Vol.2018-DBS-167 No.17
Vol.2018-IFAT-132 No.17
2018/9/13

A Comparative Study of Deep Learning Approaches for Visual Question Classification in Community QA

Hsin-Wen Liu[1,a)] Avikalp Srivastava[2,b)] Sumio Fujita[3,c)] Toru Shimizu[3,d)] Riku Togashi[3,e)]
Tetsuya Sakai[1,f)]

Abstract: Tasks that take not only text but also image as inputs, such as Visual Question Answering (VQA), have received growing attention and become an active research field in recent years. In this study, we consider the task of Visual Question Classification (VQC), where a given question containing both text and an image needs to be classified into one of predefined categories for a Community Question Answering (CQA) site. Our experiments use real data from a major Japanese CQA site called Yahoo Chiebukuro. To our knowledge, our work is the first to systematically compare different deep learning approaches on VQC tasks for CQA. Our study shows that the model that uses HieText for text representation, ResNet50 for image representation, and Multimodal Compact Bilinear pooling for combining the two representations achieved the highest performance in the VQC task.

Keywords: Comparative Study, Deep Neural Networks, Question Categorization, Visual Question Classification

1. Introduction

Deep learning approaches, such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), have been widely used in classification tasks in the field of computer vision and natural language processing due to promising results [9][11] [12] in the past few years. Recently, tasks that combine vision with language and reasoning, such as Visual Question Answering (VQA) [1] [16], which take both text and image as inputs, have expanded rapidly and become a popular research field. This study considers the task of Visual Question Classification (VQC), where the goal is to classify a given question containing both text and image to one of predefined categories from a major Japanese Community Question Answering (CQA) site called Yahoo Chiebukuro (i.e., a Japanese equivalent of Yahoo! Answers).

CQA sites offer services that allow users to post questions and those posted questions are generally organized into categories. Our experiments used real data posted from 2013 to 2014 in Yahoo Chiebukuro. As the VQA task takes both text and image as inputs, we extracted a subset of the questions that contain an image for our experiments. Figure 1 shows an example of a posted question containing both text and image in Yahoo Chiebukuro site. There are approximately 5% of the posted questions containing both text and image from 2013 to 2014 in the Yahoo Chiebukuro corpus. In our study, we analyze the VQC task from three aspects: text representation, image representation, and a method for combining for the two representations. Our study shows that the model that uses HieText for text representation, ResNet50 for image representation, and Multimodal Compact Bilinear pooling for combining the two representations achieved the highest performance in the VQC task.

Fig. 1 An example of a question containing both text and image posted in Yahoo Chiebukuro.

2. Related Work

Several recent papers have begun to explore tasks that consider both vision and language as inputs. One famous example is VQA [1][6][16], where given an image with a question about the im-

1 Waseda University, 341, Shinokubo, Tokyo, 1690072, Japan
2 Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA
3 Yahoo Japan Corporation, 13, Kioityo, Tokyo, 1028282, Japan
a) stephanie1125@toki.waseda.jp
b) avikalps@andrew.cmu.edu
c) sufujita@yahoo-corp.jp
d) toshimiz@yahoo-corp.jp
e) rtogashi@yahoo-corp.jp
f) tetsuyasakai@acm.org

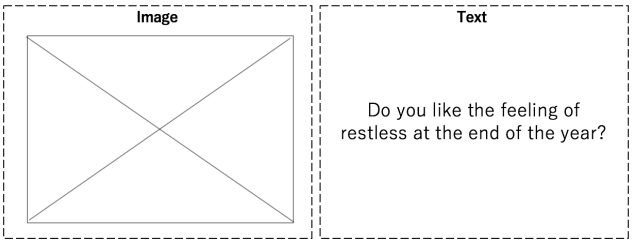


Fig. 2 Real example from Yahoo Chiebukuro dataset, where the posted text is not related to the posted image. (An image of Ultra-Man. Due to copyright reasons, we can not show the image.)

age, VQA task is to provide an answer to the question. There are also other examples such as image captioning [4][7], where the task is to automatically describe the content of an image, and visual grounding [18][3], where the task is to localize an object in an image by a textual query phrase. These task are related to our VQC task as these tasks combine vision with language. We are not aware of any existing studies on VQC except for the work by Tamaki *et al.* [8], who explored different methods for combining text and image features. In contrast to their work, we consider not only the combination methods but also how the two representations effect the performance of our experimental models in the VQC task.

