11-777 Report 2: Related Work and Model Proposal

Hao Wu[*] **Jiayi Shen**[*] **Yanlin Feng**[*] **Yinghuan Zhang**[*] **Yuwei Wu**[*]
{haowu3, jiayis2, yanlinf, yinghuan, yuweiwu}@andrew.cmu.edu

1 Related Work and Background

1.1 Related Datasets

TextVQA, TextVQA-X, and TextCaps TextVQA[1] contains 28,408 images from OpenImages and 45,336 questions that require reading and reasoning about text in images. Each question comes with 10 ground truth answers. TextVQA-X (Rao et al., 2021) contains 11,681 images and 15,374 questions from TextVQA. For each question, up to 5 distinct human annotators provide visual and textual explanations for why a given answer is correct. TextCaps (Sidorov et al., 2020) is another dataset that is built upon TextVQA. 5 captions are collected for each image in the TextVQA dataset.

ST-VQA Created with a similar purpose as that of TextVQA, ST-VQA (Biten et al., 2019) comprises 23,038 images sourced from six datasets and 31,791 questions that can be unambiguously answered using text in the image. (In contrast, 39% of the answers in TextVQA do not contain OCR tokens.) Each question comes with up to 2 ground truth answers.

VQA v2.0 As an updated version of the first large-scale VQA dataset, VQA v2.0[2] contains 265,016 images and abstract scenes, 204,721 of which are COCO images. While there are a total of 1,105,904 questions, only 8k (or less than 1%) of these questions require reading text in the image (Biten et al., 2019). Each question comes with 10 ground truth answers.

OCR-CC Another large-scale dataset worth mentioning is the OCR-CC dataset (Yang et al., 2021). It contains 1.367 million scene text-related image-caption pairs from the Conceptual Captions dataset. The scene text detected per image has a mean and median of 11.4 and 6, compared with 23.1 and 12 in TextVQA, and 8.03 and 6 in ST-VQA.

VisualMRC Recently Tanaka et al. (2021) proposed the VisualMRC dataset which is a machine reading comprehension dataset based on document images. VisualMRC contains long abstractive answers which do not correspond to spans in the documents. Images in VisualMRC are sourced from multiple domains which makes it more challenging than DocVQA.

1.2 Unimodal Baselines

Question When only using the question module and predicting from the 8000 most frequent answers, LoRRA (Singh et al., 2019) achieves validation and test accuracies of 8.09% and 8.70%.

Image When only using the image module and predicting from the 8000 most frequent answers, LoRRA achieves validation and test accuracies of 6.29% and 5.58%.

OCR LoRRA uses Rosetta-ml to produce OCR tokens. By predicting a random OCR token present in an image, validation and test accuracies of 7.72% and 9.12% are achieved. By predicting the most frequently occurring OCR token in an image, validation and test accuracies of 9.76% and 11.60% are achieved.

1.3 Relevant techniques

1.3.1 Feature extraction

Question Embedding Pretrained language models (Devlin et al., 2018) have become the dominant text encoders in recent work. These models are trained on massive amount of unlabeled text by minimizing unidirectional or bidirectional language modeling loss and later finetuned on specific domains.

[*] Everyone Contributed Equally – Alphabetical order
[1] See https://textvqa.org/ for details.
[2] See https://visualqa.org/ for details.

Regional Image Feature Faster R-CNN model is the dominant approach used in VQA related tasks to extract region-based object features, including visual information (convolutional features), positional information (bounding box coordinates) and class labels of detected objects. Apart from feature information on detected objects, there are papers such as Gao et al. (2020) that extract grid-based features as visual embedding, which mainly relies on ResNet and its variants to learn image representation as 2048-D vectors corresponding to 196 grids.

Whole Image Feature Rao et al. (2021) uses Feature Pyramid Network (FPN) to construct a semantic segmentation of the image and obtain visual explanations. In addition, although not mentioned in the VQA papers covered in this literature review, depth map is another potentially meaningful representation of the whole image besides RGB information, towards which Dense Prediction Transformer (DPT) (Ranftl et al. (2021)) is the state-of-the-art approach to estimate fine-grained depth information based on an architecture combining vision transformer encoder and convolutional decoder.

