DiagNet: Bridging Text and Image

Mark Jin **Shengyi Qian** **Xiaoyang Shen** **Yi Wen** **Xinyi Zheng**
University of Michigan, Ann Arbor
{kinmark, syqian, xyshen, wennyi, zxycarol}@umich.edu

Project ID: 1
Authors contributed equally and are listed alphabetically.

Abstract

Visual Question Answering (VQA) is to answer open-ended natural language questions related to images. Modern VQA tasks require reading and reasoning of both images and texts. We propose DiagNet, an attention-based neural network model that can effectively combine multiple evidence of texts, images and texts in images. Within DiagNet, a novel multi-task training strategy is used to combine answer type evidence in a hybrid fusion. We conduct comprehensive evaluation on multiple VQA tasks, and achieve competitive results. Our code is available at https://github.com/WYchelsy/DiagNet.

1 Introduction

Visual Question Answering (VQA) is a multi-modal task which combines Computer Vision and Natural Language Processing. A VQA system takes an image, and a free-form, open-ended, natural language question as input, and produces a natural language answer as the output [1]. Different kinds of VQA models and tasks are proposed step by step these years [2–4]. Those models are designed from perception, to reading and reasoning. In this paper, we proposed DiagNet, an attention-based model that has the ability to detect objects, read text in images, and answer questions by reasoning over them. Figure 1 illustrates the overall pipeline of our approach.

Endowing VQA models with the reading and reasoning ability is challenging mainly in two folds. Firstly, there are three reasoning channels: question texts, image features and texts in images, while current vision and language tasks mostly only consider first two sources. Moreover, since texts in images are typically generated from an Optical Character Recognition (OCR) module, the model will suffer propagation error from the earliest stage during inference: recognizing texts in images. Multiple feature sources with noise raise challenges in reasoning ability and robustness of the model. Secondly, answers can come from either frequent answers or texts in images, which requires a single model needs to simultaneously handle questions from both (a) frequent answer set and (b) texts in images. As pointed out in [4], current VQA models suffer from such heterogeneous answer sources.

Towards the first challenge, we propose the main architecture of DiagNet, as shown in Figure 2. The intuition behind DiagNet is that we want to effectively combine three channels of inputs: question texts, images, and texts in images. We build insights from state-of-the-art image captioning and VQA tasks, where visual attention mechanism is widely adopted. We use a Bottom Up Top Down (BUTD) mechanism for features interplay [5]. Specifically, we let question texts features have BUTD co-attention with image features, and OCR tokens features respectively. Towards these attention feature outputs, one key building block is the multi-modal feature fusion. Most existing approaches simply use linear models (e.g., concatenation or element-wise addition) to integrate the visual feature from the image with the textual feature from the question, even their distributions may vary dramatically [6]. Such linear modals may fail to integrate the complex information from different modals effectively. We propose to use a bi-linear pooling based methods to achieve hierarchy multi-modal feature fusions.

Preprint. Work in progress.

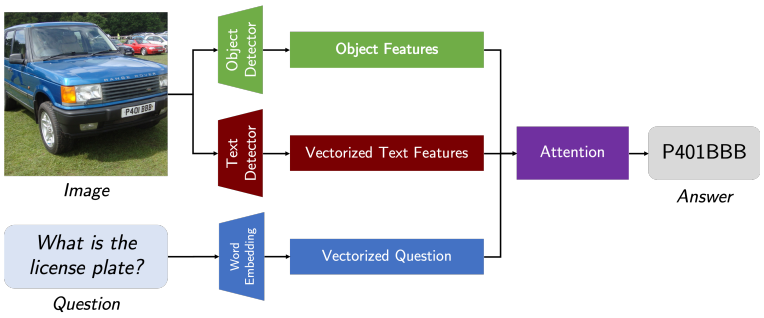

Figure 1: DiagNet detects objects and reads text in images, and then answers questions by reasoning over them.

To the best of our knowledge, we are the first to use multiple bi-linear pooling based modules[3, 2] to achieve hierarchy multimodal feature fusions in VQA tasks.

