Adapting Visual Question Answering Models for Enhancing Multimodal Community Q&A Platforms

Avikalp Srivastava
Carnegie Mellon University
avikalps@cs.cmu.edu

Hsin-Wen Liu
Waseda University
stephanie1125@toki.waseda.jp

Sumio Fujita
Yahoo Japan Corporation
sufujita@yahoo-corp.jp

ABSTRACT

Question categorization and expert retrieval methods have been crucial for information organization and accessibility in community question & answering (CQA) platforms. Research in this area, however, has dealt with only the text modality. With the increasingly multimodal nature of web content, we focus on extending these methods for CQA questions accompanied by images. Specifically, we leverage the success of representation learning for text and images in the visual question answering (VQA) domain and adapt the underlying concept and architecture for automated category classification and expert retrieval on image-based questions posted on Yahoo! Chiebukuro, the Japanese counterpart of Yahoo! Answers.

To the best of our knowledge, this is the first work to tackle the multimodality challenge in CQA, and to adapt VQA models for tasks on a more ecologically valid source of visual questions. Our analysis of the differences between visual QA and community QA data drives our proposal of novel augmentations of an attention method tailored for CQA and use of auxiliary tasks for learning better grounding features. Our final model markedly outperforms the text-only and VQA model baselines for both tasks of classification and expert retrieval on real-world multimodal CQA data.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; **Question answering**;

KEYWORDS

Visual question answering, community question answering, multimodal Q&A platforms

ACM Reference Format:

Avikalp Srivastava, Hsin-Wen Liu, and Sumio Fujita. 2019. Adapting Visual Question Answering Models for Enhancing Multimodal Community Q&A Platforms. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM '19)*, November 3–7, 2019, Beijing, China. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3357384.3358000>

1 INTRODUCTION

Community question & answering (CQA) platforms enable users to crowd-source answers to posted queries, search and explore

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM '19, November 3–7, 2019, Beijing, China

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6976-3/19/11...\$15.00

<https://doi.org/10.1145/3357384.3358000>

questions, and share knowledge through answers. As the number of users increases, so does the information content; this makes it imperative to carefully design methods for categorizing and organizing information and identifying relevant content for personalized recommendations. Such end-tasks are of significant practical importance to CQA platforms, making them a big focus in information retrieval and natural language processing domains.

The CQA task of automatic question classification is useful for tagging newly posted questions and suggesting an appropriate question category to the asking user. It plays a crucial role in enabling users to find and answer questions in their area of expertise, thereby also facilitating effective answering. Another key problem to solve is that of retrieving “experts”. Here, the aim is to identify and retrieve users from the community who are likely to provide answers to a given question. This provides an efficient way to make the community well-knit, provide better content to askers, and recommend only the relevant questions from a gigantic pool of queries to the potential answerers.

A recurring feature in these tasks has been that the data is comprised only of text. Research datasets from Stack Exchange, Quora, and collections like TREC-QA rarely contain questions with a combination of text and images. In this work, we tackle data from the CQA website Yahoo! Chiebukuro (YC-CQA) ¹, where questions accompanied by an image form a considerable percentage (~10%) of the total posted questions (Fig. 1(a)). With Stack Exchange sites supporting images (~7%, 11%, 12%, and 20% image-based questions on computer science, data science, movies, and anime Stack Exchange sites, respectively), not to mention the numerous image-based threads on discussion platforms like Reddit, the advantages of our solutions for multimodal CQA are not limited to Chiebukuro.

Models using only text can give reasonable performances for multimodal CQA tasks (as we will see in our results), but there is potential to gain substantial improvements by utilizing the image data. It is easy to identify a couple of broad categories where image data will be essential for our end-tasks: i) where the image contains the actual question, and the question loses meaning without the image (Fig. 1(a) bottom-left example), and ii) where the image is necessary to make sense of the question text (top-mid & top-right examples in Fig. 1(a)). Images can also help reinforce the inferences from textual features (Fig. 1(a) top-left), or provide disambiguation over multiple topics inferred from text (‘plants’ and ‘shoes’ in Fig. 1(a) bottom-mid example).

Therefore, we focus on methods to best exploit the combined image-text information from multimodal CQA questions. Considering existing research at the intersection of vision and text, visual question answering models are dependent on deriving rich representations that encode a combined understanding of a question’s

¹<https://chiebukuro.yahoo.co.jp>

Figure 1: Questions sampled from (a) left: YC-CQA site (translated), and (b) right: VQA, taken from [9].

text-image pair. Thus, so as not to reinvent the wheel, we leverage the success of VQA architectures in deriving such joint representations and build novel augmentations to adapt them for CQA tasks.

In its most common form, the VQA task [3] is modeled as a classification task involving an image-question pair (Fig. 1(b)) and the selection of an answer from a fixed set of top possible answers. The main idea behind its proposal has been to connect the advances in computer vision and NLP, so as to provide an “AI-complete” task. However, given the nature of the questions and images, its direct practical applicability is limited. The questions are short, direct, and query the image, or at the most require common sense or objective encyclopedic knowledge. This is in contrast with the nature of questions found on the web where askers seek human expertise and the question texts provide context outside the input image, or are supported by the image. It is therefore important to properly identify and resolve the shortcomings of VQA models to enable better understanding of the image-text data from CQA.

