Finetuning Language Models for Multimodal Question Answering

Xin Zhang*
Harbin Institute of Technology,
Shenzhen
zhangxin2023@stu.hit.edu.cn

Wen Xie*
Harbin Institute of Technology,
Shenzhen
xiewen354@gmail.com

Ziqi Dai*
Harbin Institute of Technology,
Shenzhen
ziqidai2024@gmail.com

Jun Rao
Harbin Institute of Technology,
Shenzhen
rao7jun@gmail.com

Haokun Wen
Harbin Institute of Technology,
Shenzhen
whenhaokun@gmail.com

Xuan Luo
Harbin Institute of Technology,
Shenzhen
gracexluo@hotmail.com

Meishan Zhang†
Harbin Institute of Technology,
Shenzhen
zhangmeishan@hit.edu.cn

Min Zhang
Harbin Institute of Technology,
Shenzhen
zhangmin2021@hit.edu.cn

ABSTRACT

To achieve multi-modal intelligence, AI must be able to process and respond to inputs from multimodal sources. However, many current question answering models are limited to specific types of answers, such as yes/no and number, and require additional human assessments. Recently, Visual-Text Question Answering (VQTA) dataset has been proposed to fix this gap. In this paper, we conduct an exhaustive analysis and exploration of this task. Specifically, we implement a T5-based multi-modal generative network that overcomes the limitations of traditional labeling space and provides more freedom in responses. Our approach achieve the best performance in both English and Chinese tracks in the VTQA challenge.

CCS CONCEPTS

• **Information systems** → **Question answering**; • **Computing methodologies** → *Computer vision representations.*

KEYWORDS

Visual Question Answering, Multi-modal Fusion, T5 Finetuning

ACM Reference Format:
Xin Zhang, Wen Xie, Ziqi Dai, Jun Rao, Haokun Wen, Xuan Luo, Meishan Zhang, and Min Zhang. 2023. Finetuning Language Models for Multimodal Question Answering . In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3581783.3612837

1 INTRODUCTION

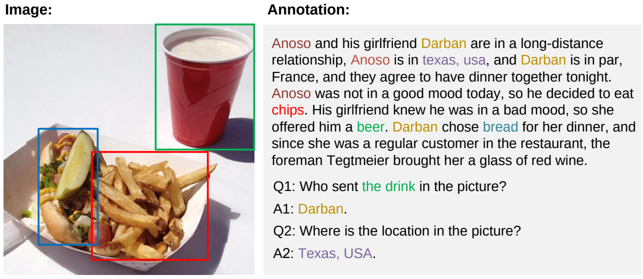
Multi-modal research has long aimed to achieve an understanding and reasoning that spans different modes of communication. Recent benchmark tasks such as image-text retrieval [19, 26], visual question answering [9], phrase grounding [16] and visual commonsense reasoning [30] focus on models' ability to comprehend multiple ways. This requires models to align semantic information from visual and textual content [6, 22] to answer questions. While previous benchmarks have focused on solving simple common sense tasks such as color, counting, and spatial relationships, newer datasets like OK-VQA [17], Science-QA [15], and VCR [30] require models to possess general knowledge, such as historical events and sequential logic of events. These datasets need models to store more text-related common sense knowledge rather than relying solely on connections between images and text. As a result, existing models may overfit text-related common sense knowledge at the expense of visual content.

Existing VQA models typically require predefined labels of answer space, e.g., VQA V2 [9] have 1,000 classes for answering. This limits the answers to the questions, reducing the difficulty of answering the model and making it more susceptible to bias [12] in the dataset. Such problems limit further development in the field of VQA. Unlike the previous dataset, the newly proposed Visual-Text Question Answering [4] challenge focuses on more open answers, i.e., the model needs to answer the corresponding appropriate open-ended answer based on a deeper multi-modal understanding rather than searching through a given pattern of answers. An example from the competition website is shown in Figure 1. It involves image input, text segment descriptions, multiple questions and answers. The goal of this task is to achieve three objectives: 1) Recognize entities in both images and text relevant to the question. 2) Synchronize multimedia representations of the same entity. 3) Carry out multi-step reasoning between text and image to provide an open-ended answer.