To overcome heterogeneous answer sources, there are two main methods. LoRRA [4] integrate the OCR output of texts in images into the frequent answer set before feeding information into the model. The model is trained end-to-end, letting the model learn itself which answer source to choose. We call this approach *early fusion*. [7] proposed Multi-Output Model (MOM) to determine whether a question is a yes/no question firstly, and trained two different branches for yes/no question and chart-specific questions respectively. We call this approach *late fusion*, because the final output can be viewed as an aggregation of two separate models, where they are integrated in the last step of inference. Inspired by those two methods, we proposed a novel *hybrid fusion* strategy, where we introduce answer type evidence in both early and late stages. We integrated the OCR outputs of texts in images into the frequent answer set before feeding into the model, serving as an early stage evidence for answer source selection. Moreover, to let the model have self-assessment ability on whether the question related to the image is answerable from the frequent answer set, we train another binary classifier determining if the question is answerable from frequent answer set, together with the answer prediction task in a multi-task training style. The model can learn the information of answer source selection both in accurate answer prediction, and the binary classification. The architectures for the two outputs share parameters except for the last linear layer. We believe that pair the main training task (answer prediction) with a related task (answer type prediction) can help the model converge better, and learn more side information from the feature outputs.

Our main contributions can be summarized as follows.

(i) We propose a new neural architecture called DiagNet, that has the ability to read text in images and answer questions by reasoning over the text, objects, and questions.

(ii) We propose a novel multi-task training strategy combining answer type evidence in a hybrid fusion.

(iii) We conduct comprehensive evaluation on multiple VQA tasks, and achieve competitive results.

2 Related Work

Multimodal Machine Learning. Our work is related to general multimodal machine learning. Families of Dynamic Bayesian networks [8] and discriminative sequential models [9] were proposed earlier this century, however they suffer from learning deep representations of multimodal features. Recently, a family of deep multimodal architectures have been proposed, such as Bimodal Deep Belief Network [10], Deep Boltzmann Machine [11], etc. In terms of information fusion, both model-agnostic approaches and model-based approaches are proposed. Our work can be categorized as an application of model-agnostic deep multimodal architectures designed specifically for difficult VQA tasks. In this context, our work is novel in overcoming fusion challenge, where we propose a hybrid fusion strategy, distinct from solely early or solely late fusion.

Object and Text Detection. Two-stage object detectors was first introduced and popularized by R-CNN [12]. By introducing a region proposal network (RPN), Faster RCNN becomes an important

benchmark for object detection [13]. Faster-RCNN is also adopted to build a text detection system, called Rosetta [14].

VQA Models. Our work is directly related to current VQA models. The commonly used method is the 2-channel vision + language model concurrently proposed by [15] and [16]. The Bayesian-based approaches [17, 18] try to model the co-occurrence statistics of both the visual and the semantic features, which intends to infer the relationships between the image and the questions. The attention-based models attempt to overcome the limitations that using global features alone may obscure the question-related regions in the image. These models are trained to attend the important features respecting question answering, ignore the irrelevant information and model the interactions across the vision and language modalities. In recent works, the stacked attention networks (SANs) [19] and Dense Symmetric Co-attention (DCN) [20] perform well in VQA tasks using the attention mechanism. Our work lies under the category of attention based models.

3 Preliminaries

3.1 Problem Setup

The task we are tackling can be formalized as follows. Given a question $Q = (q_1, q_2, \cdots, q_W)$, and an image I, the VQA system should generate an free-form answer Ans. Notice that the image I is allowed to have multiple text tokens $I_{t_1}, I_{t_2}, \cdots, I_{t_N}$ presented. In the training step, the only supervision signal is the answer ground truth Ans, and there is no ground truth for $I_{t_1}, I_{t_2}, \cdots, I_{t_N}$.

3.2 OCR: Optical Character Recognition

In the task of OCR, given an image, the OCR system should correctly extract the text overlaid or embedded in the image. Challenges to such a task compound as a number of potential fonts, languages, lexicons, and other language variations including special symbols, non-dictionary words, or specific information such as URLs and email ids, and images tend to vary in quality with text in the wild appearing on different backgrounds. Since we need to endow the model with the ability to read texts in images, an OCR system is serving as an upstream module for downstream reasoning tasks over question texts and vision contents. Given an Image I, the output of the OCR system on this image is formalized as $OCR(I)$, where $OCR(I) = (o_1, o_2, \cdots, o_N)$ is a sequence of tokens. We use $OCR(I)$ to approximate the $I_{t_1}, I_{t_2}, \cdots, I_{t_N}$, where I_{t_i} is a token of text in images.