While our contributions can be viewed under a more general lens, it is worth noting that given the significant percentage that image-based questions occupy on Chiebukuro, and the current policy on the site making it mandatory for asking users to provide a category from among hundreds of choices, improving automated category classification simplifies the introduction of the feature that suggests appropriate categories to the askers. It can even allow them to skip this part by assigning the predicted category automatically. Providing better expert retrieval has the obvious benefit of improving the responsiveness and quality of the QA service as a whole. More generally, the decision to use VQA-inspired end-to-end learning architectures makes our models generalizable and usable for image-based sections on other QA/discussion platforms, along with possessing the potential for extension to question answering. Therefore, in this paper,

- We closely analyze the differences between VQA and image-based CQA tasks, and identify the challenges in multimodal CQA that may hinder the performance of VQA models.
- Following this, we propose modifications to VQA-inspired models for a better CQA-task performance. Our key contributions include learning an additional global weight for the

image in the image-text combination step and introducing auxiliary tasks to learn better grounding features.

- We evaluate our model against baselines from text-only & VQA models, and other frequently used methods for image-text combination, on the Chiebukuro dataset.
- Finally, we conduct an ablation study to quantify the contributions of each of our suggested changes. The source code for our models is publicly available ².

A natural counterpart to VQA is answering image-based CQA questions. However, this is far more difficult and subjective compared to answering in VQA due to varying answer lengths and composition, the requirement of non-trivial external knowledge that must be modified according to the question’s context, and the necessity of human opinions. Therefore, we do not tackle answering in this work, but remain optimistic about our ability to use the results, inferences, and models from this work to answer a subset of simpler factoid-based CQA image questions in our future work.

To the best of our knowledge, our work is the first to tackle the challenges of multimodal CQA, and also to adapt VQA models for tasks on a more ecologically valid source of visual questions posted by humans seeking the expertise of the community (as opposed to straightforward questions that query the image). It is worth noting that [30] deals with the same dataset source, comparing joint embedding methods for a basic classification task, but does not attempt to address any CQA-specific challenges. By identifying and targeting such specific idiosyncrasies, we get a > 8% jump in classification accuracy and > 25% relative MRR increase on expert retrieval compared to the model in [30].

2 RELATED WORK

CQA Tasks. Initial approaches for question categorization and expert retrieval were heavily based on supervised learning approaches utilizing hand-crafted features ([29], [23], [21]), language models ([20], [39]), topic models ([41], [33]), and network structure information ([37], [28], [27]). A recent shift to end-to-end deep learning approaches ([38], [30], [11], [25]) has shown successful results for

²<https://github.com/avikalp7/VQAtoCQA>

Table 1: Statistics for VQA (English) and YC-CQA (Japanese) datasets.

Dataset	Total # of Questions	Question Text's Vocab Size	Answer Text's Vocab Size	Avg. Question Length (#words)	Avg. Answer Length (#words)	Total # of Question Categories	Average Categories per Question
DAQUAR	12,468	2520	823	11.53	1.15	3	1.00
COCO-QA	117,684	12,047	430	8.65	1.00	4	1.00
VQA v2	658,111	26,749	29,548	6.20	1.16	65	1.00
YC-CQA	1,018,833	176,921	335,658	71.54	62.15	38	2.73

these tasks and the related task of question similarity ranking. Apart from [30], all works focused on only the text modality.

Joint Image-Text Representation Learning. The most prevalent applications of works combining computer vision (CV), natural language processing (NLP), and knowledge representation & reasoning (KR) have been in image captioning and VQA tasks. While many methods for image captioning use the image representation as context for the decoder segment ([35], [4]), it can also be casted into an encoder-decoder framework that uses joint image-text representations [12]. We choose to instead focus on VQA for a number of reasons. First, there has been significantly more progress in VQA due to the hardness of evaluation of the image captioning task, along with the fact that the captioning task lacks the need for reasoning and requires only a single coarse glance at the image [3]. Second, the usage of joint representations in VQA (classification over answers) is more similar to our use case of classification and ranking, compared to text generation in captioning. Third, many recent works have developed models that can be used for both captioning and VQA ([31], [2]). Therefore, there seems no apparent reason to favor captioning, and we streamline our work by focusing on VQA methods.

VQA Methods. The two most common approaches for combining image-text representations in VQA are joint embedding methods and attention-based mechanisms [32]. One of the most basic methods is to simply concatenate the derived text and image embeddings [40]. A better method is concatenation of the element wise sum and product [24]. A bilinear pooling-based method was found to be effective by [6]. Attention-based approaches learn a convex combination of spatial image vectors as the contributor to the final joint embedding. A simple and popular attention model for VQA is the stacked attention network [34], where the text is seen as a query for retrieving attention weights for image regions. Methods involving attention over both image and text include [15], [18] among others. [6] also describes a version of its model that utilizes attention. More recent improvements in performance come from the use of bottom-up attention features [2], intricate attention mechanisms like bilinear attention maps [10], and careful network tuning and data augmentation methods [8].

3 UNDERSTANDING VQA-CQA DIFFERENCES

It is crucial to understand the differences between the question-image pairs in VQA and CQA in order to identify the unique challenges posed by the new dataset and address them by means of appropriate modifications. The dataset consists of questions posted

over one year on YC-CQA ³, which allows questions with and without an image. In this work, we only deal with questions accompanied by an image. Table 1 presents a simple comparison highlighting different aspects of the data, contrasting YC-CQA with the most commonly used VQA datasets. To better understand the contrast, we first analyze the quantitative differences and next discuss some of the more fundamental differences that are the driving influences for our proposed modifications.