OCR System OCR methods used in prior work include Rosetta-en OCR, Rosetta-ml OCR, SBD-Trans OCR, and Google-OCR. While LoRRA (Singh et al., 2019) only uses FastText to extract word embedding from the OCR tokens, M4C and many later models employ a multi-feature representation, including appearance feature extracted using Faster R-CNN, character feature extracted using Pyramidal Histogram of Characters (PHOC), and a 4-dimensional location feature of the token's relative bounding box coordinates. In addition, Gao et al. (2021) introduced a 512D CNN feature called RecogCNN, which is extracted from text visual patches and trained on a text recognition task.

1.3.2 Multimodal Multitask Learning

There has been progresses made in the area of multimodal multi-task learning. Lu et al. (2020) jointly trained 12 vision-and-language tasks with a multi-task transformer based on VILBERT (Lu et al., 2019). Hu and Singh (2021) further expanded beyond fixed input modalities and jointly handled different single modal and multimodal tasks with a unified transformer model. They also pointed out the multimodal tasks such as VQA benefit from multi-task traning with uni-modal tasks.

Besides jointly training on tasks from different domains, several works designed a variety of highly correlated tasks to enhance the performance of VQA models.

Bounding Box Prediction Han et al. (2020) introduces a bounding box prediction task to prove its confidence of answer prediction. Specifically, when the model generates the answer by copying from OCR tokens, the IoU between the predicted bounding box and the ground truth OCR bounding box is calculated, serving as an evidence. The loss the IoU is added into training loss to urge the model to predict credible answer.

ANLS As Reward Average Normalized Levenshtein Similarity (ANLS) is another popular metric in TextVQA task. It is used to measure the similarity between predicted answer and ground truth answer instead of binary comparison. It is common case that OCR system provides an incorrect token that model fail to get the exact correct answer due to the systematic error. ANLS guides the model to make the right prediction even when the predicted answer does not exactly match it ground truth. Many prior works(Zhu et al., 2020; Liu et al., 2020) introduced ANLS as a reward into the training loss.

1.3.3 Attention mechanisms

Graph Attention Networks (GAT)(Veličković et al., 2017) is broadly used in prior works to encode visual relationship between objects, which has proven to be crucial to many computer vision tasks. ReGAT(Li et al., 2019a) introduces a Relation-aware GAT to model multi-type inter-object relations, including positional relation, semantic interactions and implicit relations. SA-M4C(Kant et al., 2020) improves M4C by introducing spatially aware self-attention layer where objects attend each other in a spatial graph.

Visual relation is important, though, TextVQA task requires a better understanding of relationship across multiple modalities. Many previous works explored relations between objects and OCR under the question text supervision. CRN(Liu et al., 2020) feeds question text features, OCR text features, and visual features into the Progressive Attention Module in turn and update informative features gradually. SSBaseline(Zhu et al., 2020) encodes multimodal features with three attention blocks, in each block OCR visual features, OCR text features, and object features attended with question embedding respectively. SMA(Gao et al., 2021) uses a Question Conditioned Graph Attention Module to encoder the object-object, object-OCR, OCR-OCR relationships under the question's guidance.

1.3.4 Pre-training

Given that our target task requires reasoning over both text and images, our main focus here is the Vision-and-Language Pre-training (VLP). There are several strategies commonly used to train VLP models.

Image Text Matching This task requires the model to generate high-quality instance-level representations. Given a random pair of image and text descriptions, the model predicts whether the pair is matched. This task is widely used in a vartity of VLP models including ViLT, TAP, VISUALBERT, VILBERT, and LXMERT (Kim et al., 2021; Yang et al., 2021; Li et al., 2019b; Lu et al., 2019; Tan and Bansal, 2019). LayoutLM (Xu et al., 2020) further modified this task into Multi-label document classification task and used the document tags to supervise the document-level representation learning.

Masked Language Modeling Following the mechanisms proposed by Kenton and Toutanova (2019), this objective aims at predicting masked text tokens from the given contextualized vector and vectors corresponding to image regions. Kim et al. (2021) introduced whole word masking for MLM task to train VLP models, while Powalski et al. (2021) proposed to use T5-like salient span masking schema. Lu et al. (2019); Tan and Bansal (2019) extended this idea to masked multi-modal modeling task. In this manner, 15% of both words and image region inputs are masked and the model is required to reconstruct given the remaining inputs.