3.3 MFB & MFH

Multi-modal Factorized Bilinear Pooling (MFB) [3] is to get an o-dimensional output vector z that integrates information from different modals. Say we have $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$ from two modals, MFB will project x and y onto a higher-dimensional space with projection matrix $\tilde{U} \in \mathbb{R}^{m \times ko}$ and $\tilde{V} \in \mathbb{R}^{n \times ko}$. ko is the latent dimension of the high-dimensional feature space. Then elementwise multiplication is used to fuse the high-dimensional feature vectors $\tilde{U}^T x$ and $\tilde{V}^T y$. In order to get the low-dimensional fused representation $z \in \mathbb{R}^o$, we perform SumPool.

$$z_{exp} = f_1(\tilde{U}^T x \circ \tilde{V}^T y) \in \mathbb{R}^{ko} \qquad z = f_2(\mathrm{SumPool}(z_{exp}, k)) \in \mathbb{R}^o \tag{1}$$

SumPool uses a one-dimensional non-overlapped window with size k to perform sum pooling over the input vector $\in \mathbb{R}^{ko}$ to get an o-dimensional output vector. Activation f_1, f_2 can be applied before the SumPool and after.

To further improve the multi-modal representation, Generalized Multi-modal Factorized High-order Pooling (MFH) [2] cascades the MFB blocks. To make p MFB blocks cascadable, we modify Eq.(1) as

$$z^i_{exp} = z^{i-1}_{exp} \circ \left[f_1((\tilde{U}^i)^T x \circ (\tilde{V}^i)^T y) \right] \qquad z^i = f_2(\mathrm{SumPool}(z^i_{exp}, k)) \in \mathbb{R}^o \tag{2}$$

where $i = 1, \ldots, p$ is the index for the i^{th} MFB block. $z^{i-1}_{exp} \in \mathbb{R}^{ko}$ is the internal feature of the $i-1^{\text{th}}$ MFB block and $z^0_{exp} \in \mathbb{R}^{ko}$ is an all-one vector. After the internal feature is calculated for the i^{th} block, we can use Eq.(1) to get the output z^i for that block. The final output z of the high-order MFH^p is obtained by concatenating the z^i in different blocks, $z = [z^1, \ldots, z^p] \in \mathbb{R}^{op}$ and will be passed to the downstream tasks.

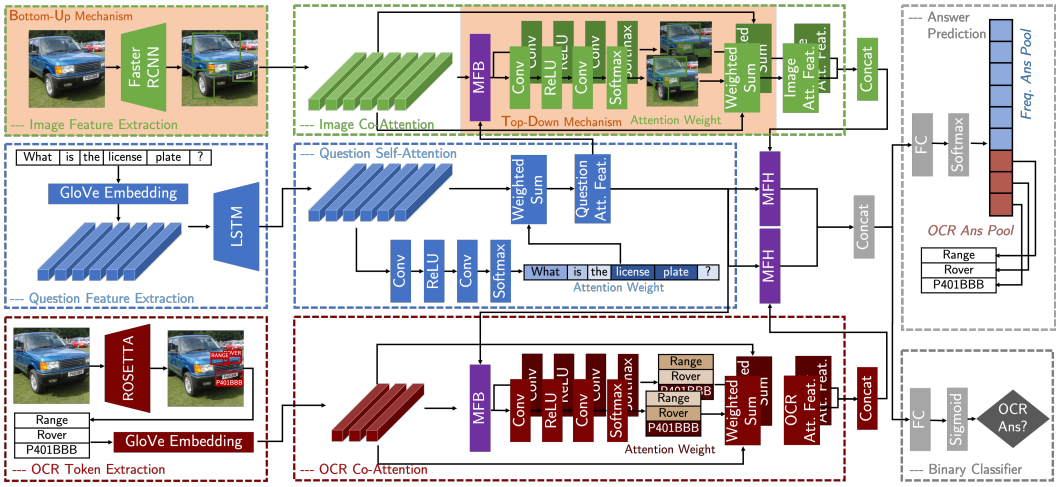

Figure 2: The architecture of DiagNet. The first row corresponds to the image object branch. The second row is the question branch. The third line is the OCR token branch.

4 Our Approach

4.1 Image Features

We use exactly the same method as [5] to extract object-level features. Faster-RCNN is used to extract the features of all detected objects. The output of each image is a $K \times 2048$ tensor, where $K \in [10, 100]$. Meanwhile, we use Rosetta [14] to extract the text in images. The output of each image is a few words and their corresponding bounding boxes.