3.1 Quantitative Dataset Differences

Table 1 highlights the complexity of CQA data in terms of a significantly larger vocabulary set and average question and answer lengths (despite different languages, the magnitude of difference is sufficient to drive the point). The CQA dataset presents a significantly higher noise in its text and image data. Common methods for question generation in VQA are to either automatically convert image caption data into questions or to have human annotators produce the questions on the basis of predefined guidelines. These lead to a sense of homogeneity in the questions - one that is missing in CQA, where question authors comprise a large number of different individuals. These differences are partly quantified by our experiment, where we select a subset of samples and retrieve the nearest neighbors for each text sample using Jaccard-Needham dissimilarity as the distance metric. We compare the mean average distance of the neighbors for the VQA and YC-CQA datasets. The result of this experiment, performed for randomly sampled 1k, 2k, and 3k sized subsets, demonstrates closer distances between a similar set of questions in VQA than in CQA (Fig. 2(b)).

The image diversity in CQA is expected to be greater as well. Images in the DAQUAR [17] dataset comprise indoor scenes, COCO-QA [19] images contain common-objects-in-context, and the VQA dataset [3] contains abstract scenes and clipart images. All these categories are subsumed by the images on the CQA platform, as users are not restricted in terms of the type or attributes of the image and the question they post. The results for an experiment similar to the one for texts, but using pre-trained ResNet-derived embeddings for sampled images, are shown in Fig. 2(a).

3.2 CQA Tasks

To clarify further differences, we first take a closer look at our two intended end-tasks.

3.2.1 Category Classification. The category assignment for a question is provided by the asking user. The available category choices

³Data available for purchase from <https://nttdata-nazuki.jp/data/chiebukuro.html>

are arranged in a hierarchical fashion, with each category having a single parent category. The number of level-0 categories is 14, followed by 95 level-1 and 415 level-2 categories. Each question's most specific category can come from any of these levels, with the condition that a level-1 or level-2 category labeled question is also labeled with the parent category. For example, for the category hierarchy '*Life Sciences > Plants & Animals > Plants*', the possible category assignments are '*Life Sciences*', or '*Life Sciences*' & '*Plants & Animals*', or '*Life Sciences*' & '*Plants & Animals*' & '*Plants*'.

Most level-2 and many level-1 categories have an extremely sparse presence in the dataset. Little training data is available, and for practical reasons it makes sense to skip such rare categories and settle for predicting only their parent category. Our final classification is done on 38 categories, selected by eliminating ones occurring in less than 5k samples. We treat this as a flat multi-label classification problem. Thus, a question tagged as '*Life Sciences > Plants & Animals*' is labeled as both '*Life Sciences*' and '*Plants & Animals*'. This also leads to lower training loss for over- and under-generalized predictions compared to completely wrong ones.

3.2.2 Retrieving Experts. We define our candidate pool of experts to contain users with more than 50 answers in the initial six-month period from which our dataset is drawn. Therefore, the relevant set of experts for a question is comprised of users that both answered the question and are present in the candidate pool. Similar to [13], we use mean reciprocal rank (MRR) as the evaluation measure. Note that in practical settings, we append features such as 'last active time of user', 'asking user's reputation', etc. to our feature set, but here we compare the models on their ability to retrieve experts based solely on text-image pairs of questions answered by users. This makes sense since the information contribution from other features is mostly orthogonal to this.

3.3 Identifying Fundamental Challenges

While more noise in the data poses a problem to any learning model, it is important to identify more pressing CQA challenges that question the fundamental assumptions of VQA models.

The first challenge stems from the difference in the role of images. In VQA, the question generally queries the image, and it is imperative to gain a visual understanding in order to answer it. This strong dependence on image is almost completely non-existent in CQA tasks for most questions. For many samples, the text contains enough information to successfully perform the task, and/or the image contains relatively little information, and at times is just posted as a placeholder, or is irrelevant for gaining an understanding of the question. The combination of image and text embeddings in VQA models has the implicit assumption of balanced information content from both for the end-task. Therefore, to deal with information imbalance between text and image in CQA samples, our first intended modification is to model this difference by learning an additional global weight for the image, which would signify its contribution towards the final joint embedding.

The second challenge is to correctly ground the text to image relation. Grounding here implies pairing the relevant objects or regions in an image to the corresponding references to them in the accompanying text. VQA questions are mostly single sentences

Figure 2: Image & text sample proximity in CQA vs. VQA. (a): Mean average distance to K-nearest neighbors for image representations. (b): Mean average distance to K-nearest neighbors for BoW text representations.

with keywords referring to objects in the image, and the final answer is dependent on such references. This leads to sound learning of grounding features. This is much more difficult for CQA because i) large question texts hamper identification of text regions where the image is referred, ii) low contribution of image towards the final task means that the model tends to skip grounding, and iii) the CQA tasks are simpler compared to VQA, as VQA needs the multimodal features to interact, while CQA tasks tend to focus more on textual features, thus impeding the model's ability to learn grounding features well. Thus, our second intended modification is to design tasks that help to learn these features better, and to use those features to improve performance on our main tasks. Intuitively, with this modification, we stand to benefit in the scenarios where multiple topics can be inferred from the textual features. In such cases, identifying the terms in the text that refer to the image can help the model to understand the question's subject better, and thus improve the image-text combination embedding.