Deep learning approaches are widely used in tasks that involved with both vision and languages [13][4][3]. CNN has proven to be very effective and has shown its power for learning image features [10][5]. ResNet [5], a pre-trained CNN network from large-scale image corpus, is widely used to extract features from images in VQA task [13][3][14]. Hence, we also apply ResNet50 [5] to our VQC task as one of the image representations and compare its performance with a 3-layers CNN.

RNN, especially Long Short Term Memory (LSTM), is widely used to represent sentences or phrases in VQA and visual grounding tasks [18][3]. To compare the performance between LSTM and CNN, we apply both CNN and LSTM as text representations in our VQC task. In VQA task, Lu *et al.* [13] introduce a hierarchical architecture question representation, which applied word embedding, 1-dimensional CNN and LSTM encoding. We concatenate outputs of word embedding, 1-dimensional CNN, and LSTM from their question representation and use it as one of the text representations in our VQC task. This text representation outperforms the other two text representations and is the best text representation in our experiments.

In VQA and image grounding tasks, Fukui *et al.* [3] introduce MCB, a method of combining image and text representations by randomly projecting the image and text representations to a higher dimensional space and convolve both vectors using element-wise product. Tamaki *et al.* [8] applied a method called SP for combining image and text representations, which simply concatenates the sum and element-wise product of the two representations. Our experiments explore both MCB and SP.

Although tasks that involve both visual and language share similar network structures, there is one big difference between the VQC task and other VQA, image captioning, and image grounding tasks. Different from other tasks, VQC tasks evaluate models using real data posted in a CQA site. Questions from CQA tend to be longer, and more diverse in terms of quality, than those from other tasks. For example, some users may post a question with an image that are not related to the question. Figure 2 shows a real example posted in Yahoo Chiebukuro where the posted text is not related to the posted image. In contrast, text and image from VQA, image captioning, and image grounding tasks are highly related to each other because these tasks require the understanding of reasoning from both image and text.

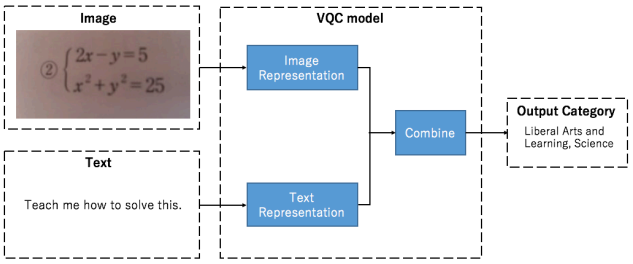


Fig. 3 Given both text and image (real example from Yahoo Chiebukuro dataset) as inputs, the task of VQC is to classified the text and image into one of the predefined categories.

3. Experiments

This section gives an introduction of our task, experimental models, dataset, and some setting in the experiments.

3.1 Task and models

In this study, we consider the task of VQC. Given a question containing both text and an image, the task is to provide a category for the question from one of predefined categories. The VQC task in our experiment is 14 categories VQC task, where questions need to be classified into 14 top-level categories obtained from a major Japanese CQA site called Yahoo Chiebukuro. An example of a VQC task is shown in Figure 3. Refer to Figure 3, the network structure of VQC model can be divided into three part: text representation, image representation, and combination methods. As a result, we prepared our experiments and models based on the following three aspects:

(1) **Text Representation** 3CNN, LSTN, and HieText are prepared as text representations in our experiments. We freeze the choice of image representation and combination method to 5CNN and SP for our VQC models.

