model [19]. In the image captioning model, the image features are firstly extracted by inputting the image to the deep neural network (DNN) [20]. Then the image captioning model is trained by inputting the image features, and captions corresponding to the images into the network consisted of long short-term memory (LSTM) [21], which is a kind of recurrent neural networks (RNNs) [22]. RNNs make it possible to consider the time series between words. In the LSTM network, captions for viewed images are generated from fMRI data by converting fMRI data to text features after converting to image features.

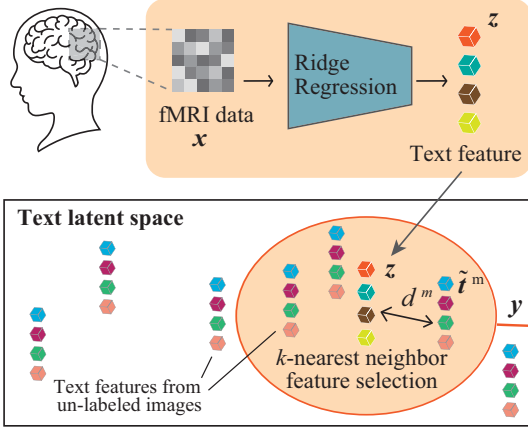
However, there remain two problems in conventional methods. First, although captions are generated by conversion of fMRI data into text features after conversion into image features in conventional methods, there is a possibility that the key information of fMRI data is lost in these two-stage conversions. We think that the use of more suitable features is desirable for the caption generation than image features in conventional methods. Second, it is difficult to collect large-scale data because measurement of fMRI data requires heavy burdens for subjects. Since only pairs of fMRI data and image features included in the training data are used to estimate their relationship in the previously reported methods, it is difficult to generate captions for images in categories not used during the training. By solving these problems, it is expected to improve the applicability of the caption generation from fMRI data in conventional methods in terms of features used in the training and unlabeled data not used in training.

In this paper, we propose a method of caption generation from fMRI data via unsupervised text latent space. We solve the above existing problems based on the following approaches. First, we convert fMRI data into text features corresponding to the intermediate layer of LSTM instead of mediating image features. Then the performance improvement of the caption generation is expected by directly converting fMRI data to more semantic text features contrary to previously reported methods. Second, we newly make use of neighborhood features, where these features are obtained from a large number

This work was partly supported by the MIC/SCOPE #181601001.

Image Caption Generation from fMRI Data

Conversion of fMRI data into text features (2.2)



Text feature transformation with un-labeled images (2.3)

Image Captioning Model (2.1)

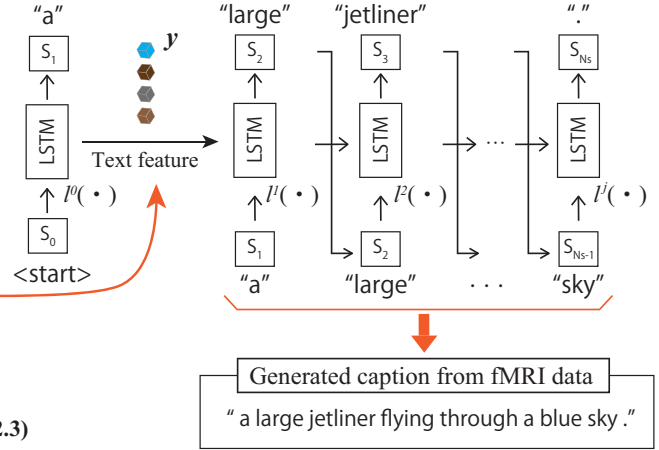


Fig. 1: Overview of our proposed method. Note that construction of the image captioning model and the two novelties of the proposed method are explained in 2.1 - 2.3 respectively.

of unlabeled images that are not used for the training. This approach enables more successful caption generation for images in categories not used in training. We expect that using more linguistic semantic features contributes to the performance improvement of the caption generation. Finally, we summarize our contributions in this paper as follows.

- We use the connection from fMRI data to the corresponding text features from the LSTM cell, which are in the text latent space to convert fMRI data without loss of key information.
- We introduce a new approach utilizing neighboring features calculated from unlabeled images in the feature space to realize successful caption generation for images whose categories are not used in training.