Another intuitive observation is that attention cannot be expected to give the same impressive improvements for CQA tasks as it does for VQA. The reasons are similar: attention is suited for VQA, where salient characteristics of different image regions play differently important roles in both understanding the question and inferring the answer. For CQA, along with absence of such dependencies, poor grounding makes it harder to learn good attention weights. Therefore, overcoming the second challenge can also help to utilize the attention mechanism a bit better.

4 ADDRESSING CQA CHALLENGES

Now we present our solutions for the two identified challenges: varying information contribution from images across samples, and difficulty in learning grounding features.

4.1 Learning a Global Image Feature Weight

The strong image dependence in VQA makes it feasible to use methods such as element-wise sum-product and concatenation for image and text representations. Attention mechanisms learn attention weights for different image regions that are derived using a final softmax layer and so, sum to 1. These methods, however, provide no way for the model to learn to weigh the contributions of text and image separately for each sample, which becomes important for CQA,

where these two contribute significantly different amounts of information in different samples. We therefore introduce the learning of a global weight for image, both with and without attention.

4.1.1 Global Weight w/o Attention. Given the text and image vectors, we want to learn a parameter α that acts as the scalar weight for the image vector's contribution. This parameter is derived by the contribution of both the derived image vector $\mathbf{v}_I \in \mathbb{R}^d$ and the derived text vector $\mathbf{v}_T \in \mathbb{R}^d$, as:

$$\mathbf{h}_A = \tanh(\mathbf{W}_{IA}\mathbf{v}_I + \mathbf{W}_{TA}\mathbf{v}_T + \mathbf{b}_A), \quad (1)$$

$$\alpha = \sigma(\mathbf{W}_{A\alpha}\mathbf{h}_A + b_{A\alpha}), \quad (2)$$

where $\mathbf{W}_{IA}, \mathbf{W}_{TA} \in \mathbb{R}^{k \times d}$, and $\mathbf{W}_{A\alpha} \in \mathbb{R}^{1 \times k}$. Simply multiplying α by $\mathbf{v}_I$ to get the image contribution $\tilde{\mathbf{v}}_I$ can be problematic for the image-text embedding product we plan to use in the joint embedding in Eq. 5. Therefore, we distribute the α and $1 - \alpha$ parameters between $\mathbf{v}_I$ and a 'fall back' option $\mathbf{v}_{T'}$ obtained by a non-linear transformation on $\mathbf{v}_T$:

$$\mathbf{v}_{T'} = \tanh(\mathbf{W}_{TT'}\mathbf{v}_T + \mathbf{b}_{T'}), \quad (3)$$

$$\tilde{\mathbf{v}}_I = \alpha * \mathbf{v}_I + (1 - \alpha) * \mathbf{v}_{T'}. \quad (4)$$

The final image-text embedding is derived as

$$\mathbf{v}_{IT} = [\mathbf{v}_T + \tilde{\mathbf{v}}_I, \mathbf{v}_T * \tilde{\mathbf{v}}_I]. \quad (5)$$

4.1.2 Global Weight with Attention. Given the spatial image embedding $\mathbf{v}_{spI} \in \mathbb{R}^{d \times m}$, where d is the representation dimension for m image regions, attention weights and image contribution in [34] are derived as

$$\mathbf{h}_A = \tanh(\mathbf{W}_{IA}\mathbf{v}_{spI} \oplus (\mathbf{W}_{TA}\mathbf{v}_T + \mathbf{b}_A)), \quad (6)$$

$$\boldsymbol{\alpha}_I = \text{softmax}(\mathbf{W}_{A\alpha}\mathbf{h}_A + \mathbf{b}_\alpha), \quad (7)$$

$$\tilde{\mathbf{v}}_I = \sum_i \alpha_i \mathbf{v}_{spI_i}, \quad (8)$$

where $\mathbf{W}_{IA}, \mathbf{W}_{TA} \in \mathbb{R}^{k \times d}$, $\mathbf{W}_{A\alpha} \in \mathbb{R}^{1 \times k}$, $\oplus$ denotes the addition of a matrix and a vector.

To introduce the global image weight, we adopt an approach similar to the one used in [14]. For the m image regions, instead of learning $\{\alpha_1, \alpha_2, \dots, \alpha_m\}$ attention weights with $\sum_{i=1}^m \alpha_i = 1$, we learn an additional weight α_{m+1} such that now $\sum_{i=1}^{m+1} \alpha_i = 1$. This allows the model to attribute more weight to α_{m+1} (assigned to $\mathbf{v}_{T'}$) when the image contribution is determined to be low. The attention weights are derived as follows:

$$\mathbf{v}'_{spI} = [\mathbf{v}_{spI}, \mathbf{v}_{T'}], \quad (9)$$

$$\mathbf{h}'_A = \tanh(\mathbf{W}_{IA}\mathbf{v}'_{spI} \oplus (\mathbf{W}_{TA}\mathbf{v}_T + \mathbf{b}_A)), \quad (10)$$

$$\boldsymbol{\alpha}'_I = \text{softmax}(\mathbf{W}_{A\alpha}\mathbf{h}'_A + \mathbf{b}'_\alpha), \quad (11)$$

where $\mathbf{v}'_{spI} \in \mathbb{R}^{d \times (m+1)}$, $\mathbf{h}'_A \in \mathbb{R}^{k \times (m+1)}$, $\boldsymbol{\alpha}'_I \in \mathbb{R}^{1 \times (m+1)}$. The image contribution $\tilde{\mathbf{v}}_I$ is derived as in Eq. 8 with $\mathbf{v}'_{spI}$ replacing $\mathbf{v}_{spI}$. The joint embedding is then obtained as in equation 5.