3CNN Text representation from the output of a 3-layers CNN model, which contains randomly initialized word embedding, 2-dimensional convolution, and a dropout applied in the fully connected layer;

LSTM Text representation from an LSTM model's last hidden layer features, which contains randomly initialized word embedding, LSTM units, and a dropout applied in the fully connected layer;

HieText Concatenation of the 3 levels language representations: word level, phase level, and sentence level from [13], which contains randomly initialized word embedding, 1-dimensional convolution, max pooling across different n-grams at each word, LSTM encoding,

and a dropout applied in the fully connected layer;

(2) **Image Representation** 5CNN and ResNet50 are prepared as image representations in our experiments. We freeze the choice of text representation and combination method to 3CNN and SP for our VQC models.

5CNN Image representation from a 5-layer CNN model, which contains 2-dimensional convolution, batch normalization, and max pooling;

ResNet50 Image representation from the last feature maps with input shape of (7, 7, 2048) of a pre-trained ResNet-50 model [5];

(3) **Combination methods** MCB and SP are prepared as combination methods in our experiments. We freeze the choice of text and image representations to 3CNN and 5CNN for our VQC models.

SP Concatenate the sum and element-wise product of text and image features and pass it to the fully connected layer. This method was introduced and used in [8];

MCB Generate a joint image and text representation (1,024*1,024 → 2,048) using MCB [3] and pass it to the fully connected layer;

3.2 Dataset

We evaluate our models using a subset of the Yahoo Chiebukuro corpus from 2013 to 2014. The subset contains 1,018,833 questions. Each question contains a text-image pair, along with corresponding categories assigned by human judges. In our experiment, we use the 14 major top-level categories shown in Table 1 and divide the subset into 3 splits with the 8:1:1 ratio: train, validation, and test, each of which contains 815,066, 101,884, and 101,883 questions, respectively.

Table 1 14 major top-level categories (originally in Japanese).

Label	Category Name	# of questions
0	Entertainment and Hobbies	279069
1	Liberal Arts and Learning, Science	171986
2	Health, beauty and fashion	130465
3	Life and living guide	126664
4	Internet, PC and home appliances	114200
5	Sports, outdoor, cars	69613
6	Life and romance, worries of human relations	39629
7	News, politics, international situation	24669
8	Region, Travel, Outing	23656
9	Parenting and school	13909
10	Manners, ceremonial occasions	11173
11	Computer technology	5269
12	Business, economy and money	4833
13	Occupation and career	3698

3.3 Dataset imbalance problem

Table 1 shown that there is a class imbalance problem in our evaluation dataset, that is, the size of 'Entertainment and Hobbies' category is ten times larger than half of all categories. The 'Entertainment and Hobbies' category is a major class in our dataset, where it has more number of data in the dataset, while half of other categories are minor classes due to their relatively less number of data [2]. This will lead to a problem that VQC models are more focusing on classifying data from the major class while ignoring or misclassifying data from minor classes. It is also possible that VQC models predict most data as major class and ignore other minor classes, which will lead to poor classification rates on minor classes.

We tackle this problem by adding some class weights in the cost function of our models during training. For example, for a question x and the corresponding model output y, the cross entropy loss L can be calculated in Equation 1, where j is the correct category, w_j is the category weight for the category j and N_j is the number of questions in category j.

$$\begin{aligned} L &= -w_j \log y_j \\ w_j &\propto \frac{1}{N_j} \end{aligned} \tag{1}$$

Hence, misclassification of minor classes will have higher weights than the major class, that is, errors are considered more costly in minor classes than the major class. This will prevent the biases towards the major class in our VQC models.

3.4 Experimental Setting

For all experimental models, we set the randomly initialized word embedding dimension to be 256 and the input text length to be 160 in word-level, which covered 90% of the questions used in our experiment. As for the image representation, We use the input image size 224*224 for both 5CNN and pre-trained ResNet50 [5]. For training, we use a mini-batch size of 128.

4. Results and Analysis

This section report our experimental results and analyze reasons behind those results.