2. GENERATION OF VIEWED IMAGE CAPTIONS FROM HUMAN BRAIN ACTIVITY

The overview of the proposed method is shown in Fig. 1. Our method consists of three phases. In the first phase, we construct an image captioning model composed of a DNN and LSTM cells. In the second phase, we introduce direct regression to the latent text space in the task of image caption generation from fMRI data. In the third phase, we generate captions from fMRI data in unsupervised text latent space. We explain the image captioning model, which generates captions for input images in Subsec. 2.1. Our novel frameworks, the conversion of fMRI data to the text features and the text feature estimation with unlabeled images are introduced in Subsecs. 2.2 and 2.3.

2.1. Construction of Image Captioning Model

First, we construct an image captioning model motivated by [19]. Our image captioning model is composed of the image feature extraction phase and the caption generation phase. We extract the image feature with a pre-trained DNN model. DNNs trained on a large

scale dataset [23] are often used in general object recognition tasks and can extract semantic features suitable for object classification of images. Given an image I^i ($i = 1, 2, \dots, N_c$; N_c being the total number of images used in the training phase), its highly representative image feature $v^i \in \mathbb{R}^{D_v}$ (D_v being the dimension of the features of the intermediate layer of the DNN) can be calculated as follows:

$$v^i = \text{DNN}(I^i). \quad (1)$$

In order to decrease the computing cost of the training, we convert v^i to $v'^i \in \mathbb{R}^{D'_v}$ through the linear combining layer $\mathbf{W}_{\text{linear}}$ as follows:

$$v'^i = \mathbf{W}_{\text{linear}} v^i, \quad (2)$$

where $\mathbf{W}_{\text{linear}}$ is a linear combining layer¹ which can transform the features from the D_v -dimension space into the D'_v -dimension space, where $D'_v < D_v$. We enable decreasing the computing cost by the linear transformation based on $\mathbf{W}_{\text{linear}}$.

Next, we construct the captioning network to generate captions corresponding to the input images. The network consists of multiple LSTM cells $l^j(\cdot)$ ($j = 0, 1, \dots, N_l$; N_l being the number of LSTM cells). This network is trained on the pairs of the features extracted from images v'^i and their corresponding captions. Specifically, in order to input each word in a caption to LSTM cells, we convert them into vectors based on word2vec [18]. The model word2vec can convert words into semantic vectors that can capture the semantic information of a language. The captioning network is constructed by learning the relationship between the image features and the captions based on the LSTM.

Specifically, we define n th word of image I^i as S_n^i ($n = 0, 1, \dots, N_s^i$; N_s^i being the number of words included in a caption). Image features v'^i are input into the first LSTM layer architecture as follows:

$$t^i = l^0(v'^i, \text{word2vec}(S_0)), \quad (3)$$

¹https://pytorch.org/docs/stable/_modules/torch/nn/modules/linear.html

where the starting word S_0 means the beginning of the sentence input to the first LSTM cell $l^0(\cdot)$. The first LSTM cell $l^0(\cdot)$ converts the input image feature v^i into the text feature $t^i \in \mathbb{R}^{D_t}$ (D_t being the dimension of the features of the intermediate layer of the LSTM cell). The obtained text feature t^i is input to the next LSTM $l^1(\cdot)$ and used to generate captions. From the text feature t^i , the network calculates the probability of the next word, and the word whose probability is the highest is decided as the next word S_1^i . In this process, the parameters of LSTM cells are randomly initialized. According to $l^j(\cdot)$, the same process is repeated until the end-of-sentence token appears. In the training of the captioning network, the loss is calculated based on the cross-entropy loss, and the network is optimized by Adam [24]. In this way, the captioning model can generate captions from the target image.

2.2. Conversion of fMRI data into Text Features

We aim to generate natural language sentences describing viewed images from human brain activity. Conventional regression models convert fMRI data into image features, and the estimated image features are transformed into text features via a captioning model. However, we think that this two-stage approach is not suitable for the captioning task. We believe that the direct regression to text latent features will improve the quality of the caption generation.

First, we convert fMRI data $x^l \in \mathbb{R}^{D_f}$ ($l = 1, 2, \dots, N_f$; N_f is the total number of viewed images used in the measurement of fMRI data and D_f being the dimension of the fMRI data) to the semantic text feature $y \in \mathbb{R}^{D_t}$ by performing the regression to the text latent space. Given a viewed image I^l , the text features t^l are extracted from the first LSTM cell $l^0(\cdot)$ of the captioning network as follows:

$$t^l = l^0(v^l, S_0), \quad (4)$$

where $v^l = W_{\text{linear}} v^l$ and $v^l = \text{DNN}(I^l)$. Next, we estimate the relationship between fMRI data x^l and the corresponding text feature t^l based on the ridge regression [25]. The direct regression from fMRI data to text features without losing the key information is one of novel point of our method.