Other Cross-Modality Tasks Addition to the above prediction tasks, a few models also introduce several tasks that need strong cross-modality representations. Xu et al. (2021) proposed a text-image alignment task to encourage the model learn the alignment of detected objects among different modalities. Yang et al. (2021) designed a relative (spatial) position prediction (RPP) task. The RPP task aims to predict the relative spatial position between an object region and a scene text region. Tan and Bansal (2019) proposed image question answering tasks on the pretraining stage to further enhance the cross-modality representations.

1.3.5 Multireference

Rao et al. (2021) uses the sample one technique to leverage the multiple textual explanations collected for each question. In each training epoch, one of the available textual explanations is randomly selected.

1.3.6 Copy Mechanism and Pointer Network

Many sequence learning tasks requires *copying*, which refers to selectively replicating segments of the input to generate the output. For example, in a dialogue system, the agent needs to generate responds by referring to entities in the input utterance. Another example is text summarization, where the model is required to extract text from the original documents. Similar phenomenon is also observed in real-world language communication where humans tend to repeat long phrases in conversation.

Various pointer network (Vinyals et al., 2015) architectures have been proposed to address the challenge of *copying*. Gu et al. (2016) proposed CopyNet to address copying in seq2seq learning tasks. CopyNet generates an answer token at each timestep based on a mix of generation-mode probabilities and copy-mode probabilities, where the copy-mode probabilities are computed by attending to the input tokens. See et al. (2017) generalized CopyNet by modeling the selection between generation and copyping with a binary classifier.

Copying is also critical in text-related VQA tasks, where the answers often include OCR tokens in the image. LORRA (Singh et al., 2019) concatenates OCR tokens to the common word vocabulary and compute normalization over the new vocabulary. M4C (Hu et al., 2020) also augments transformer with a dynamic pointer network which computes OCR token probabilities with a bilinear layer.

1.3.7 Data Augmentation and Generalizability

Apart from designing efficient and robust model structures, there are papers targeting the generalization capabilities of VQA models and proposing methods on data augmentation, evaluation and training procedure.

Data Augmentation For image augmentation, in addition to techniques commonly used in computer vision community (rotation, scaling, adding noises, etc), which typically don't affect the semantic meaning of the visual content, VQA image augmentation also involves transformations that deliver significantly different visual information. For instance, Gokhale et al. (2020) has proposed to mutate images via removing object instances or inverting colors, which is critical enough to lead to different answers to the questions. It is worth noting that removing object instances requires additional procedures using inpainting network based on Generative Adversarial Network (GAN) to make transformed images photorealisitic.

As for questions augmentation, Gokhale et al. (2020) has adopted template-based questions operators to perform tasks such as negation on yes-no questions and substitution of critical words with antonyms. Moreover, Rosenberg et al. (2021) followed another template-based approach that converts "what/how" questions into "yes/no" questions.

Evaluation To measure and boost model performance on Out-Of-Distribution (OOD) datasets, several papers have come up with new metrics to evaluate generalizability. Rosenberg et al. (2021) has introduced Robustness to Augmented Data (RAD), which calculates the proportion of correct answers on augmented samples among all correct answers, and has demonstrated the indicating power of RAD on model robustness against unseen data through quantitative results. Moreover, Kervadec et al. (2020) group questions on frequencies and argue that accuracy over infrequent question-answer pairs is more suitable in delineating generalization capabilites.

Training Procedure The concept of grouping questions is not merely utilized in designing evaluation metrics, but adopted in the training procedure as well. For example, Gokhale et al. (2020) has partitioned training samples by question types and optimized a different copy of the model under each cluster, which has been proved to lead to better out-of-distribution generalization.

2 Model Proposal

2.1 Overall model structure

We present our overall model structure as shown in Fig.1. Given a question and an image as input, we firstly extract the OCR tokens and object by an off-the-shelf OCR system and detector. We separately obtain the uni-modal features of all four modalities, including question, whole image, detected objects, and OCR tokens, and use one multimodal encoder to learn the multimodal embedding.