4.1 Experimental results

Table 2 shows the macro-averaged F1 scores for different experimental models on our VQC test set. Equation 2 shows the formula for computing macro-averaged F1 scores, where N is the number of categories, TP_i, FP_i, FP_i is True Positive, False Positive, and False Negative respectively for each category i.

$$\begin{aligned} Precision_i &= \frac{TP_i}{TP_i + FP_i} \\ Recall_i &= \frac{TP_i}{TP_i + FN_i} \\ F1_i &= 2\frac{Precision_i * Recall_i}{Precision_i + Recall_i} = \frac{2TP_i}{2TP_i + FP_i + FN_i} \\ F1(Macro-averaged) &= \frac{1}{N}\sum_{i=1}^{N} F1_i \end{aligned} \tag{2}$$

It can be observed that for different text representations: 3CNN, LSTM, HieText, with the same image representation (5CNN) and combination methods (SP), the performance in terms of macro-averaged F1 scores improves 17.1% and 17.5% when we replace text representation from 3CNN to LSTM and HieText, respectively.

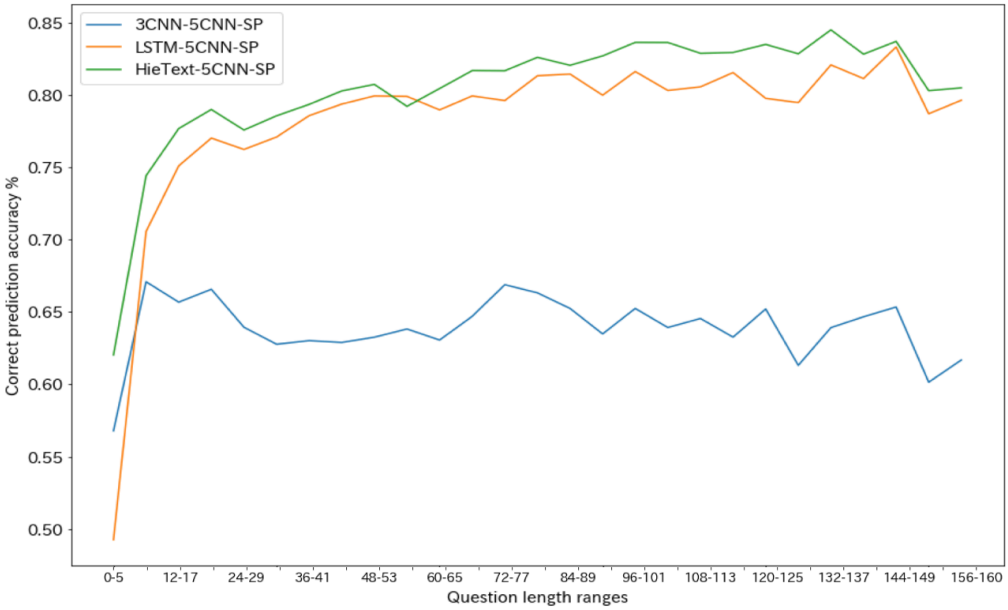


Fig. 4 Distributions of correct prediction accuracy for different text representation models over different question length ranges.

Table 2 Experimental results on VQC test set.

Models (Text-Image-Combine)	macro-averaged F1 scores (%)
3CNN-5CNN-SP	50.2
LSTM-5CNN-SP	67.1
HieText-5CNN-SP	67.5
3CNN-ResNet50-SP	51.1
3CNN-5CNN-MCB	53.1
HieText-ResNet50-MCB	**68.2**

For image representations, we compare the performance of 5CNN with a pre-trained ResNet50 using the same text representation (3CNN) and combination methods (SP). The performance in terms of macro-averaged F1 scores improves 0.9% when we replace image representation from 5CNN to a pre-trained ResNet50. For two combination methods used in our experiments with the same text representation (3CNN) and image representation (5CNN), MCB improve the performance in terms of macro-averaged F1 scores by 2.9% compare to SP.

Furthermore, we choose the best text representation (HieText) along with the best image representation (ResNet50) with the best combination methods (MCB) in our experiments, we can further improve the performance in terms of macro-averaged F1 scores of our baseline model (3CNN-5CNN-SP) by 18%.

Table 3 shows the Tukey HSD (Honestly Significant Differences) p-values and effect sizes (i.e., standardised mean differences) based on two-way ANOVA (without replication) [15]. It can be observed that our best model (HieText-ResNet50-MCB) statistically significantly outperforms 3CNN-5CNN-SP, 3CNN-ResNet50-SP, and 3CNN-5CNN-MCB.