Specifically, in the training phase, we compute the weight vector W and the bias b in the regression function as followed:

$$t^l = W^T x^l + b. \quad (5)$$

These parameters are obtained by simply solving the following optimization problem:

$$\min_{W, b} \sum_{l=1}^{N_f} \|t^l - (W^T x^l + b)\|_2^2 + \alpha \|W\|_2^2. \quad (6)$$

where α is a regularization strength parameter. Note that the model is regularized to avoid over-fitting phenomena by using the L2 norm of the regression parameters. The parameters are calculated by the relationship between features in the text latent space and the fMRI data corresponding to the viewed images based on the ridge regression. In this way, our approach can directly regress the fMRI data to the text features without redundant transformation.

2.3. Text Feature Transformation with unlabeled images

Even if we can construct a model that can estimate answers to questions about viewed images from fMRI data, obtaining large-scale annotated data is still challenging, and a lack of training images can

cause performance degradation. To address this problem, we introduce the utilization of unlabeled images and realize the improving caption generation for viewed images. We generate captions from fMRI data $x_{\text{test}} \in \mathbb{R}^{D_f}$ measured when subjects viewed test images via unsupervised text features. By introduction the approach enhancing robustness, we can obtain more robust features by compensation of the estimated image features in Subsec. 2.2 by using unlabeled images that are not used for training the models.

First, we create the text latent space with unlabeled images. We input unlabeled images \tilde{I}^m ($m = 1, 2, \dots, N_a$; N_a being the number of unlabeled images), which are not used for the measurement of fMRI data and the training of the captioning model, into this model. Then we extract text features $\tilde{t}^m \in \mathbb{R}^{D_t}$ from unlabeled images \tilde{I}^m . Next, we obtain features $z \in \mathbb{R}^{D_t}$ by inputting the test fMRI data x_{test} as follows:

$$z = W^T x_{\text{test}} + b. \quad (7)$$

Then we calculate the Euclidean distance d^m between z and \tilde{t}^m in the text latent space. Based on this distance d^m , k -nearest neighbor features existing in the vicinity of the feature quantity z obtained from the fMRI data x_{test} are selected. Finally, the following new feature y can be calculated by utilizing the above selected features as follows:

$$y = \beta z + \frac{(1 - \beta)}{k} \sum_{m=1}^k \tilde{t}^m. \quad (8)$$

In Eq. (8), the first term represents the original test text feature derived from the fMRI data, and the second term represents the unsupervised learning features. Also, β is a weighted coefficient. Utilizing the features in the neighborhood enables compensation for the estimated image features close to real image features.

Although it is difficult to generate captions for images in categories not used during the training in the conventional methods, the use of the k selected features provides a solution to this problem. By making use of unlabeled data, we can obtain more sophisticated sentence features for the caption generation task. Consequently, by inputting y to the captioning network in Subsec. 2.1, the proposed method can generate the corresponding caption.

3. EXPERIMENTS

In this section, we show the effectiveness of our method. The details of the dataset are explained in Subsec. 3.1. Furthermore, we show evaluation indices and confirm the performance of our method from the experimental results in Subsec. 3.2.

3.1. Experimental Settings

As fMRI data, we used the fMRI dataset published in [6]. In the experiment of [6], a total of 1,200 images from 150 categories, including eight images, were presented to each subject in the training phase. In the test phase, a total of 50 images from 50 categories, including one image, were presented to each subject, where each image was selected from each category. Note that the categories for the test phase were not used for the training phase. Categories attached to these images are corresponding to the ontology of WordNet [26]. Furthermore, fMRI activity was measured once for each training image and 35 times for each test image. The fMRI data were measured with the fMRI equipment (Siemens MAGNETOM Prisma²).

²<https://www.healthcare.siemens.co.jp/magnetic-resonance-imaging/research-systems/magnetom-prisma/>

In this experiment, as unlabeled images for the caption generation, we used 38,532 images that were not included in the viewed images for the test phase in the same 50 categories as the presentation images for the test phase collected from ImageNet. For the evaluation, we used the fMRI data obtained from a subject. The image captioning model was trained with Microsoft Common Objects in Context (MSCOCO) dataset [27]. fMRI decoder was trained by using labeled 1,200 samples of the fMRI data measured when the subject was viewing images in the same manner as [6].