In this paper, we implement a unified framework to tackle the challenging VTQA task, which includes a multi-modal encoder to process the multi-modal input, *i.e.*, the image, the text description, and the question, and a text decoder to generate the answer in the

[*] Authors contributed equally.
[†] Corresponding author: Meishan Zhang.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612837



Anoso and his girlfriend Darban are in a long-distance relationship, Anoso is in texas, usa, and Darban is in par, France, and they agree to have dinner together tonight. Anoso was not in a good mood today, so he decided to eat chips. His girlfriend knew he was in a bad mood, so she offered him a beer. Darban chose bread for her dinner, and since she was a regular customer in the restaurant, the foreman Tegtmeier brought her a glass of red wine.

Q1: Who sent the drink in the picture?

A1: Darban.

Q2: Where is the location in the picture?

A2: Texas, USA.

Figure 1: An example of VTQA task.

open-ended form. Specifically, we first analyze the statistical information of the VQTA dataset, such as text length, response type, etc. We use corresponding models on different tracks (Chinese and English) to address the multilingualism problem. We also use a 2-size model to show more results. The results show that larger models and the appropriate language model can achieve good results on the competition list, i.e., a score increase of 1.0+. Experiments show that our framework can greatly detect the entities from the image given the question and use the multi-modal data to answer the entities-related questions generatively. We offer in-depth analysis and ablation studies on the visual feature and task formulation, i.e., generation v.s. classification.

In summary our contribution is as follows:

- We implement a generative multi-modal network as a solution for the VQTA challenge to achieve a first-place finish.
- Extensive experiments and analyses are carried out to verify that our system could extract or generate answers with some relevance based on the content provided.

2 RELATED WORK

Our work is closely in line with the studies on **visual question answering**. Conventional methods generally cast VQA as a **classification task**, where the image and question are first respectively encoded by the visual encoder and the textual encoder, and then various multimodal fusion methods [5, 21] are exploited to derive the fused multimodal features. Finally, the fused multimodal features are further processed by a classifier to generate the candidate answers [1, 3]. For example, Yang *et al.* [28] utilized VGGNet [24] and LSTM [11] to encode the image and question, respectively. Then an attention mechanism is designed to fuse the multimodal features, followed by a classifier to generate the answers. Moreover, Anderson *et al.* [2] resorted to Faster R-CNN [23] and GRU [7] as the visual encoder and the textual encoder, respectively. Then a novel combined bottom-up and top-down visual attention mechanism is employed to derive the multimodal features. Thereafter, Ye *et al.* [29] adopted Faster R-CNN and LSTM to encode the image and question, respectively, and resorted to the self-attention mechanism [25] to model the multimodal intra- and inter-modal interactions. Similarly, the fused features are processed by a classifier to derive the answer. Nowadays, SoTA pretraining methods like ALBEF [13] adopts knowledge distillation [10, 20] to enhance multimodal understanding.

Although these methods have made prominent progress, they can only generate the answers based on the pre-defined answer candidate in a discriminative manner, which limits the flexibility of the answer and restricts the real-world application scenarios. On the contrary, in this work, we aimed to generate open-ended answers in the natural language form, which is more challenging and valuable.

3 METHOD

In this section, we first formulate the research problem and subsequently detail the proposed T5 based approach.

3.1 Problem Formulation

In this work, we aim to tackle the challenging VTQA task, which can be formally defined as given an image-text pair, the goal is to answer a question in an open-ended form. Suppose we have a set of quadruples, denoted as $\mathcal{D} = \left\{(v, t, q, a)_i\right\}_{i=1}^{N}$, where v, t, q, and a refer to the image, text description, question, and answer, respectively. N is the total number of quadruples. Based on $\mathcal{D}$, we aim to optimize a function that can answer the question q based on the image-text pair (v, t) correctly. This can be formally defined as follows,

$$\mathcal{F}\left(q, v, t | \Theta\right) \rightarrow a, \tag{1}$$

where Θ are the to-be-optimized parameters.