4.2 Text Features

We embed question words $q_1, q_2, \cdots, q_W$ with GloVe[21] vectors, and get an embedding matrix Q_{emb}. Then we take the last hidden state of a one layer as the question feature vector q for a given question Q, where $q = LSTM(Q_{emb})$. Similarly, we embed $OCR(I)$ (the OCR output of images) with GloVe[21] vectors, and get an embedding matrix OCR_{emb} for each image I.

4.3 Image and Text Interplay

Question Self-attention Since the questions are interpreted in natural languages, the contribution of each word will be definitely different. We mimic human's behavior to train the question attention mechanism with the assumption that human can infer the question attention (i.e., the key words in the question) without seeing the image. *Question Self-attention* module will produce different attention weight to each word q_i of the question and yield the attended question feature q_{att}. In case we have several *Question Self-attention* modules, we concatenate the q_{att} to get the final attended question vector q_{att}

Question-Image Co-attention For the same image, different question will result in completely different answers, which means that different attention should be cast on the proposed regions in the image with the different question. The *Question-Image Co-attention* module, which can predict the relevance between each proposed object region of the image with the question, is beneficial for predicting best-matching answer accurately [2]. For each vector representation i_{λ_i} of the proposed object region λ_i in the image, we fuse the representation with the attended question vector q_{att} and generate the attention weight α_{λ_i} for each object region using the *Question-Image Co-attention* module. Then, we can get the attended image feature i_{att} with softmax normalization (Eq.4).

$$v_{\lambda_i, fused} = \text{MFB}(q_{att}, i_{\lambda_i}) \qquad \alpha_{\lambda_i} = \text{BUTD}(v_{\lambda_i, fused}) \tag{3}$$

$$i_{att} = \sum_{\lambda_i} \frac{\exp(\alpha_{\lambda_i})}{\sum_{\lambda_j} \exp(\alpha_{\lambda_j})} i_{\lambda_i} \tag{4}$$

In case we have several *Question-Image Co-attention* modules, we concatenate the i_{att} for each module to get the final attended image vector i_{att}.

Question-OCR Co-attention. We perform a co-attention mechanism between the attended question vector and each vector representation of the OCR tokens, in the similar way as the *Question-Image Co-attention* to produce the attended OCR vector o_{att}.

For the *Question Self-attention Module*, *Question-Image Co-attention Module* and *Question-OCR Co-attention Module*, they share the same architecture where the input vectors will go through some feature transformations and softmax normalization to predict the attention weight α for each input vector. However, the input for the *Question Self-attention Module* are question word embeddings while the input for *Question-Image Co-attention Module* and *Question-OCR Co-attention Module* are the fused question-image vector and fused question-OCR vector for each each proposed object region and each OCR token embedding.

4.4 Answer Prediction

We fuse the attended question vector q_{att} with the attended image vector i_{att} using MFH. We also fuse the attended question vector q_{att} with the attended OCR vector o_{att} using MFH. Then we concatenate the two fused vector as the input vector v_{final} for answer prediction.

$$v_{i,q,fused} = \mathrm{MFH}(i_{att}, q_{att}) \qquad v_{o,q,fused} = \mathrm{MFH}(o_{att}, q_{att})$$
$$v_{final} = \mathrm{Concat}(v_{i,q,fused}, v_{o,q,fused})$$

For the answer prediction. We simply regard the frequent answer pool and the OCR answer pool as the same and produce the confidence score for each answer in the answer pool. The frequent answer pool is composed of the frequent answers from the training set and the OCR answer pool is composed of the OCR tokens extracted from the image. Note that the OCR answer pool will change according to the different OCR tokens extracted for each image but the frequent answer pool will not.

$$\mathrm{AnswerPoolScore} = \mathrm{softmax}(\mathrm{FC}(v_{final}))$$

4.5 Multi-task Training

We integrate the answer type evidence in a hybrid fusion. In the early stage, we feed OCR_{emb} into *Question-OCR* module. In the late stage, whether $OCR(I)$ is in answer set serves as the ground truth for binary classifier. We use another fully-connected layer to produce the binary prediction. Formally,

$$\mathrm{BinScore} = \mathrm{sigmoid}(\mathrm{FC}(v_{final})) \tag{5}$$

We minimize two loss functions simultaneously by using a linear combination of them.

$$Loss = \mathrm{KLDIV}(\mathrm{AnswerPoolScore}, \mathrm{Ans}) + \beta \mathrm{CrossEntropy}(\mathrm{BinScore}, \mathrm{BinaryLabel}) \tag{6}$$

where β is a hyperparameter.