Next, we evaluated the results, *i.e.*, the obtained captions, by comparing captions generated by the test fMRI data and captions directly generated by inputting the viewed images into the image captioning model [19]. For the evaluation, we used the following cosine similarity $w_{c,g}$ between s_c and s_g obtained by inputting captions into Sent2Vec [28]:

$$w_{c,g} = \frac{s_c \cdot s_g}{|s_c| |s_g|}, \quad (9)$$

where s_c and s_g are features obtained from captions generated from the fMRI data and those generated by the image captioning model, respectively. Note that Sent2Vec is an unsupervised method to train and infer sentence embedding. We can extract semantic expression by inputting sentences into Sent2Vec, and the quality of the generated captions is evaluated based on how generated captions are similar to the ground truth. Ideally, the use of manually annotated ground truth is expected for the evaluation. However, this also causes subject biases. Hence, we generate the ground truth captions through the model in [19] to avoid human biases.

In this experiment, we compared the accuracy of the proposed method (PM) using (A) the text latent space regression and (B) the unlabeled images with the following comparative methods to confirm their effectiveness.

- CM1 [17]: image latent space regression without unlabeled images
- CM2: text latent space regression using unlabeled images (*i.e.*, (A) is not adopted)
- CM3: image latent space regression without unlabeled images (*i.e.*, (B) is not adopted)
- CM4: text latent space regression using unlabeled images (*i.e.*, $\beta = 0$ in Eq. (8))

Note that the paper [17] is the state-of-the-art method for the task of caption generation for viewed images.

3.2. Experimental Results

Figure 2 shows examples of generated captions from fMRI data in different category images not included in the training data. We can see that estimated captions are not perfect, but generated captions by our model are reasonable compared to comparative methods including the state-of-the-art method [17].

Table 1 shows the average value of the evaluation metric for 50 test images, where the performances were determined in such a way that each method output the best results. From Table 1, the proposed method outperforms the other comparative methods. Specifically, direct regression to text features can improve the performance of generating captions for viewed images from the fMRI data according to the comparison between CM1 and CM3. In addition, the use of unlabeled images can improve the performance of the caption generation for images whose categories were not included in the training data according to the comparison between CM3 and CM4. The similarity of PM is higher by using direct regression to text features and the





Image	Method	Caption
	PM	a small bird is sitting on a branch.
	CM1	a bird is sitting on a branch in the background.
	CM2	a bird sitting on a branch with a bird on it.
	CM3	a close up of a bird with a bird on its head.
	CM4	a close up of a person holding a banana.
	PM	a bird is sitting on a branch in the water.
	CM1	a bird sitting on a wooden bench outside.
	CM2	a black and white bird sitting on a table.
	CM3	a bird sitting on a wooden table next to a tree.
	CM4	a black and white bird sitting on a table.
	PM	a black and white photo of a black and white photo.
	CM1	a pair of scissors sitting on a table.
	CM2	a close up of a oersin holding a cell phone.
	CM3	a pair of scissors and a pair of scissors.
	CM4	a small brown and white bird sitting on a table.
	PM	a bird sitting on a wooden table with a bird.
	CM1	a black and white dog with a red bow tie.
	CM2	a black and white dog sitting on a wooden table.
	CM3	a small bird is sitting on a bench.
	CM4	a bird sitting on a branch in the grass.

Fig. 2: Samples of generated captions.

Table 1: Cosine similarity between captions directly generated from viewed images and captions generated from fMRI data by each method.

Method	Cosine similarity
PM (best result: $k = 11, \beta = 0.4$)	0.464
CM1 [17]	0.364
CM2 ($k = 11$)	0.405
CM3	0.377
CM4 ($k = 11, \beta = 0.0$)	0.421

utilization of unlabeled images in comparison with the conventional method [17], CM1. It was confirmed that the difference between the similarity of captions via PM and CM1 was statistically significant with $p < 0.05$ by Welch's t-test [29]. These results suggest the effectiveness of using text features from the intermediate layer of LSTM and the use of unlabeled images in the task of the image caption generation from fMRI data.

4. CONCLUSION

In this paper, we have proposed the image caption generation method using fMRI data. We validated the effectiveness of the caption generation using regression in the text latent space and utilizing selected features to the feature space consisting of unlabeled features from unlabeled images. We showed that our method outperformed the state-of-the-art caption generation method.

