DEIM Forum 2018 E7-3

Classifying Community QA Questions That Contain an Image

Kenta TAMAKI†, Riku TOGASHI††, Sumio FUJITA††, Sosuke KATO†, Hideyuki MAEDA††, and Tetsuya SAKAI†

† Waseda University,
3–4–1, Shinokubo, Tokyo, 169–0072, Japan
†† Yahoo Japan Corporation,
1–3, Kioityo, Tokyo, 102–8282, Japan
E-mail: †madao@akane.waseda.jp, sow@suou.waseda.jp, tetsuyasakai@acm.org,
††{rtogashi,sufujita,hidmaeda}@yahoo-corp.jp

Abstract We consider the problem of automatically assigning a category to a given question posted to a Community Question Answering (CQA) site, where the question contains not only text but also an image. For example, CQA users may post a photograph of a dress and ask the community "Is this appropriate for a wedding?" where the appropriate category for this question might be "Manners, Ceremonial occasions." We tackle this problem using Convolutional Neural Networks with a DualNet architecture for combining the image and text representations. Our experiments with real data from Yahoo Chiebukuro and crowdsourced gold-standard categories show that the DualNet approach outperforms a text-only baseline ($p = .0000$), a sum-and-product baseline ($p = .0000$), Multimodal Compact Bilinear pooling ($p = .0000$), and a combination of sum-and-product and MCB ($p = .0000$). where the p-values are based on a Tukey Honestly Significant Difference test with $B = 5000$ trials.

Key words Classification, Convolutional Neural Network, Community Question Answering

1. Introduction

In Community Question Answering (CQA) sites, posted questions are organised by *category* so that users can find their desired questions easily. Each CQA site typically has its own hierarchy of mutually exclusive categories, where the top-level categories might be "Entertainment and Hobbies," "News, Politics, International affairs," and so on. In *Yahoo Chiebukuro*(注1), the most widely-used CQA site in Japan, when a questioner posts her question, the site presents her with an automatically selected list of possible categories, from which she can select one and tag her own question with it. Our goal is to automate the problem of assigning a top-level category to a given question, where the question contains not only text but also an image. For example, CQA users may post a photograph of a dress and ask the community "Is this appropriate for a wedding?" where the appropriate category for this question might be "Manners, Ceremonial Occasions." Clearly, this task is more challenging than classifying purely textual questions, and is practically important: in the real CQA data used in our study, approximately 10.2% of the questions actually contain an image.

We tackle the aforementioned classification problem using Convolutional Neural Networks (CNNs) [5] with a DualNet architecture for combining the image and text representations. Using real data from Yahoo Chiebukuro, we conduct a large-scale question classification evaluation where the categories actually assigned by the questioners are considered to be the gold standard, and a smaller-scale experiment where the gold standard is constructed based on the views of crowd workers. The latter experiment shows that the DualNet approach outperforms a text-only baseline ($p = .0000$), a sum-and-product baseline ($p = .0000$), Multimodal Compact Bilinear pooling ($p = .0000$), and a combination of sum-and-product and MCB ($p = .0000$). where the p-values are based on a Tukey Honestly Significant Difference test with $B = 5000$ trials.

2. Related Work

2.1 Back Ground

Handling information from multiple modalities has been

(注1) : https://chiebukuro.yahoo.co.jp/

studied in various fields. Among these studies, *Visual Question Answering* (VQA) [1], [2], *visual grounding* [2], [7], and *image captioning* [6] are closely related to our task in that these tasks all involve handling of both image and text. In VQA, the system is given an image and a question about that image, and is required to output an answer in natural language. In visual grounding, the system is given an image and a natural language description, and is required to return a bounding box within that image that corresponds to the description. In image captioning, the system is given an image, and is required to output a natural language description of that image.