4.2 Learning Grounding Features through Auxiliary Tasks

We discussed the problem of failing to learn grounding features in CQA. Using hints, i.e., predicting the features as an auxiliary task, is one of the proposed approaches for the problem of learning features that might not be easy to learn using only the original

task ([1], [22]), with success shown in recent work on sentiment analysis in [36] and on name error detection in [5]. We propose two auxiliary tasks to learn better grounding features and outline the training pipeline for utilizing these towards the final tasks.

Image-Text Matching Auxiliary Tasks. A comparatively more challenging task on the CQA data is matching a question's image to its corresponding text from among a pool of candidate texts, and vice-versa. To do this well, it's necessary to learn the regions in the text that refer to salient regions in the image, thus providing an effective logical solution to the problem of poor grounding. Furthermore, this task relies simply on clever data usage for training, requiring no extra labels or samples.

Formally, given our questions dataset $\mathcal{D} = \{(\text{Im}_i, T_i)\}_{i=1}^N$, where Im_i and T_i are the associated image and text with the i^{th} question, respectively, we construct two new training sets for the image-to-texts and text-to-images matching tasks. For the former, we set up the task as follows: given a question image and five candidate texts, the aim is to correctly identify the question text corresponding to the image among the candidates. We construct $\mathcal{D}^{IT} = \{\text{Im}_i, \mathbf{T}_{\text{cand}_i}\}_{i=1}^N$, where $\mathbf{T}_{\text{cand}_i} = \{T_{i1}, T_{i2}, T_{i3}, T_{i4}, T_{i5}\}$ such that $T_i \in \mathbf{T}_{\text{cand}_i}$, and the other four texts are negatively sampled. Similarly, for text-to-images matching, we construct the dataset $\mathcal{D}^{TI} = \{\mathbf{Im}_{\text{cand}_i}, T_i\}_{i=1}^N$, s.t. $\mathbf{Im}_{\text{cand}_i} = \{\text{Im}_{i1}, \text{Im}_{i2}, \text{Im}_{i3}, \text{Im}_{i4}, \text{Im}_{i5}\}$ and $\text{Im}_i \in \mathbf{Im}_{\text{cand}_i}$. The training is described in subsections 5.3 and 5.4.

5 FINAL MODEL DESCRIPTION

We now present the full picture of our model that utilizes the solutions we have proposed.

5.1 Text Representation

The text data is in the Japanese language. We perform some elementary preprocessing by removing HTML characters and replacing URLs with a special token. Tokenizing Japanese text is challenging since words in a sentence aren't separated by spaces. Therefore, we use the morphological analyzer Janome⁴ for word splitting.

We use randomly initialized word embeddings (trained end-to-end, similar to [34]), followed by a CNN-based architecture from [11] to derive the high-level text representation $\mathbf{v}_T$. CNN-based architectures have shown successful results in previous VQA works ([34], [16], [15]), and can be particularly useful for extracting features important for CQA tasks. We learn filters of sizes 1, 2, and 3 over the sequence of embeddings with max-pooling over each full-stride of a filter to obtain the text representation $\mathbf{v}_T$.

5.2 Image Representation

Most VQA works use networks pre-trained on ImageNet, such as the ResNet [7] or VGGNet [26]. Here, we use the pre-trained ResNet network, utilizing the final spatial representation for attention-based networks, and the final flat embedding for other models. Images are resized to 224 x 224, giving a 7 x 7 x 2048 dimensional spatial embedding $\mathbf{v}_{spI}$, and 2048-D flat embedding $\mathbf{v}_I$.

⁴<http://moco-beta.github.io/janome/en/>

Figure 3: Training pipeline for auxiliary and final tasks. Presence of the translucent light blue layer implies “frozen” weights, i.e., absence of backpropagation and weight updates through those channels.

5.3 Joint Representation and Final Layers

We use the global image weight with attention mechanism from Section 4.1.2 to get the joint embedding v_{IT} . v_{IT} can be input to different final layers for different tasks, which are described below.

5.3.1 Category Classification Layer. A fully connected layer with sigmoid activation is used for the multi-label classification task.

5.3.2 Expert Retrieval Layer. For expert retrieval, we try to score each candidate expert for each given question. Hence, we use an architecture inspired by [25], utilizing a matching matrix to score the candidate pool. Formally, given the joint embedding $v_{IT} \in \mathbb{R}^h$, for an expert with embedding representation $e_i \in \mathbb{R}^h$ (randomly initialized, learned end-to-end), the score for this expert is

$$\text{match}(e_i, v_{IT}) = e_i^T M v_{IT}, \quad (12)$$

where $M \in \mathbb{R}^{h \times h}$ is the randomly initialized, end-to-end learned matching matrix, as shown in Fig. 4.