5 Experiments

5.1 Datasets

We conduct experiments on two datasets: We train our model on both TextVQA v0.5 dataset and VQA v1.0 dataset [1, 4]. In TextVQA v0.5 dataset, there are 28,000 images from Open Images, and the images are allowed to have texts. VQA v1.0 dataset has 200,000 images from MS-COCO. We provide results on VQA v1.0 as complementary experiments to show that our model can generalize well on other domains, while robust models on VQA v1.0 do not need to have the ability of reading texts in images.

5.2 Implementation Details

We use the ResNet-152 [22] pretrained on ImageNet [23] to extract `2048d` vectors or the Faster R-CNN [24] used by BUTD Attention Model [25] pretrained on Visual Genome [26] to extract `k x 2048` tensors as image features. We use the `300d` GloVe [27] vectors as part of the word embedding for tokens in the images and the questions to ask. These parameters are frozen.

In the question branch, the question words are embed with the pre-trained `300d` GloVe [27] and `300d` embedding to be trained. The `600d` embedded sequence will be fed into an LSTM layer of `hidden_size=1024, num_layers=1`, followed with dropout `p=0.3`. The output is the unattended question features of `T x 1024`. For every word, the `1024d` feature is fed to `fc-512, relu, fc-2` and use `softmax` to generate the attention weight for two *Question Self-attention* modules. We concatenate the two weighted sum to derive `2048d` attended question feature.

In the object branch, the `K x 2048` tensors consist of the `2048d` features of the region for K detected objects. Then the attended question vector and the image feature vector at each region go through a MFB block with pre-activation `f1=Dropout(p=0.1)`, window size `k=5`, out dim `o=1000` and post-activation `f2` as signed square root followed by l_2-normalization activation. For every region, the `1000d` feature is fed to `fc-512, relu, fc-2` and use `softmax` to generate the attention weight for two *Question-Image Co-attention* modules. We concatenate the two weighted sum to derive `4096d` attended image feature. Then the image vector co-attended with the question and the attended question vector go through a `p=2` MFH block with pre-activation `f1=Dropout(p=0.1)`, window size `k=5`, out dim `o=1000` and post-activation `f2` as signed square root followed by l_2-normalization activation. The output is `2000d` image-question feature vector.

In the OCR token branch, the `N x 300` tensors are the N OCR tokens embedded by `300d` GloVe [27]. Then the attended question vector and the token feature vector for each OCR go through a MFB block with pre-activation `f1=Dropout(p=0.1)`, window size `k=5`, out dim `o=1000` and post-activation `f2` as signed square root followed by l_2-normalization activation. For each fused OCR token vector, the `1000d` feature is fed to `fc-512, relu, fc-2` and use `softmax` to generate the attention weight for two *Question-OCR Co-attention* modules. We concatenate the two weighted sum to derive `600d` attended OCR token feature. Then the token vector co-attended with the question and the attended question vector go through a `p=2` MFH block with pre-activation `f1=Dropout(p=0.1)`, window size `k=5`, out dim `o=1000` and post-activation `f2` as signed square root followed by l_2-normalization activation. The output is `2000d` token-question feature vector.

We concatenate image-question feature and token-question feature. The `4000d` vector is fed into `fc-1` for the binary classifier and `fc-3104` for 3000 frequent vocabulary words and no more than 104 flexible OCR tokens.

5.3 Experimental Results

Experiments on TextVQA v0.5 dataset. We train DiagNet on TextVQA v0.5 dataset, which requires both reading text and identifying objects in images. The results are shown in Table 1. Our model significantly outperform Pythia [28], which does not contain a text detection branch. However, our model is not as good as LoRRA + Pythia [4]. We attribute the performance difference to the following reasons: (1) LoRRA + Pythia uses fine-tuned state-of-the-art VQA model Pythia as one module of whole pipeline, while we need to train all three branches from scratch; (2) LoRRA uses a different sophisticated pointer network on OCR token prediction, while due to limited time and computing resources, we only use linear layer to predict OCR token; (3) LoRRA + Pythia uses a lot of training tricks such as different learning rates for different layers and data augmentation, while we do not; (4) LoRRA + Pythia is trained on 8 GPUs and uses Visual Genome [29] and Visual Dialog [30] as additional training data, while we don't have so many computing resources. We believe that with more computing resources, time and engineering tricks, our model could achieve similar results.