Of the above, VQA and visual grounding, require the understanding of both image and text inputs. However, note that the input text in these tasks concerns whatever is featured *within* the input image: for example, an input text for visual grounding may be "(locate) a small white dog (within the picture)" [7]. In contrast, our question classification task involves question texts that generally provide a context *outside* the input image: for example, in the aforementioned example with a photograph of a dress, the accompanying question "Is this appropriate for a wedding?" does not describe any feature within that photograph; rather, it *complements* the information conveyed in the image, while referring to the dress with the demonstrative pronoun "this." In short, our classification task is different from the aforementioned tasks that involve both text and images.

2.2 Multimodal Compact Bilinear pooling

In the context of VQA and visual grounding, Fukui et al. [2] proposed *Multimodal Compact Bilinear* (MCB) pooling for joint representation of image and text. The *Compact Bilinear* pooling model [3] is a technique to compress the high-dimension output of a traditional bilinear pooling model [10]; MCB is a multimodal version of this technique. Fukui et al. [2] argue that MCB complements basic operations such as vector concatenation. In the present study, we apply MCB to the problem of question classification where each question contains text and an image. Moreover, we propose a combination of MCB with the simple sum and element-wise product approach.

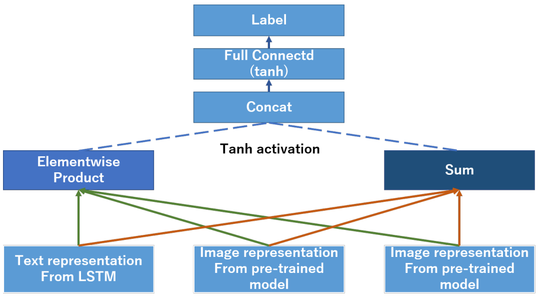
Label
Full Connectd (tanh)
Concat
Tanh activation
Elementwise Product
Sum
Text representation From LSTM
Image representation From pre-trained model
Image representation From pre-trained model

Figure 1 Image of DualNet for real images category

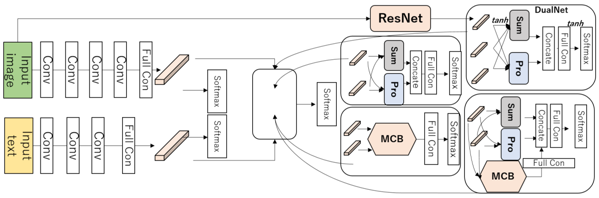

Figure 2 Proposed CNN architecture.

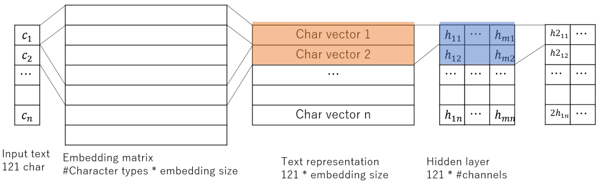

Figure 3 Character level convolution for the text classification network.

2.3 DualNet

DualNet [8] was proposed in 2016 as a method of VQA, and it outperfomed the state of the art in VQA Challenge 2016. Figure 1 depicts the DualNet method in the real images category. In this method, image and text information are combined by fusing the last hidden layer of some pre-trained models. For the real images task, they obtained the text representation from the LSTM's last hidden layer, and the image representation from the hidden layer of multiple pre-trained models (such as *ResNet* [4]). In the present study, we propose to use a model pre-trained with images actually posted to Yahoo! Chiebukuro, along with a pre-trained *ResNet* model.

3. Proposed Classification Methods

Figure 2 shows our proposed CNN architecture for question classification. The image classification network (top left) is a simple 5-layer CNN, where the input image size is 128*128, with a mini batch size of 64. Each convolution layer applies convolution, batch normalization, ReLU activation, and max pooling. Dropout is applied in the fully connected layer.

The text classification network (bottom left) is a character-level CNN [11]. Figure 3 depicts our character-level convolution method. The input text size was set to 121, which was the average number of characters in our training data. Questions whose text part was longer than 121 characters were excluded from our experiment. The vocabulary size (i.e., the number of distinct characters) of the training data was 5,206; accordingly, we set the size of the embedding matrix to 5,208*embedding-size (namely, 200), after adding "white space" and "unknown" to the vocabulary. Questions in the validation and test data that were shorter than 121 characters were stuffed with white spaces; Characters that never appeared in the training data were treated as "unknown." Our text classification network first prepares an embedding matrix that consists of distributed representations corresponding to each character that can be input. Next, we compute a convolution with a 200*height-size*#channel kernel to the matrix obtained by concatenating the distributed representations of the input characters.