2.2 Encoders

Describe encoders for each modality and at least one alternatives for each. Explain the relative strengths of each option (e.g. coverage, efficiency, ...)

Embedding of question words. For question words, we follow the previous work and use the first 3 layers of BERT-BASE, rather than all its 12 layers, in order to save computation. Given a question with W words, we embed it into a sequence of $d-$dimentional feature vector $\{x_i^{\text{ques}}\}_{i=1}^{W}$. Alternatively, we can also use other pretrained language model of larger size like T5-Large model to get better question representations.

Embedding of whole image. We introduce the entire image as a modality. We embed the entire image as a semantic segmentation map predicted by Feature Pyramid Network (FPN). Alternatively, the whole image can be represented as a depth map estimated by Dense Prediction Transformer (DPT), although semantic segmentation map may be able to encode more information and potentially advantageous in contributing to model performance.

Embedding of detected objects. Given an image, we obtain the region-based object features including visual features and bounding box features. These two parts features are extracted by an off-the-shelf Faster R-CNN. Visual features of M objects are embedded into a sequence of same $d-$dimentional feature vector $\{x_{v,i}^{\text{obj}}\}_{i=1}^{M}$ and another sequence of 4-dimensional location feature vector $\{x_{bb,i}^{\text{obj}}\}_{i=1}^{M}$ is used for bounding box location feature embedding. Alternatively, as implemented in Gao et al. (2020), a pretrained ResNet model can be used to extract grid-based image feature as a list of 196 $d-$dimentional vectors corresponding.

In addition, we want to incorporate detected object labels. We use the same embedder as question texts to extract the label features of objects. A sequence of $d-$dimension feature vector $\{x_{label,i}^{\text{obj}}\}_{i=1}^{M}$ is expected.

Embedding of OCR tokens For OCR tokens, we also plan to follow M4C and extract tokens using Rosetta-en. Specifically, we plan to use FastText for word embedding, Faster R-CNN for appearance feature, a 604-dimensional Pyramidal Histogram of Characters (PHOC) for character feature, and a 4-dimensional location feature of the token's relative bounding box coordinates. Alternatively, we