4.2 Text length analysis

Figure 4 shows the correct prediction accuracy with different question length ranges for different text representations with the same image representation and combination methods. For text representation, we found that 3CNN and LSTM are comparable when question lengths are small, e.g., <10, then LSTM gets increasing advantage over 3CNN when meet longer questions. Moreover, our best model, HieText-5CNN-SP outperforms 3CNN and LSTM for both short and long questions.

4.3 Case studies

We also did some case studies for comparing CNN with LSTM. Table 4 shows two examples from Yahoo Chiebukuro. One is an example in which 3CNN predicts correctly while LSTM predicts incorrectly. The other is an example in which LSTM predicts correctly while 3CNN predicts incorrectly. Generally speaking, CNNs are considered good at extracting keywords in text, while LSTM tend to classify text based on the entire sentence [17]. It can be observed that our 3CNN-5CNN-SP model extracts keywords such as “car”, “window”, or “water” and it predicts “Sports, outdoor, car” category. On the other hand, LSTM predict the first example wrong as category “Life and living guide”. One possible reason for this incorrect prediction is that LSTM seems to capture feature from “Please lend me your wisdom” in the first example as we use the last hidden state from LSTM to represent the question. In the second example, one possible reason that CNN predicts incorrectly is that it capture some keywords such as “listening”, “people”, “laughing”, while LSTM predict the correct category based one the entire sentence.

4.4 Confusion matrix heatmap analysis

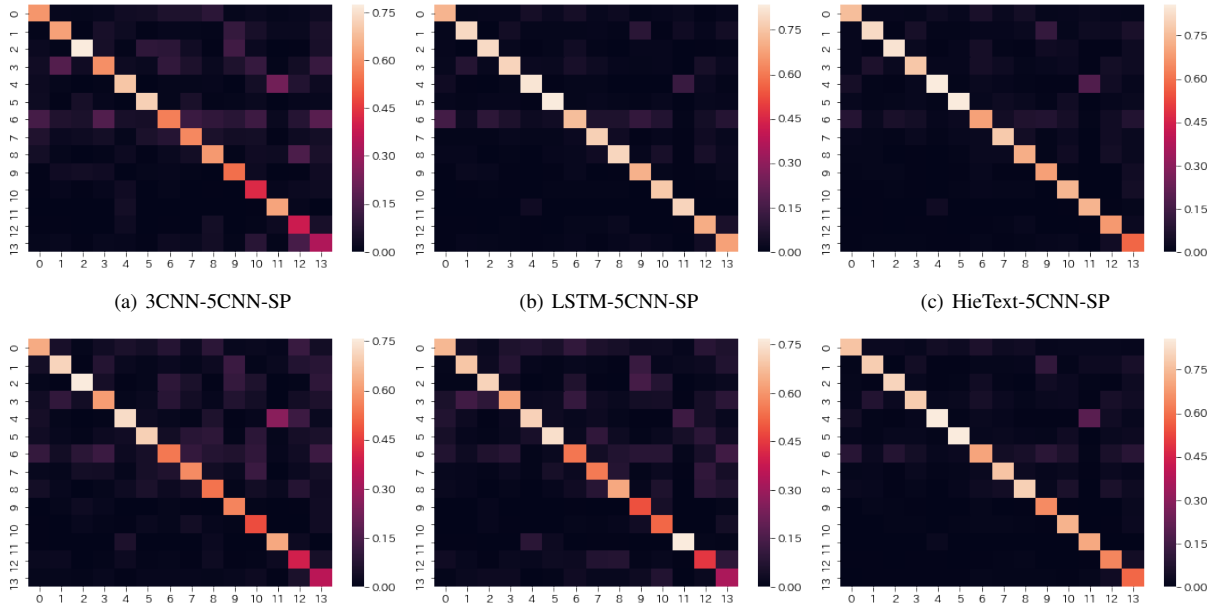
Figure 5 shows the confusion matrix heatmap our experimental models. It can be observed that our experimental models tends to misclassify questions from Category 11 (“Computer technology”) into Category 4 (“Internet, PC and home appliance”). The above tendency is because the questions posted in “Computer technology” is similar to the questions posted in “Internet, PC and home appliance”. In fact, it is probably difficult for humans to judge whether a given questions belongs to one of these categories and not the other. Figure 6 shows a real example

Table 3 Tukey HSD p-values and effect sizes (i.e., standardised mean differences).