Our final network (Figure 2 middle) utilises pre-trained models of both of the aforementioned networks and pre-trained *ResNet*[4] network. Using the hidden layers from the image and text networks, we consider the following four methods for combining the image and text representations (Figure 2 right):

SP Calculate the Sum and element-wise Product; concatenate the results and pass it to the fully connected layer;

MCB Use MCB to generate a joint image and text representation (256*256 → 2,048); then pass the joint representation to the fully connected layer;

SP+MCB Combination of the above two. As shown in Figure 2, the MCB representation goes through a fully connected layer and then is concatenated with the Sum and element-wise Product. The result of concatenation is then passed to another fully connected layer.

DualNet Calculate the sum and element-wise Product of text representation, image representation and the hidden layer of a pre-trained *ResNet* model with tanh activation, and concatenate the results. The processing of this method is shown below (I_1 = image representation from our image classification network, I_2 = image representation from *ResNet*, Q = text representation from our text classification network).

$$I'_1 = \tanh(W_{I1}\ I_1) \tag{1}$$

$$I'_2 = \tanh(W_{I2}\ I_2) \tag{2}$$

$$Q' = \tanh(W_Q\ Q) \tag{3}$$

$$F_S = I'_1 + I'_2 + Q' \tag{4}$$

$$F_P = I'_1 \circ I'_2 \circ Q' \tag{5}$$

$$F = Concat(F_S, F_P) \tag{6}$$

4. CQA and Ground Truth Data

To evaluate our question classification methods, we used real data from Yahoo Chiebukuro. Table 1 shows some statistics of our data: of the approximately 11M questions, approximately 1.12M (10.2%) contains an image; from this set, we extracted 693,519 image-attached questions that cover the ten major top-level categories shown in Table 2. For our experiments, this set was divided into training, test, and validation data sets with the 8:1:1 ratio, and as a result, we obtained 69,355 test questions. By treating the category that is already attached to each test question as the gold standard, we can compute the *classification accuracy* for each of our methods. Note that, since our problem setting considers ten top-level categories, a random system would only achieve a classification accuracy of 10% on average.

Table 1 Number of questions in our CQA data.

all questions	11,074,960
questions with an image	1,128,167
questions with an image used in the experiment	693,519
training questions	554,777
test questions	69,355
test questions with crowdsoured category labels	5,190
validation questions	69,387

Table 2 Ten categories used in our experiment (originally in Japanese).

Manners, Ceremonial occasions
Entertainment and Hobbies
News, Politics, International affairs
Internet, PCs, Home appliances
Life and romance, Worries of human relations
Life, and Living guide
Health, Beauty, Fashion
Liberal Arts, Learning, Science
Sports, Outdoor,Cars
Region, Travel, Outing

We argue, however, that the above approach of utilising the actual question category as the gold-standard is not necessarily the best way to evaluate our classification systems. This is because different people may have different views about which category is most appropriate to a given question, and the category assigned by the questioner may not be the same as the one a CQA user, who is looking for an existing question by category, might choose. As the goal of our automatic question classifiers is to provide quick access to CQA users, the views of the CQA users may in fact be more important than that of the questioner. Moreover, the category assigned by the questioner reflects one person's point of view: the above evaluation methodology cannot take multiple viewpoints into account.

Based on the above argument, we constructed a set of crowdsourced category labels for a subset of the above test questions. Due to a budget and time constraint, we randomly selected 519 questions from each category, and thereby constructed a set of 5,190 questions for this second experiment. For each of these questions, five crowd workers were assigned, who independently labeled the questions with a category.

The crowd workers were shown a test question containing an image, and were asked to select the most appropriate category from the list shown in Table 2. They were instructed that the purpose of assigning a category was to enable quick access to the desired question on a CQA site. A total of 205 crowd workers participated in constructing the $5,190 * 5 = 25,950$ category labels.