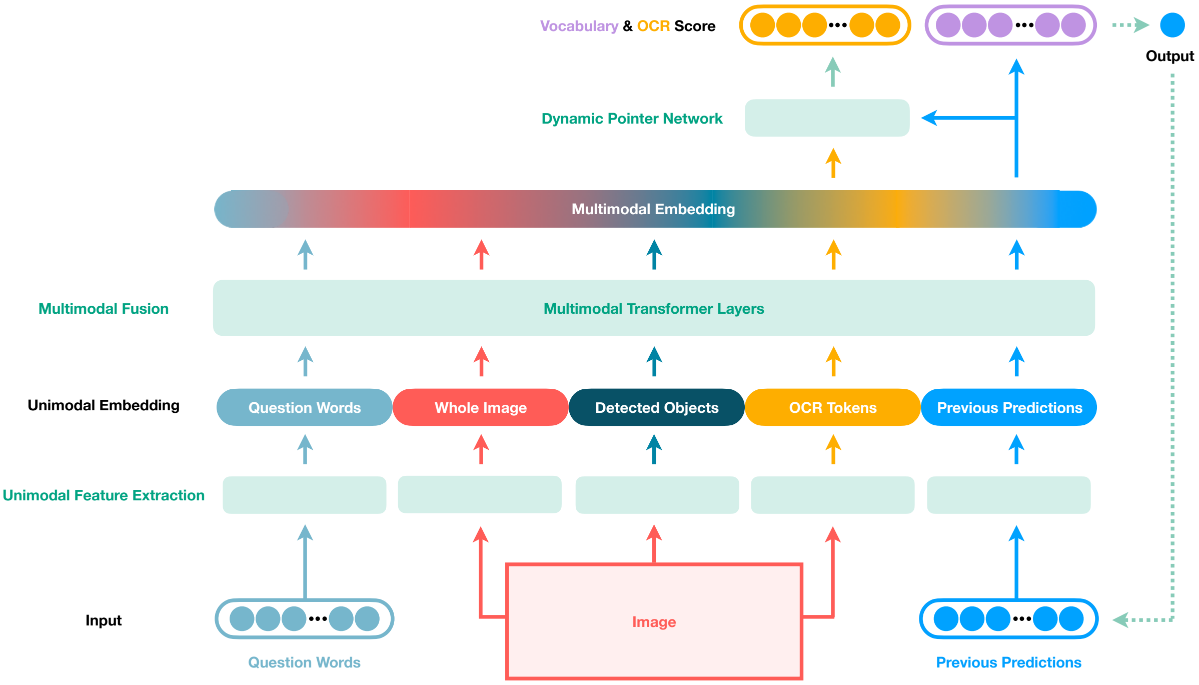

Figure 1: Overall Model Structure

can use pretrained encoder like LayoutLM to encode the text, position and image features of OCR tokens. In this manner, the embedding of OCR tokens is jointly optimized compared with combining features from different modules like in M4C.

Multimodal Encoder After obtaining embedding features of these four modalities, we use a typical encoder with $L-$layers stacked transformer layers as our multimodal encoder.

2.3 Decoders

We predict an answer with variable length based on M4C's dynamic network. Our multimodal encoder fuses features coming from four modalities, *i.e.*, question words, whole image, detected objects, and OCR tokens, and outputs a multimodal feature vector as mentioned. In the decoding, we use exactly the same transformer layers as the multimodal encoder. A dynamic network takes this multimodal embedding as input to iteratively generate the answer. We construct a fixed vocabulary list of V words which frequently appear in the training set answers and a dynamic case-ware vocabulary list of N OCR tokens detected in the test image. The whole answer space at each inference is composed with these two vocabulary lists and additional two special tokens, `<begin>` and `<end>`.

At each time step, our decoder predicts one word from the case-aware answer space. The prediction begins with `<begin>` at time step 0 which only takes multimodal embeddings as the input. For time step t, we feed the decoder both the multimodal embedding and the embedding of its previous prediction, and the next answer word coming from the case-aware answer space is generated based on the decoder output with a dynamic pointer network. We stop the decoding process after the decoder predict another special token `<end>`.

2.4 Pretraining

According to Yang et al. (2021), pretraining on the same downstream task datasets, their system achieved an improvement of the absolute accuracy on the TextVQA dataset by +5.4%. Inspired by this, we would like to implement several pretraining strategies over the TextVQA dataset to enhance our system's performance.

Masked Language Modeling Inspired by Kim et al. (2021); Powalski et al. (2021), we would like to use all input OCR tokens to perform the MLM tasks and in a salient whole word masking schema i.e. the whole words for named entities are preferred rather than random tokens . By masking these tokens, we encourage the system to predict meaningful entities from both the whole image and detected object features.

Masked Image Reconstruction Similar to Lu et al. (2019), we would like to mask 15% of the image region inputs and task the model with reconstructing the maksed image given the remaining inputs. The masked image regions are selected in the regions for detected objects and scene texts from the whole image.

Image text matching We would also like to implement the commonly used ITM task to help the model learn the correspondence between image and textual content. For ITM task, we would like to explore three different ways of constructing negative sample. The first one is to follow Yang et al. (2021)'s way of polluting text words. The second way is to randomly replace detected objects while the last way is to randomly replace the whole image with ones from other training pairs.

Masked Box Position Prediction In order to enhance the learning of relative positions among detected objects and OCR scene texts, we proposed a novel pretraining strategy called **masked box position prediction**. As shown in the figure 3, we would randomly mask some of the position for detected objects and OCR scene texts. The system is required to predict the original coordinate for the masked positions. In this manner, we encourage the model to learn the relative position around the masked detection box.

Image Text Alignment This is another cross-modality pretraining task we would like to explore. In order to properly answer the questions in VQA dataset, it is essential to learn the alignment between OCR texts and whole image, the alignment between detected objects and whole image. So we proposed a new alignment task: **detection box mask prediction**. For detection box mask prediction, as illustrated in the figure 2, the image is processed in the same manner as in the masked image reconstruction task. The system is then required to predict whether each of the input OCR tokens and detected objects is masked in the whole image.