	LSTM-5CNN-SP	HieText-5CNN-SP	3CNN-ResNet50-SP	3CNN-5CNN-MCB	HieText-ResNet50-MCB
3CNN-5CNN-SP	p = 0.0000 ES = 1.8354	p = 0.0000 ES = 1.9411	p = 0.6323 ES = 0.1309	p = 0.2906 ES = 0.2944	p = 0.0000 ES = 2.0627
LSTM-5CNN-SP	-	p = 0.9474 ES = 0.0763	-	-	p = 0.5608 ES = 0.2802
HieText-5CNN-SP	-	-	-	-	p = 0.7519 ES = 0.2137
3CNN-ResNet50-SP	-	-	-	-	p = 0.0000 ES = 2.1122
3CNN-5CNN-MCB	-	-	-	-	p = 0.0000 ES = 1.6152

Table 4 Real example from Yahoo Chiebukuro dataset.

Gold	3CNN-5CNN-SP	LSTM-5CNN-SP	Posted text (originally in Japanese).
Sports, outdoor, cars	Sports, outdoor, cars	Life and living guide	When I wash my car, I forgot to roll up the window and the window got stained with water and it could not be removed even if you wiped it. How to remove water stains from the window? Please lend me your wisdom.
Liberal Arts and Learning, Science	Life and living guide	Liberal Arts and Learning, Science	If you are confident in English listening, please let me know. People start laughing around 2:33 in this video. Why is that? Please let me know as much as possible.



(a) 3CNN-5CNN-SP (b) LSTM-5CNN-SP (c) HieText-5CNN-SP

(d) 3CNN-ResNet50-SP (e) 3CNN-5CNN-MCB (f) HieText-ResNet50-MCB

Fig. 5 Confusion matrix heatmap.

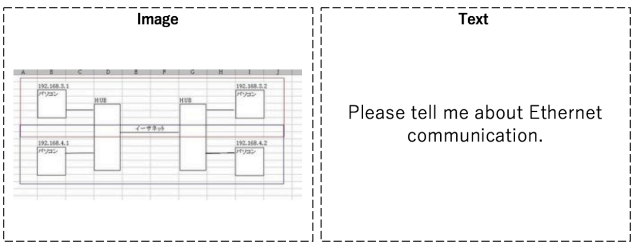


Fig. 6 Real example posted in Yahoo Chiebukuro, where "Computer Technology" is assigned by human judges while our best model (HieText-ResNet50-MCB) classified it as "Internet, PC and home appliances".

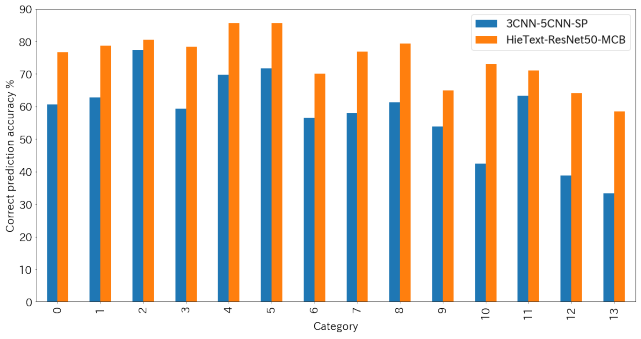


Fig. 7 Correct prediction accuracy of 3CNN-5CNN-SP and HieText-ResNet50-MCB for different categories.

posted in Yahoo Chiebukuro dataset, where "Computer Technology" is assigned by human judges while our best model (HieText-ResNet50-MCB) misclassified it as "Internet, PC and home appliances". Different people may have different views about the category for this question. From our point of view, "Internet, PC and home appliances" seems to be more appropriate for the question, as the question in Figure 6 is asking information about Ethernet communication, which is more related to Internet.