3.2 Model

Our model is essentially a T5 [18] model with visual embedding inputs from Faster RCNN [23]. Readers with sufficient background knowledge can skip the following methodology section and proceed directly to the experiments (section 5).

The proposed approach consists of two key modules: (a) multimodal encoder and (b) text decoder. The former aims to encode the multi-modal input, *i.e.*, the image, the text description, and the question into hidden states (detailed in Section 3.2.1). Based on the hidden states, the latter is responsible for generating the answer (described in Section 3.2.2).

3.2.1 ***Multimodal Encoder.*** For the visual image, we utilize a frozen Faster RCNN [23] as the vision encoder to extraction region features and a learnable linear layer to transform the features to match the text input embedding dimension. Formally, we have,

$$\mathbf{E}^v = \mathbf{W} \cdot \text{FasterRCNN}\left(x\right) + \mathbf{b}, \tag{2}$$

where $\mathbf{E}^v = \left[\mathbf{e}_1^v, \mathbf{e}_2^v, \cdots, \mathbf{e}_L^v\right] \in \mathbb{R}^{L \times D}$ is the visual embedding. L is the number of image tokens (regions) and D denotes the feature dimension.

As for the text description and question, which are both in textual modality, we first tokenize them into standard vocabularies, and then apply an embedding layer to derive the textual embeddings, which are denoted as follows,

$$\begin{cases} \mathbf{E}^t = \left[\mathbf{e}_1^t, \mathbf{e}_2^t, \cdots, \mathbf{e}_M^t\right], \\ \mathbf{E}^q = \left[\mathbf{e}_1^q, \mathbf{e}_2^q, \cdots, \mathbf{e}_N^q\right], \end{cases} \tag{3}$$

where $\mathbf{E}^t \in \mathbb{R}^{M \times D}$ and $\mathbf{E}^q \in \mathbb{R}^{N \times D}$ refer to the text description embedding and question embedding, respectively. M and N denote the length of the text, respectively. Note that to maintain the relation information, we also add the positional embeddings to the textual embeddings.

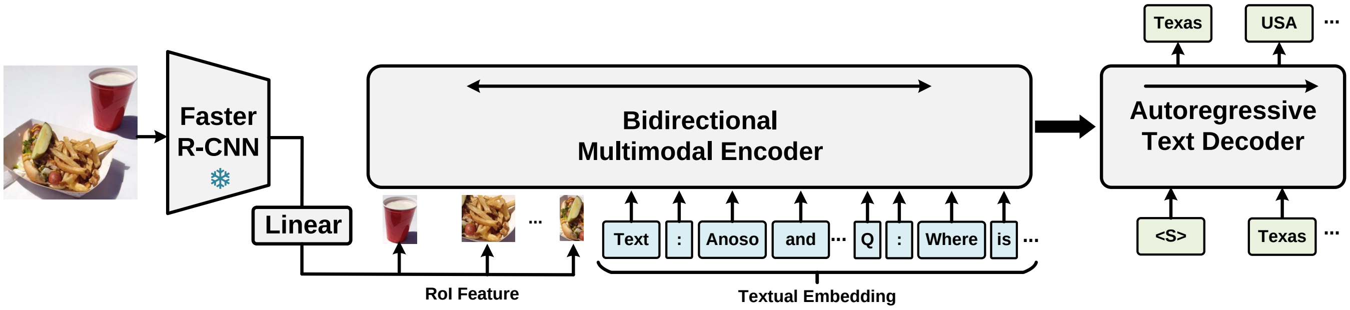


Figure 2: An illustration of our MGN architectures for VQTA task.

Table 1: Dataset statistics. Length by T5 tokenizer.