2.5 Detecor/OCR-Robust Finetuning by Masking

While all TextVQA questions can be answer solely with the raw image, almost all models that have been proposed rely on pretrained object detectors and off-the-shelf OCR toolkits to provide information that is more fine-grain than the raw image. Such a strong dependency might hurt model performance if (1) the dataset domain differs significantly from the domain that the detector is trained on (2) the OCR recognizer produces poor results due to rotation or noises. Therefore, we propose to make our model more robust with respect to both the object detector and the OCR recognizer with a modality-specific masking strategy.

As illustrated in Figure 4, the input image is represented in three different modalities: the raw image, objects as well as OCR tokens. The objects (and the bounding box positions) are predicted by a pretrained detector like RCNN and the OCR tokens are extracted by an OCR recognizer. In textVQA, the question can *always* be answered with only the raw image, while in most cases, can be answered with either of the other two modalities, given that the outputs from detector/OCR outputs are correct. [3] Therefore, to avoid overfitting to objects or OCR tokens and thus make our model more robust w.r.t. to detector/OCR, we propose the following masking strategy during finetuning.

At each iteration, we randomly select one or two of the three modalities. For the selected modalities, we randomly mask $x\%$ of the features by setting them to zeros. The model is required to predict the correct answer from the masked features. We finetune the model by minimizing the cross-entropy loss as in previous work. We believe this masking strategy will encourage model to utilize information in all three modalities thus reducing the risk of overfitting. Specifically, suppose the answer is one of the OCR tokens, masking it out will force the model to utilize information in the raw image to answer the question. As as result, our model will be more robust when the object detector or OCR recognizer behaves poorly.

2.6 Loss Functions

During finetuning on the TextVQA dataset, we train our model by minimizing the cross-entropy loss. The loss is calculated between the predicted answer and the ground-truth answering. As described in Sec. 2.5, we also incorporate a modality-specific masking strategy to improve robustness with respect to the object detector and OCR recog-

[3] Note that some questions (e.g. clock reading questions) cannot be answered with only OCR even if the OCR results are correct.

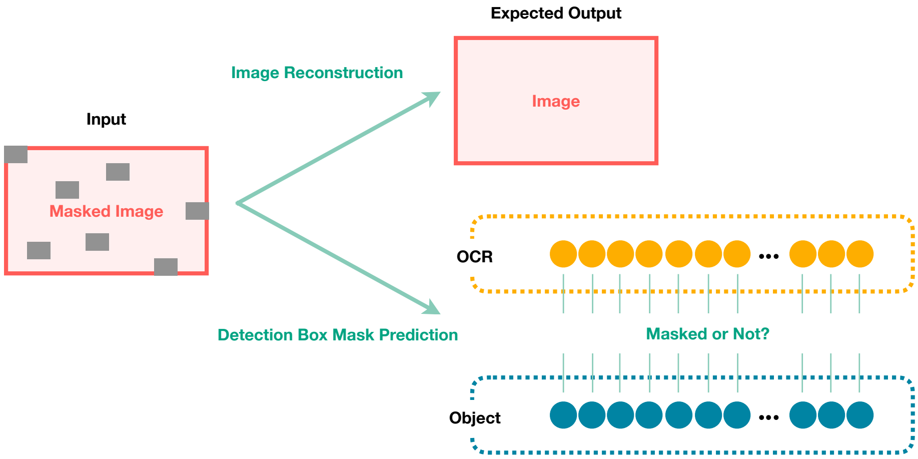

Figure 2: Pretraining: Masked Image Reconstruction Detection Box Mask Prediction.

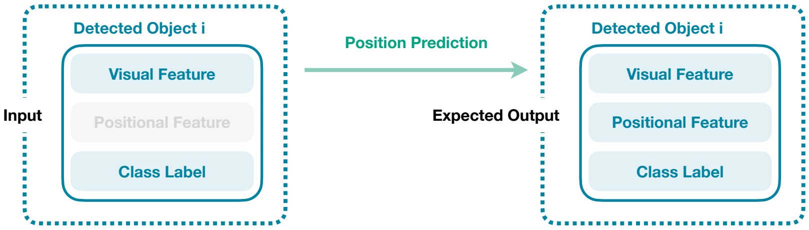

Figure 3: Pretraining: Position Prediction.

nizer. Our loss function can be formally written as follows:

$$\mathcal{J}_{\text{finetune}} = \mathbb{E}_{s \subset \{\text{img, obj, ocr}\}} \sum_{\mathbf{x},\mathbf{y}} \mathcal{L}(\texttt{Mask}(\mathbf{x}, s), \mathbf{y}) \tag{1}$$

where $\mathcal{L}$ denotes the cross-entropy loss, s is the set of modalities that are selected and `Mask` denotes our modality-specific masking operation.