Although there are problems in "Computer technology" category and "Internet, PC and home appliances" category. Figure 7 shows that our best model (HieText-ResNet50-MCB) improves the correct prediction accuracy in all categories compared with our baseline (3CNN-5CNN-SP) model.

5. Conclusions

Our experiments compare different deep learning approaches from text representation, image representation, and combination methods. To our knowledge, our work is the first to compare models based on the above three aspects in VQC tasks. Our findings from the above three aspects can be summarized as follows:

- For text representation, we evaluate three different text representations, 3CNN, LSTM, HieText. HieText outperforms the other two on average, although the difference between HieText and LSTM is not statistically significant. Moreover, the choice of text representation has the largest impact on the overall VQC performance.
- For image representation, we evaluate two different image representations, 5CNN and ResNet50 [5]. While ResNet50 slightly outperforms 5CNN on average, the difference is not statistically significant. Moreover, the choice of image representation on the overall VQC performance seems small.
- We evaluate two methods for combining text and image representations, Multimodal Compact Bilinear (MCB) pooling [3] and Sum and element-wise Product (SP) concatenation [8]. While MCB outperforms SP, the difference is not statistically significant.
- Among the approaches that we explored, the model that used HieText or text representation, ResNet50 for image representation, and MCB for combining the two representations achieved the highest performance.

Acknowledgments This study was undertaken with the supports from Yahoo Japan Corporation.

References

[1] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C. and Parikh, D.: VQA: Visual question answering, *Proceedings of the IEEE ICCV 2015*, pp. 2425–2433 (2015).

[2] Chawla, N. V., Japkowicz, N. and Kotcz, A.: Editorial: special issue on learning from imbalanced data sets, *SIGKDD Explorations*, Vol. 6, pp. 1–6 (2004).

[3] Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T. and Rohrbach, M.: Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding, *EMNLP* (2016).

[4] Google, O. V.: Show and Tell : A Neural Image Caption Generator (2014).

[5] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (2016).

[6] Kafle, K. and Kanan, C.: Visual question answering: Datasets, algorithms, and future challenges, *Computer Vision and Image Understanding*, Vol. 163, pp. 3–20 (2017).

[7] Karpathy, A. and Fei-Fei, L.: Deep Visual-Semantic Alignments for Generating Image Descriptions, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3128–3137 (2015).

[8] Kenta TAMAKI, Riku TOGASHI, S. F. S. K. H. M. and SAKAI, T.: Classifying Community QA Questions That Contain an Image, *ICTIR* (2018).

[9] Kim, Y.: Convolutional Neural Networks for Sentence Classification, *EMNLP* (2014).

[10] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, *Advances in Neural Information Processing Systems 25* (Pereira, F., Burges, C. J. C., Bottou, L. and Weinberger, K. Q., eds.), Curran Associates, Inc., pp. 1097–1105 (online), available from ⟨http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf⟩ (2012).

[11] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, *Commun. ACM*, Vol. 60, No. 6, pp. 84–90 (2017).

[12] Lee, J. Y. and Dernoncourt, F.: Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks, *HLT-NAACL* (2016).

[13] Lu, J., Yang, J., Batra, D. and Parikh, D.: Hierarchical Question-Image Co-Attention for Visual Question Answering, *NIPS* (2016).

[14] Saito, K., Shin, A., Ushiku, Y. and Harada, T.: DualNet: Domain-invariant network for visual question answering, *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 829–834 (2017).

[15] Sakai, T.: Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power, *Springer* (2018).

[16] Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A. R. and van den Hengel, A.: Visual question answering: A survey of methods and datasets, *Computer Vision and Image Understanding*, Vol. 163, pp. 21–40 (2017).

[17] Yin, W., Kann, K., Yu, M. and Schütze, H.: Comparative Study of CNN and RNN for Natural Language Processing, *CoRR*, Vol. abs/1702.01923 (2017).

[18] Zhu, Y., Groth, O., Bernstein, M. S. and Fei-Fei, L.: Visual7W: Grounded Question Answering in Images, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4995–5004 (2016).