Split	#Image-text	#question	MAX / MIN / AVG text length	#answer type: YN / E / G
Train	4200	11312	1099 / 75 / 383.5	3931 / 6930 / 451
Val	469	1245	739 / 73 / 398.9	436 / 763 / 46
Test_dev	842	2189	950 / 81 / 328.3	-
Test	-	9035	-	-

Table 2: Text length statistics of VTQA dataset by BERT tokenizer.

	description			Question			Answer		
len	avg	max	min	avg	max	min	avg	max	min
en	187	914	52	11	34	3	2	87	1
zh	239	949	91	14	49	5	3	107	1

Thereafter, we concatenate the visual and textual embeddings to derive $\overline{\mathbf{E}} = \left[\mathbf{E}^{v}, \mathbf{E}^{t}, \mathbf{E}^{q}\right]$, which is processed by off-the-shelf Transformer to derive the multi-modal fusion hidden states. Formally, we have,

$$\mathbf{E} = \mathrm{Transformer}_{\mathrm{E}}\left(\overline{\mathbf{E}}\right). \tag{4}$$

3.2.2 ***Text Decoder.*** Based on the hidden states, the text encoder can generate the answer in an auto-regressive manner. Specifically, we also employ the Transformer architecture in this module, which iteratively attends to previously generated token via self-attention and the encoder outputs via cross-attention. Specifically, we train the model parameters Θ by minimizing the negative log-likelihood of ground-truth answer a tokens given image v, text description t, and question q as follows,

$$\mathcal{L}_{\Theta} = -\sum_{j=1}^{|a|} logP_{\Theta}\left(a_j | a_{<j}, v, t, q\right). \tag{5}$$

4 DATASET

4.1 Overview

The VTQA dataset consists of 10124 image-text pairs and 23,781 questions. The images are real images from MSCOCO dataset, containing a variety of entities (Figure 1 demonstrates a data instance). For each questions, there are three types of answers: 1) yes/no, 2) extracted from the image description text, and 3) open-ended answer that need generate new text.

We count the information on the length of the text in the data set, using BERT or T5 tokenizer. As Table 2 shown, although the average text length is within 300, there are still some texts with 500+ tokens that need to be considered.

4.2 Metric

As the answers divided into three types, the VTQA task sets different metrics for the three types of answers.

Exact match (EM). This metric measures the percentage of model predictions that match the ground truth answer exactly (string exact match, with fuzzy match for synonyms of yes/no). It is used in all types of answers.

(Macro-averaged) F1 score (F1). This metric measures the average overlap between the prediction sentence and ground truth answer. The prediction and ground truth are treated as bags of tokens to compute their F1. This metric will be used for the 'E' and 'G' types of answers.

YN accuracy (YNAcc). This metric is only used for the 'YN' type of answers. The answer will be transformed into 'yes' or 'no' by a pre-defined yes-or-no dictionary from the organizers.

In the VTQA competition, the EM metric is used for ranking models. The other metrics are presented to show the performance of the algorithm.

4.3 Dataset Analysis

Dataset statistics are shown in Table 1 & 2. Based on the answer type statistics in Table 1, we can observe that the majority of the answers in the VTQA dataset are of "Extraction" type, which indicates that the questions are focused on identifying objects, attributes, or relationships in the image descriptions. The "Yes/No" type answers are the second most frequent, indicating that there are also questions that require binary answers. The "Generation" type answers are relatively infrequent, suggesting that questions that require new texts are less common in this dataset.

5 EXPERIMENT

5.1 Setup

We implement our model based-on the T5 [18] of transformers [27]. For the base-sized pre-trained text model, we adopt the `t5-base` and `mengzi-t5-base` [32] for the English track and Chinese track,

Table 3: Main results from the competition leaderboard.