As described in Sec. 2.4, we also propose five auxiliary losses during unsupervised pretraining, including mask language modeling, masked image reconstruction, image text matching, masked box position prediction and image text alignment:

$$\begin{aligned}\mathcal{J}_{\text{pretrain}} = \sum_{\mathbf{x}} \mathcal{J}_{\text{MLM}}(\mathbf{x}) + \mathcal{J}_{\text{MIR}}(\mathbf{x}) + \\ \mathcal{J}_{\text{ITM}}(\mathbf{x}) + \mathcal{J}_{\text{MBPP}}(\mathbf{x}) + \mathcal{J}_{\text{ITA}}(\mathbf{x})\end{aligned} \tag{2}$$

We believe these auxiliary losses encourage our model to learn better multimodal representation that encode useful information.

2.7 Data Augmentation

2.7.1 Image Augmentation

To build up model robustness against the heterogeneity of input images, training images will be augmented through rotation, scale, adding noises and color editing. Due to the scope of this project, we focus on transformations that do not affect answers, and leave more in-depth image augmentations that require answer changes and photorealistic editing as future work)

Moreover, positional information of detection boxes extracted from OCR modules and unimodal models such as Faster R-CNN will be perturbed by altering coordinates, which will serve to construct model robustness against varying performance of unimodal models and OCR modules.

2.7.2 Localization-aware QA Enrichment

Visual relations are proven to be essential part in either uni-modal visual tasks or multi-modal tasks. Many prior works use GAT to learn the spatial relations between objects in the images. However, such relation-aware GAT is not learned across different modalities. Though model is able to learn

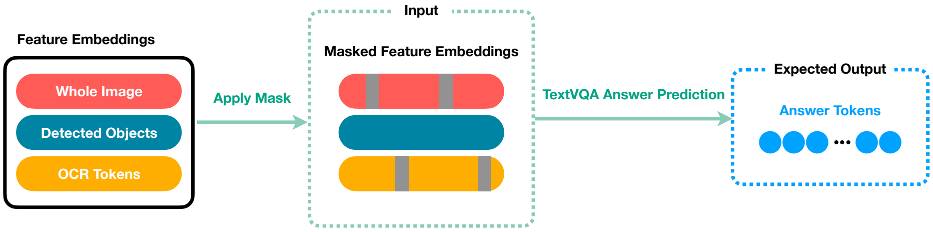

Figure 4: Detecor/OCR-robust finetuning by modality-specific masking

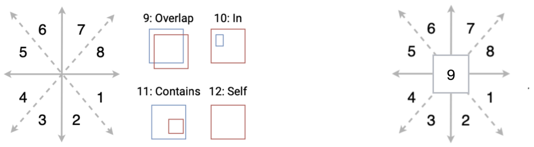

Figure 5: Positional Relationships. (left) Space of potential positional relationship between object/OCR and object/OCR. (right) Space of potential positional relationship between object/OCR and whole image.

the spatial relationship well in visual modality, it is hard to respond to the question correctly. Inspired by the idea of implicit relation, several researchers utilizes the vanilla attention mechanism to explore the potential implicit relationship between object, OCR, and questions, but the cross-modality learning happen after the uni-modal embedding, which may compromise the cross-modality learning of relationships.

We propose a data augmentation method called Localization-aware QA Enrichment. Following SA-M4C(Kant et al., 2020), we defines 12 positional relations which can happen between `<object, object>`, `<object, OCR>`, and `<OCR, OCR>`, shown as Fig.5(left). Since we introduce whole image as a new modality, we also define 9 positional relations between `<object, IMG>` or `<OCR, iMG>`, shown as Fig.5(right). We design two templates of questions respectively:

1. Template-1: Is A [relative position] to B?

2. Template-2: Is C [relative position] to the whole image?

By adding these two types of localization-aware question-answer data, we urge the model to learn positional relationship between different modalities under the natural language supervision.