Rather than deciding on one true category based on a majority vote, we fully utilised the labels from the five assessors as follows. Let C be the set of categories, and let $c_i \in C$ be the i-th category. Let $votes(q, c_i)$ be the number of crowd workers ($\leqq 5$) who assigned c_i to a given test question q. If the system assigns a category c_i to q, we give $votes(q, c_i)/5$ points to the system for q. For example, if the system agrees with two assessors for q, the *partial score* for q is $2/5 = 40\%$.

5. Experimental Results

5.1 Overall Results

Table 3 Experimental results. (a) is based on the correct categories as defined by the questioners. (b) is based on the gold labels obtained from the crowd workers.

ns.)

	(a) Accuracy (%) (69,355 questions)	(b) Mean partial score (5,190 questions)
image only	54.19	-
text only	73.48	0.5114
SP	77.37	0.5165
MCB	78.00	0.5193
SP+MCB	78.16	0.5210
DualNet	82.86	0.5411

Table 4 p-values / effect sizes (standardised mean differences with $V_E = .0419$ [9]) for the differences in mean partial scores.

	SP	MCB	SP+MCB	DualNet
text only	$p = .7140$ $ES = .0248$	$p = .2810$ $ES = .0386$	$p = .1150$ $ES = .0471$	$p = .0000$ $ES = .1449$
SP	-	$p = .9652$ $ES = .0137$	$p = .7924$ $ES = .0222$	$p = .0000$ $ES = .1201$
MCB	-	-	$p = .9942$ $ES = .0085$	$p = .0000$ $ES = .1063$
SP+MCB	-	-	-	$p = .0000$ $ES = .0979$

Table 3 summarises our experimental results. It can be observed that, in terms of both classification accuracy and mean partial score, the performance improves as we move down the table.

Table 4 shows the p-value for the difference in mean partial score for each system pair (excluding the substantially underperforming image-only baseline) based on Tukey HSD (Honestly Significant Differences) test with $B = 5000$ trials, along with effect sizes (standardised mean differences) [9]. It can be observed that the DualNet approach outperforms a text-only baseline ($p = .0000, ES = .1449$), a sum-and-product baseline ($p = .0000, ES = .1201$), MCB pooling ($p = .0000, ES = .1063$), and a combination of sum-and-product and MCB ($p = .0000, ES = 0.0979$).

5.2 Evaluation by Questioners' Categories

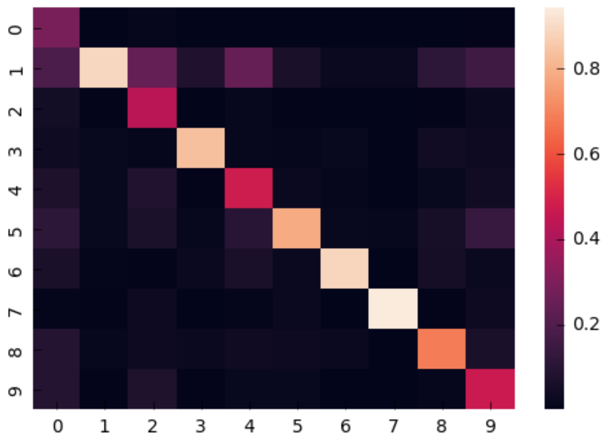

Figure 4 Heapmap of DualNet architecture results, where the horizontal axis represents the true category as defined by the questioners.

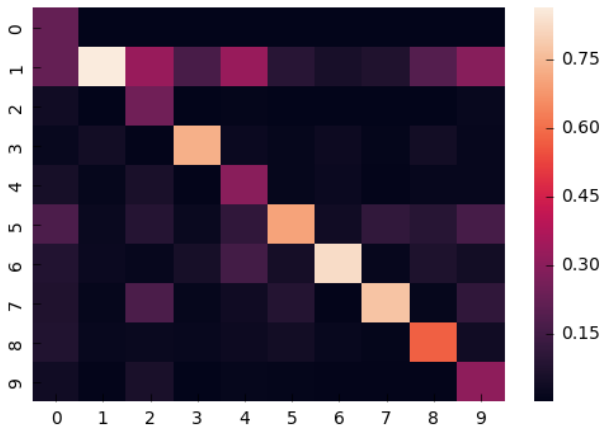

Figure 5 Heapmap of text-only architecture results, where the horizontal axis represents the true category as defined by the questioners.