Complimentary experiments on VQA v1.0 dataset. We also train DiagNet on VQA v1.0 dataset to show our DiagNet is competitive on task that does not need text detection. The results are shown in Table 2. DiagNet is comparable to MFB, with or without BUTD mechanism.

Model	Accuracy
Pythia [4]	13.04
DiagNet	18.77
LoRRA+Pythia [4]	26.56

Table 1: Single model performance (in %) on TextVQA v0.5 dataset.

Model	all	yes/no	number	other
MFB [3]	55.97	79.99	33.48	43.65
DiagNet w/o BUTD	56.47	80.78	33.10	44.14
MFB+CoAtt+BUTD [5]	64.63	83.74	40.09	56.48
DiagNet	65.01	84.64	39.50	56.72

Table 2: Our single model performance (in %) on VQA v1.0 validation set.

5.4 Ablation Study

We conduct experiments shown in Table 3 to investigate the contributions of various components in DiagNet. DiagNet-OCR means the model without the multi-task binary classifier train strategy. DiagNet without BUTD means that the model only uses image features without object detection. DiagNet-late means we use late fusion to integrate answer type information, in other words, we aggregate two separate models, and train a binary classifier to choose from the models. DiagNet-Binary means we use the multi-task training strategy, but we let the binary classifier determine which answer set to choose. Notice that experiments on VQA do not include OCR token branch.

Model	Accuracy
DiagNet w/o BUTD & OCR	11.42
DiagNet w/o OCR	11.25
DiagNet-late	15.34
DiagNet-binary	15.86
DiagNet w/o BUTD	18.22
DiagNet-OCR	18.44
DiagNet	**18.77**

Table 3: Ablation Study on TextVQA v0.5 dataset.

OCR token branch We can see that adding OCR to DiagNet can improve the performance of the model of 7.21% accuracy(DiagNet w/o BUTD & OCR versus DiagNet without BUTD, DiagNet w/o OCR versus DiagNet-OCR). OCR token branch is very significant, since we want the model to have the ability to read texts in the image.

BUTD Bottom-Up attention mechanism proposed image regions, each with an associated feature vector, while top-down mechanisms determine feature attention weights. Adding BUTD could help model combine better attentive features of images with question texts. In TextVQA, the improvement is slight, while in VQA, the improvement is a big margin (8.62%). Because VQA mostly requires model to recognize the object in the image, while TextVQA involves more sophisticated reasoning process.

Multi-task training with binary classifier Pairing the answer prediction task with a binary classifier can improve the model performance. Because information is propagated in the model in a hybrid fusion, helping model combine more side information and converge better. Performance of DiagNet-Binary & DiagNet-late is lower than DiagNet, which means that a late fusion of information combination is not ideal in TextVQA. Empirically we find that the accuracy of training a binary classifier is only around 80%, so the whole model accuracy significantly drops because the binary classifier can not choose from the correct answer set. To this extent, combining answer type in a hybrid fusion is ideal, as the model can learn side information from the binary classifier, but the inference is still largely decided by the answer prediction model, instead of the binary classifier.

5.5 Error Analysis

5.5.1 VQA

For VQA dataset, we can categorize the errors as (a) Question Understanding Failure. Fail to understand the question. (b) Image Attention Error. The attended region of the image does not correspond to the question. (c) Lack of Knowledge Base. Some of the answers don't directly come from the object detected in the image. Instead, further knowledge is needed.

Figure 3: Error sampled from the VQA v1.0 validation set. The first two samples correspond to Question Understanding Failure. The middle two samples correspond to Image Attention Error. The last two samples correspond to Lack of Knowledge Base.

5.5.2 TextVQA

For TextVQA dataset, we can also categorize the errors as Question Understanding Failure, Image Attention Error and Lack of Knowledge Base. Also, given the additional text branch in the TextVQA task, OCR Tokens Extraction Error also occur.