5.3.3 Auxiliary Tasks. This is a five-class single-label classification task. Five joint embeddings are derived, as each sample has either five candidate images or five candidate texts. The prediction of the correct candidate uses the architecture shown in Fig. 5 for the

image-to-texts matching task. The combined representations (red in the figure) are passed through two convolutional layers to derive five scores at the end. Softmax is applied to these scores to obtain probabilities over the candidates. For the text-to-images matching task, the roles of image and text in Fig. 5 are simply reversed.

5.4 Training Pipeline

Fig. 3 shows the training pipeline. The three steps are:

- (1) First, the two main tasks and the auxiliary tasks are individually trained. For the auxiliary tasks, depending on the task being optimized for the current batch, either the text or image input is five-fold the batch size. Since the other input is tiled to be quintupled, the rest of the architecture (apart from input) remains the same for the two tasks, and the two losses are optimized without any scaling (Fig. 3(a)).
- (2) Next, we ‘freeze’ the text embeddings and text CNN for all three, and training the classification and retrieval models using text embeddings derived from concatenation of the original and the ones from the auxiliary tasks’ text CNN (Fig. 3(b)). The parameter sizes of the model can be changed to take in double the usual text embedding size, or an FC

Figure 4: Final layer for ranking experts, given $|E|$ candidates and embedding size h . The question’s image-text embedding (red) is derived using any of the VQA-based/adapted models, and then multiplied with the similarity matrix and expert embedding matrix to get a score for each candidate expert.

Figure 5: Architecture for image-to-texts matching auxiliary task. The image representation (blue) is tiled and combined with five candidate text representations (green) to derive matching scores for the image with each of the five texts.

layer can be used to reduce the dimensions by half. Both approaches produced similar results in our experiments.

- (3) After a sufficient number of epochs (25 in our experiments), we fine-tune the text CNN for the main tasks to gain further minor improvements, as shown in Fig. 3(c).

6 EXPERIMENTS

6.1 Setup

For both tasks, we use 80%/10%/10% splits for training, validation and test sets respectively. Batch size is 128 for main tasks and 32 for auxiliary tasks (due to the five-fold inputs). Training components include early stopping, learning rate & weight decay, and gradient clipping. For images, basic data augmentation and flipping is applied. For the text model, embedding size of 128 and filter sizes of 128 for 1-gram and 256 for 2- and 3-grams worked best. More detailed notes for exact reproducibility are outlined in our code release.

6.2 Results and Analysis

For both tasks, we use the following models as baselines:

- **Random:** Predict a random class for classification; do random ranking of users for expert retrieval.
- **Weighted Random:** Predict randomly with probability weights based on distribution in training data for classification; for

retrieval, a ranking based on answerer frequency of users on training data is used for all test samples.

- **Text-only:** Use text CNN from [11] with a fully connected (FC) layer at the end.
- **Image-only:** Use pre-trained ResNet + FC layer.
- **Dual-net:** The model used in [30]. The text representation derivation method for this model is different from the text CNN used in other models.
- **Embedding Concatenation:** Simple concatenation of base image and text embeddings.
- **Sum-Prod-Concat:** Element-wise sum, product, and subsequent concatenation, as done in [24] and [30].
- **Stacked Attention (SAN):** Based on [34].
- **Hierarchical Co-Attention (Hie-Co-Att):** Based on [15].
- **Multimodal Compact Bilinear Pooling (MCB):** Based on the non-attention-based mechanism in [6].

Our model (called **CQA Augmented Model**) has already been described in Section 5.

The results for all models are presented in Table 2. *Random* and *Weighted Random* models help to establish the difficulty of the task with respect to the performance measures used. The strong results from the *Text-only* baseline indicate that for most of the samples, text contains sufficient information for both tasks, providing empirical validation for the first identified challenge in Section 3.3. Seeking improvement by combining image and text information, we get a $\sim 3\%$ increase by using simple embedding concatenation methods for classification, and 0.025 MRR measure increment. DualNet [30] performs worse than the *Text-only* model since it uses a different, less powerful text representation. Their model with our text-CNN is essentially the *Sum-Prod-Concat* model.

As expected, we don’t obtain substantial improvements by using attention models, which deal better with texts that query different regions of the image. The *Hie-Co-Att* model is further constructed on the premise of utilizing the image-text attention mapping at the word, phrase, and sentence levels. This generalizes poorly for CQA data, where the final tasks do not benefit from learning correlation mapping between every text and image region.

By incorporating CQA-specific augmentations, our model is able to achieve a further $>4\%$ improvement on the classification task, and >0.015 MRR score improvement. From the perspective of being able to use the image data to improve performance, we have a substantial $\sim 8\%$ classification accuracy increase. As noted in [13], in the expert finding task, the ground truth relevance judgment set is incomplete, as there are possibly many ‘experts’ that possess the knowledge about a given topic, but only a small number of them actually answered the question. With this definition, comparing MRR for *Text-only* and our model, for any question, the lower bound for the expected number of users to be sent a recommendation so that at least one of them is a potential expert is down from 5 to 4.