Figures 4 and 5 show confusion matrices with the ten categories for the DualNet architecture results and the text-only results, respectively, in terms of heatmaps. The horizontal and vertical axes represent the true (as defined by the questioners) and the predicted categories, respectively, and the number of questions in each cell has been normalized by the number of correct questions for each cateogory. Thus, cells with light colors are those containing many questions. It can be observed that the true and predicted categories are more well-aligned with DualNet. In particular, we can see that the number of questions misclassified into Category 1 ("Entertainment and Hobbies") has been reduced compared to the text-only case.

5.3 Evaluation by Crowdsourced Categories

In this section, we evaluate the systems with crowdsourced gold standard data. To examine the discrepancy between the questioners' categories and the crowd workers' majority vote categories, we first computed the mean partial score by regarding the latter as the system's output; this gave us 0.5956, which is higher than the systems' scores shown in Table 3, but not very high. Moreover, Cohen's kappa between the questioners' categories and the crowd workers' majority vote categories is 0.5984, with a 95% confidence interval of [0.5839, 0.6129]. Thus, there is indeed a discrepancy between the questioner and the crowd workers, and hence it is also possible that there will be a discrepancy between the questioners' categories and the expectations of the CQA users.

Table 5 An example question with the category assigned by the questioner and that by the crowd workers

Question (translated from Japanese)	What is the inner diameter of the circular barbell pierce that Stav Strashko on his septum?
Questioner's Category	Category 6 ("Health Beauty, Fashion")
Majority vote from crowd workers	Category 1 ("Entertainment and Hobbies")

Table 5 provides an example from the classification results. This question asks about the inner diameter of a pierce. But the crowd worker selected Category 1 ("Entertainment and Hobbies"), probably because the question mentions "Stav Strashko," a model who was popular at that time.

6. Conclusions

Our experiments with real data from Yahoo Chiebukuro and crowdsourced gold-standard categories show that the DualNet approach outperforms a text-only baseline ($p = .0000$), a sum-and-product baseline ($p = .0000$), Multimodal Compact Bilinear pooling ($p = .0000$), and a combination of sum-and-product and MCB ($p = .0000$). where the p-values are based on a Tukey Honestly Significant Difference test with $B = 5000$ trials. Thus, while these effects are small, the DualNet approach appears to be the most promising for combining image and text representations for question classification.

Figures 4 and 5 shows both methods tend to misclassify into Category 1. Category 1 contained the highest number of questions while Category 0 ("Manners, Ceremonial occasions") contained the fewest. In future work, we would like to take countermeasures against such imbalance. Moreover, as we have observed that the categories given by questioners and those given by crowd workers often disagree, we would like to try training our networks from crowdsoured labels instead of the questioners'.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE ICCV 2015*, pp. 2425–2433, 2015.

[2] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *CoRR*, Vol. abs/1606.01847, , 2016.

[3] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of IEEE CVPR 2016*, 2016.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[6] Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 747–756. Association for Computational Linguistics, 2012.

[7] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. *CoRR*, Vol. abs/1511.03745, , 2015.

[8] Kuniaki Saito, Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada. Dualnet: Domain-invariant network for visual question answering. In *Multimedia and Expo (ICME), 2017 IEEE International Conference on*, pp. 829–834. IEEE, 2017.

[9] Tetsuya Sakai. Statistical reform in information retrieval? *SIGIR Forum*, Vol. 48, No. 1, pp. 3–12, 2014.

[10] Joshua B. Tenenbaum and William T. Freeman. Separating style and content with bilinear models. *Neural Comput.*, Vol. 12, No. 6, pp. 1247–1283, 2000.

[11] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pp. 649–657, 2015.