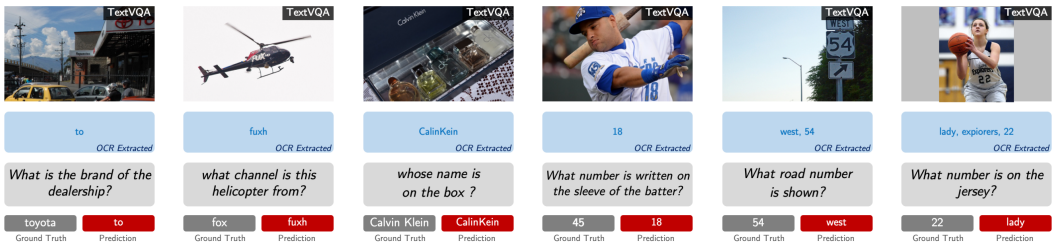

Figure 4: Error sampled from the TextVQA v0.5 validation set. The first three samples correspond to OCR Tokens Extraction Error. The last three samples correspond to Image Attention Error.

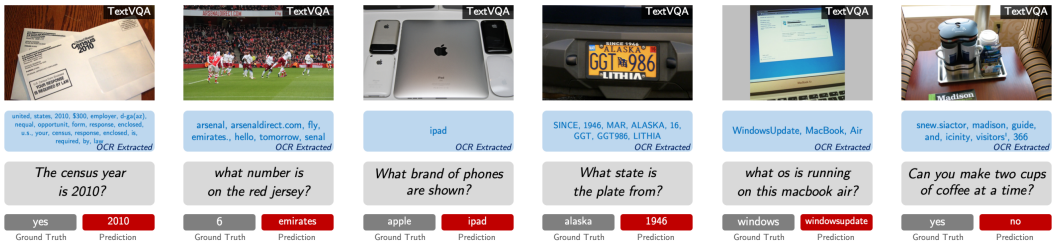

Figure 5: Error sampled from the TextVQA v0.5 validation set. The first two samples correspond to Question Understanding Failure. The last four samples correspond to Lack of Knowledge Base.

6 Conclusions

We propose DiagNet, an attention-based neural network model that can effectively combine multiple evidence of texts, images and texts in images. Within DiagNet, a novel multi-task training strategy is used to combine answer type evidence in a hybrid fusion. We conduct comprehensive evaluation on multiple VQA tasks, and achieve competitive results.

7 Author Contribution Statement

M.J. and S.Q. performed the data preprocessing. X.Z., Y.W. and X.S. designed the architecture. X.S. and S.Q performed the experiments. M.J., X.Z., Y.W. and S.Q. wrote the article.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[2] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multi-modal factorized high-order pooling for visual question answering. *CoRR*, abs/1708.03619, 2017.

[3] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. *CoRR*, abs/1708.01471, 2017.

[4] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 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*, pages 6077–6086, 2018.

[6] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Simple baseline for visual question answering. *CoRR*, abs/1512.02167, 2015.

[7] Kushal Kafle, Brian L. Price, Scott Cohen, and Christopher Kanan. DVQA: understanding data visualizations via question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5648–5656, 2018.

[8] Ashutosh Garg, Vladimir Pavlovic, and James M. Rehg. Boosted learning in dynamic bayesian networks for multimodal speaker detection. 2003.

[9] Phil Blunsom and Trevor Cohn. Discriminative word alignment with conditional random fields. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 65–72. Association for Computational Linguistics, 2006.

[10] Nitish Srivastava and Ruslan Salakhutdinov. Learning representations for multimodal data with deep belief nets. In *International conference on machine learning workshop*, volume 79, 2012.

[11] Mihai Gurban, Jean-Philippe Thiran, Thomas Drugman, and Thierry Dutoit. Dynamic modality weighting for multi-stream hmms inaudio-visual speech recognition. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 237–240. ACM, 2008.

[12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[14] Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large scale system for text detection and recognition in images. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 71–79. ACM, 2018.

[15] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. *CoRR*, abs/1505.00468, 2015.

[16] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. *CoRR*, abs/1505.01121, 2015.

[17] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *CoRR*, abs/1410.0210, 2014.

[18] K. Kafle and C. Kanan. Answer-type prediction for visual question answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4976–4984, June 2016.

[19] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. Stacked attention networks for image question answering. *CoRR*, abs/1511.02274, 2015.

[20] Duy-Kien Nguyen and Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. *CoRR*, abs/1804.00775, 2018.

[21] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *In EMNLP*, 2014.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[25] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[26] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016.

[27] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[28] Yu Jiang*, Vivek Natarajan*, Xinlei Chen*, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0.1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018.

[29] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[30] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017.