Examples in Fig. 6 help to clarify the nature of samples where SAN (taking in both image and text) performs better than the *Text-only* model. Images with characteristics that are repeated across multiple samples - like mathematical problems on paper, fashion-wear items such as shoes, cables & PC equipment - are more likely to be useful towards the final task.

Figure 6: Samples misclassified by *Text-only* model, but correctly classified by *Stacked Attention Network (SAN)* model.

Table 2: Baseline model performances vs. CQA Augmented Model on YC-CQA test split.

Model	Category Classification Accuracy (%)	Expert Retrieval: MRR
Random	2.61	0.0092
Weighted Random	12.16	0.0605
Image-only	29.88	0.0849
Text-only	68.32	0.2071
Dual-net	68.04	0.2022
Embedding Concatenation	70.52	0.2310
Sum-Prod-Concat	71.35	0.2369
SAN 1-layer	72.08	0.2375
SAN 2-layer	72.05	0.2375
Hie-Co-Att	71.87	0.2365
MCB	72.01	0.2370
CQA Augmented Model	76.14	0.2529

6.3 Ablation Analysis

To quantify the contributions of different components in our final model, we re-evaluate after performing the following ablations:

- *W/O image weight*: Global image weight (Sec 4.1.2) is removed; simple attention and auxiliary tasks are used.
- *W/O auxiliary tasks*: Both auxiliary tasks (Sec 4.2) removed.
- *W/O Image-to-Texts Matching*: Among auxiliary tasks, only text-to-images matching is done.
- *W/O Text-to-Images Matching*: Among auxiliary tasks, only image-to-texts matching is done.
- *W/O Attention*: Uses global image weight without attention (Section 4.1.1) instead.
- *W/O Fine-tuning*: The third step described in the training pipeline (Fig. 3 bottom) is not performed.
- *SAN Big Att*: Dimensions for text feature and attention layer are increased so that the stacked attention model has a similar number of trainable parameters as our full model.

Table 3: Results for bigger SAN and ablations.

Model	Category Classification Accuracy (%)	Expert Retrieval: MRR
SAN Big Att	72.48	0.2379
SAN Big FC	72.37	0.2376
W/O Image Weight	74.84	0.2499
W/O Auxiliary Tasks	74.17	0.2474
W/O Image-to-Texts	75.23	0.2510
W/O Text-to-Images	75.06	0.2505
W/O Attention	75.14	0.2504
W/O Fine-tuning	75.82	0.2518
Full Model	76.14	0.2529

- *SAN Big FC*: Two fully connected layers are added instead.

The ablation results compared to our *full model* are shown in Table 3. We can see that even with an increased parameter budget, the stacked attention network's performance doesn't improve. The dip in performance of the *W/O Attention* model confirms our intuition that attention contributes better after having learned the grounding features through auxiliary tasks.

We observe significant dips in performance compared to the *Full Model* when using *W/O Auxiliary Tasks* or *W/O Image Weight* model, providing evidence that solutions to both of the identified challenges are crucial in improving the model.

We further investigate this by looking at randomly sampled qualitative examples presented in Fig. 7. We start by evaluating the contribution of the global image weight feature. Figures 7(a) and 7(b) show examples misclassified by *W/O Image Weight* but correctly classified by *Full Model* by assigning low and high global image weight (α_{m+1} value from Sec 4.1.2), respectively. The second and the fourth images in Fig. 7(a) come from a popular Japanese smart-phone game, with similar screenshots featuring across many samples. Other common image themes are those of automobiles. Most of these samples' text can be judged as ambiguous for text-only based classification. This demonstrates the capability of the model to attribute more attention to the image when image features are useful and textual information is not conclusive enough. On

(a) Samples misclassified by *W/O Image Weight* model, and correctly classified and assigned high global image weight by *Full Model*.

(b) Samples misclassified by *W/O Image Weight* model, and correctly classified and assigned low global image weight by *Full Model*.

(c) **Samples misclassified by *W/O Auxiliary Tasks* model and correctly classified by the *Full Model*.**

Figure 7: Examples misclassified by ablations (shown as ‘Pred Cat’), but correctly predicted by *Full Model*. Image weight stated in (a) and (b) is the sum of attention weights for all image regions. Image weight=1 implies equal image-text contribution.

the other hand, the text features in Fig. 7(b) samples can be seen as strong, with difficult to interpret images; these demonstrate the cases where the model succeeds by ignoring the image and focusing on text. The effect of grounding can be seen in Fig. 7(c), especially in the third example where the image-text combination is crucial for disambiguation over the ‘*sickness*’ and ‘*gardening*’ categories that can be inferred from the text. Examples from the automobile and animation categories are frequently observed in these misclassified cases, where generally both image and text provide clues for

effective classification. These examples qualitatively validate the usefulness of our proposed CQA-specific solutions.

7 CONCLUSION

We presented the challenges and solutions for dealing with classification and expert retrieval tasks on multimodal questions posted on the YC-CQA site. Among approaches at the intersection of vision and language, using representations from VQA models suits the problem best. Upon a thorough comparison of the two datasets,

we identified two fundamental problems in direct application of VQA methods to CQA: varying image information contribution in different samples, and poor learning of grounding features. We demonstrated that our model - based on our proposed solutions of learning an additional global image weight, and better grounding features through auxiliary tasks - outperformed baseline text-only and VQA models on both tasks. The performance on the two tasks and qualitative assessment from the ablations shows that the two proposed approaches are promising for tackling noisy image-text query data in the web domain. Since we base our work off VQA models, it also opens interesting avenues for future research, including identifying the multimodal CQA questions that can be answered using modified versions of the models developed in this study.