5. REFERENCES

- [1] A. G. Huth, S. Nishimoto, A. T. Vu *et al.*, “A continuous semantic space describes the representation of thousands of object and action categories across the human brain,” *Neuron*, vol. 76, no. 6, pp. 1210–1224, 2012.
- [2] K. N. Kay, “Principles for models of neural information processing,” *NeuroImage*, vol. 180, pp. 101 – 109, 2018.
- [3] J. V. Haxby, M. I. Gobbini, M. L. Furey *et al.*, “Distributed and overlapping representations of faces and objects in ventral temporal cortex,” *Science*, vol. 293, no. 5539, pp. 2425–2430, 2001.
- [4] D. D. Cox and R. L. Savoy, “Functional magnetic resonance imaging (fMRI) “brain reading”: Detecting and classifying distributed patterns of fMRI activity in human visual cortex,” *NeuroImage*, vol. 19, no. 2, pp. 261–270, 2003.
- [5] D. E. Stansbury, T. Naselaris, and J. L. Gallant, “Natural scene statistics account for the representation of scene categories in human visual cortex,” *Neuron*, vol. 79, no. 5, pp. 1025–1034, 2013.
- [6] T. Horikawa and Y. Kamitani, “Generic decoding of seen and imagined objects using hierarchical visual features,” *Nature communications*, vol. 8, p. 15037, 2017.
- [7] T. Naselaris, C. A. Olman, D. E. Stansbury *et al.*, “A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes,” *Neuroimage*, vol. 105, pp. 215–228, 2015.
- [8] C. R. Ponce, W. Xiao, P. F. Schade *et al.*, “Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences,” *Cell*, vol. 177, no. 4, pp. 999–1009, 2019.
- [9] M. Taghizadeh-Sarabi, M. R. Daliri, and K. S. Niksirat, “Decoding objects of basic categories from electroencephalographic signals using wavelet transform and support vector machines,” *Brain topography*, vol. 28, no. 1, pp. 33–46, 2015.
- [10] Y. Miyawaki, H. Uchida, O. Yamashita *et al.*, “Visual image reconstruction from human brain activity using a combination of multiscale local image decoders,” *Neuron*, vol. 60, no. 5, pp. 915–929, 2008.
- [11] S. A. Engel, G. H. Glover, and B. A. Wandell, “Retinotopic organization in human visual cortex and the spatial precision of functional MRI,” *Cerebral Cortex*, vol. 7, no. 2, pp. 181–192, 1997.
- [12] N. K. Logothetis, “What we can do and what we cannot do with fMRI,” *Nature*, vol. 453, no. 7197, p. 869, 2008.
- [13] J. V. Haxby, A. C. Connolly, and J. S. Guntupalli, “Decoding neural representational spaces using multivariate pattern analysis,” *Annual review of neuroscience*, vol. 37, pp. 435–456, 2014.
- [14] G. Shen, T. Horikawa, K. Majima *et al.*, “Deep image reconstruction from human brain activity,” *PLOS Computational Biology*, vol. 15, no. 1, pp. 1–23, 01 2019.
- [15] Y. Lin, J. Li, and H. Wang, “DCNN-GAN: Reconstructing Realistic Image from fMRI,” *arXiv preprint arXiv:1901.07368*, 2019.
- [16] S. Nishida and S. Nishimoto, “Decoding naturalistic experiences from human brain activity via distributed representations of words,” *Neuroimage*, vol. 180, pp. 232–242, 2018.
- [17] E. Matsuo, I. Kobayashi, S. Nishimoto *et al.*, “Describing semantic representations of brain activity evoked by visual stimuli,” in *Proceedings of 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018, pp. 576–583.
- [18] T. Mikolov, W.-T. Yih, and G. Zweig, “Linguistic regularities in continuous space word representations,” in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 746–751.
- [19] O. Vinyals, A. Toshev, S. Bengio *et al.*, “Show and tell: A neural image caption generator,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.
- [20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proceedings of Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [21] F. A. Gers, J. Schmidhuber, and F. Cummins, “Learning to forget: Continual prediction with LSTM,” *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.
- [22] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.
- [23] J. Deng, W. Dong, R. Socher *et al.*, “ImageNet: A large-scale hierarchical image database,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [24] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [25] A. E. Hoerl and R. W. Kennard, “Ridge regression: Biased estimation for nonorthogonal problems,” *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.
- [26] G. A. Miller, *WordNet: An electronic lexical database*. MIT press, 1998.
- [27] T.-Y. Lin, M. Maire, S. Belongie *et al.*, “Microsoft COCO: Common objects in context,” in *Proceedings of European conference on computer vision*. Springer, 2014, pp. 740–755.
- [28] M. Pagliardini, P. Gupta, and M. Jaggi, “Unsupervised learning of sentence embeddings using compositional n-gram features,” in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 2018.
- [29] B. L. Welch, “The generalization of ‘student’s’ problem when several different population variances are involved,” *Biometrika*, vol. 34, no. 1/2, pp. 28–35, 1947.