2.8 Evaluation Metrics

Accuracy: the accuracy score between predicted answers and ground truth labels.

Average Normalized Levenshtein Similarity (ANLS): the accuracy metric awards a zero score even when the prediction is only a little different from the target answer. Since no OCR is perfect, Mathew et al. (2021) proposed the metric called ANLS for DocVQA task. Motivated by their work, we would also like to include this metric in our evaluation.

Robustness on Augmented Data (RAD): the proportion of correctly predicted augmented questions among all correctly predicted questions. As demonstrated by Rosenberg et al. (2021), RAD is a powerful indicator of model robustness on OOD data. Therefore, we would like to include this as a metric of generalization capabilities.

3 Team member contributions

Hao Wu contributed to relevant techniques (data augmentation generalizability), model structure diagrams, encoders, image augmentation, and evaluation metrics.

Jiayi Shen contributed to related datasets, unimodal baseline analysis, relevant techniques (pretraining, multireference), and encoders.

Yanlin Feng contributed to unimodal baseline analysis, relevant techniques (answer generation), model fine-tuning, and loss function.

Yinghuan Zhang contributed to relevant techniques (multitasks, attention mechanisms), model structure description, encoders, decoders, and QA enrichment.

Yuwei Wu contributed to relevant techniques (multitasks, pretraining), encoders, and model pretraining.

References

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Chenyu Gao, Qi Zhu, Peng Wang, Hui Li, Yuliang Liu, Anton Van den Hengel, and Qi Wu. 2021. Structured multimodal attentions for textvqa. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Difei Gao, Ke Li, Ruiping Wang, Shiguang Shan, and Xilin Chen. 2020. Multi-modal graph neural network for joint reasoning on vision and scene text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12746–12756.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. MUTANT: A training paradigm for out-of-distribution generalization in visual question answering. *CoRR*, abs/2009.08566.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*.

Wei Han, Hantao Huang, and Tao Han. 2020. Finding the evidence: Localization-aware answer prediction for text visual question answering. *arXiv preprint arXiv:2010.02582*.

Ronghang Hu and Amanpreet Singh. 2021. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1439–1449.

Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. 2020. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9992–10002.

Yash Kant, Dhruv Batra, Peter Anderson, Alexander Schwing, Devi Parikh, Jiasen Lu, and Harsh Agrawal. 2020. Spatially aware multimodal transformers for textvqa. In *European Conference on Computer Vision*, pages 715–732. Springer.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. 2020. Roses are red, violets are blue... but should vqa expect them to? *CoRR*, abs/2006.05121.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019a. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10313–10322.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019b. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Fen Liu, Guanghui Xu, Qi Wu, Qing Du, Wei Jia, and Mingkui Tan. 2020. Cascade reasoning network for text-based visual question answering. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4060–4069.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2200–2209.

Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. 2021. Going full-tilt boogie on document understanding with text-image-layout transformer. In *International Conference on Document Analysis and Recognition*, pages 732–747. Springer.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. 2021. Vision transformers for dense prediction. *CoRR*, abs/2103.13413.

Varun Nagaraj Rao, Xingjian Zhen, Karen Hovsepian, and Mingwei Shen. 2021. A first look: Towards explainable textvqa models via visual and textual explanations. *arXiv preprint arXiv:2105.02626*.

Daniel Rosenberg, Itai Gat, Amir Feder, and Roi Reichart. 2021. Are VQA systems rad? measuring robustness to augmented data with focused interventions. *CoRR*, abs/2106.04484.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *European Conference on Computer Vision*, pages 742–758. Springer.

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

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. Visualmrc: Machine reading comprehension on document images. *arXiv preprint arXiv:2101.11272*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. *Advances in neural information processing systems*, 28.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2021. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200.

Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. 2021. Tap: Text-aware pre-training for text-vqa and text-caption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8751–8761.

Qi Zhu, Chenyu Gao, Peng Wang, and Qi Wu. 2020. Simple is not easy: A simple strong baseline for textvqa and textcaps. *arXiv preprint arXiv:2012.05153*, 2.