ACKNOWLEDGMENTS

We would like to thank Takehiro Yamamoto, Makoto P. Kato, and Tetsuya Sakai for their feedback and comments that greatly improved this manuscript.

REFERENCES

- [1] Yaser S Abu-Mostafa. 1990. Learning from hints in neural networks. *J. Complexity* 6, 2 (1990), 192–198.
- [2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6077–6086.
- [3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.
- [4] Xinlei Chen and C Lawrence Zitnick. 2015. Mind’s eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2422–2431.
- [5] Hao Cheng, Hao Fang, and Mari Ostendorf. 2015. Open-domain name error detection using a multi-task rnn. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 737–746.
- [6] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847* (2016).
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [8] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2018. Pythia v0.1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956* (2018).
- [9] Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 1983–1991.
- [10] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *Advances in Neural Information Processing Systems*. 1564–1574.
- [11] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).
- [12] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* (2014).
- [13] Xiaoyong Liu, W Bruce Croft, and Matthew Koll. 2005. Finding experts in community-based question-answering services. In *Proceedings of the 14th ACM international conference on Information and knowledge management*. ACM, 315–316.
- [14] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 6. 2.
- [15] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*. 289–297.
- [16] Lin Ma, Zhengdong Lu, and Hang Li. 2016. Learning to Answer Questions from Image Using Convolutional Neural Network. In *AAAI*, Vol. 3. 16.
- [17] Mateusz Malinowski and Mario Fritz. 2014. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger (Eds.). Curran Associates, Inc., 1682–1690.
- [18] Duy-Kien Nguyen and Takayuki Okatani. 2018. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6087–6096.
- [19] Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. In *Advances in neural information processing systems*. 2953–2961.
- [20] Fatemeh Riahi, Zainab Zolaktaf, Mahdi Shafiei, and Evangelos Milios. 2012. Finding expert users in community question answering. In *Proceedings of the 21st International Conference on World Wide Web*. ACM, 791–798.
- [21] Kirk Roberts, Halil Kilicoglu, Marcelo Fiszman, and Dina Demner-Fushman. 2014. Automatically classifying question types for consumer health questions. In *AMIA Annual Symposium Proceedings*, Vol. 2014. American Medical Informatics Association, 1018.
- [22] Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098* (2017).
- [23] Avigat K Saha, Ripon K Saha, and Kevin A Schneider. 2013. A discriminative model approach for suggesting tags automatically for stack overflow questions. In *Proceedings of the 10th Working Conference on Mining Software Repositories*. IEEE Press, 73–76.
- [24] Kuniaki Saito, Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada. 2017. Dualnet: Domain-invariant network for visual question answering. In *Multimedia and Expo (ICME), 2017 IEEE International Conference on*. IEEE, 829–834.
- [25] Aliaksei Severyn and Alessandro Moschitti. 2015. Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. ACM, 373–382.
- [26] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [27] Amit Singh and Karthik Visweswariah. 2011. CQC: classifying questions in CQA websites. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2033–2036.
- [28] Avikalp Srivastava and Madhav Datt. 2017. Soft Seeded SSL Graphs for Unsupervised Semantic Similarity-based Retrieval. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2315–2318.
- [29] Clayton Stanley and Michael D Byrne. 2013. Predicting tags for stackoverflow posts. In *Proceedings of IJCCM*, Vol. 2013.
- [30] Kenta Tamaki, Riku Togashi, Sosuke Kato, Sumio Fujita, Hideyuki Maeda, and Tetsuya Sakai. 2018. Classifying Community QA Questions That Contain an Image. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval*. ACM, 219–222.
- [31] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van den Hengel. 2018. Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence* 40, 6 (2018), 1367–1381.
- [32] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2017. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding* 163 (2017), 21–40.
- [33] Liu Yang, Minghui Qiu, Swapna Gottipati, Feida Zhu, Jing Jiang, Huipeng Sun, and Zhong Chen. 2013. Cqarank: jointly model topics and expertise in community question answering. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 99–108.
- [34] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 21–29.
- [35] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4651–4659.
- [36] Jianfei Yu and Jing Jiang. 2016. Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 236–246.
- [37] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. 2016. Expert Finding for Community-Based Question Answering via Ranking Metric Network Learning. In *IJCAI*. 3000–3006.
- [38] Chen Zheng, Shuangfei Zhai, and Zhongfei Zhang. 2017. A Deep Learning Approach for Expert Identification in Question Answering Communities. *arXiv preprint arXiv:1711.05350* (2017).
- [39] Xiaolin Zheng, Zhongkai Hu, Aiwu Xu, DeRen Chen, Kuang Liu, and Bo Li. 2012. Algorithm for recommending answer providers in community-based question answering. *Journal of Information Science* 38, 1 (2012), 3–14.
- [40] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. 2015. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167* (2015).
- [41] Yanhong Zhou, Gao Cong, Bin Cui, Christian S Jensen, and Junjie Yao. 2009. Routing questions to the right users in online communities. In *2009 IEEE 25th International Conference on Data Engineering*. IEEE, 700–711.