Model	Dev				Test			
	EM	YN-acc	E-F1	G-F1	EM	YN-acc	E-F1	G-F1
English								
Baseline	0.5861	0.7509	0.6842	0.4938	-	-	-	-
Our-t5-base	0.6702	0.7902	0.8165	0.5216	0.6469	0.7493	0.8199	0.4718
Our-t5-large	0.6825	0.8098	0.8226	0.5445	0.6638	0.8055	0.8182	0.4681
Chinese								
Baseline	0.4884	0.7497	0.5766	0.4978	-	-	-	-
Our-t5-base	0.5916	0.7387	0.7688	0.5808	0.5967	0.7341	0.7717	0.5155
Our-t5-large	0.6240	0.7669	0.7993	0.5922	0.6127	0.7518	0.7957	0.5510

Table 4: Validation set exact match scores of different task settings.

Model	English				Chinese			
	Overall	YN	E	G	Overall	YN	E	G
Seq-to-seq w/o image features								
t5-base	0.6088	0.7156	0.5780	0.1087	0.5325	0.5046	0.5662	0.2391
t5-large	0.6193	0.7248	0.5767	0.3261	0.5205	0.4908	0.5623	0.1087
Seq-to-seq w/ image features								
t5-base	0.7655	0.8670	0.7366	0.2826	0.7012	0.6376	0.7654	0.2391
t5-large	0.7783	0.9106	0.7300	0.3261	0.7028	0.6720	0.7457	0.2826
Classification by encoder only								
t5-base	0.7173	0.8417	0.6723	0.2826	0.6466	0.6124	0.6907	0.2391
t5-large	0.7333	0.8761	0.6789	0.2826	0.6394	0.6628	0.6488	0.2608

respectively. For the large-sized models, we adopt the `t5-large` and `Randeng-T5-784M` [31]. For the image features, we just take the Faster-RCNN [23] region features provided by the competition organizer.

We use the AdamW[14] algorithm to update the model parameters, where the batch size is 64 and the total epoch number is 20. We set learning rate to $1e-4$ and adopt the linear learning rate scheduler with 200 warm-up steps. The models with best validation exact match scores are use for the final evaluation and competition submission.

5.2 Main results

We present our results on `test_dev` and `test` from the competition leaderboard in Table 3. Note that the `Our-t5-large` models in both English and Chinese tracks are the final submission. The `text` set performance of all final submissions have not been officially released by the organizer yet. Currently, our base-sized models (i.e., `Our-t5-base`) are the best on both English and Chinese tracks [1].

Our approach achieves significant improvements compared with the official baseline[2]. On the "YN" (yes or no) type instances, our models are comparable to the baseline implementation. On the "E" (extraction) type answers, our method has score increases of 19.3% and 33.3% for English and Chinese, respectively. On the "G" (generation) type, our models also provides impressive improvements. These demonstrate the effectiveness of our t5-based models.

5.3 Analysis

In this subsection, we conduct several analyses to understand our approach in-depth. It is worth noting that, since we do not aware of the word segmentation method used by the organizer, we can not compute the F1 scores of the "E" (extraction) and "G" (generation) type answers. Therefore, only exact match scores are presented in Table 4 and Table 5.

Table 5: Validation set exact match scores of finetuning t5-base models by various pre-computed image features.

Model	English				Chinese			
	Overall	YN	E	G	Overall	YN	E	G
Faster RCNN region feature								
t5-base	0.7655	0.8670	0.7366	0.2826	0.7012	0.6376	0.7654	0.2391
Faster RCNN grid feature								
t5-base	0.7703	0.8784	0.7379	0.2826	0.6980	0.6399	0.7575	0.2609
Vit-base hidden state								
t5-base	0.7614	0.8647	0.7300	0.3043	0.7004	0.6330	0.7654	0.2609

Vision information is crucial to the model. Since our approach relies on the powerful generation ability of t5 [18] series model, one important question is about the role of vision features to our model. To this end, we remove the pre-computed Faster-RCNN [23] image region features and train the base and large t5 model with only textual inputs. The evaluation results on the validation set are shown in Table 4. The performance t5 models without vision information (Seq-to-seq w/o image features) is dramatically degraded, demonstrating the our approach is rely on the multi-modal information interaction.

Generation is better than classification. The official baseline is a multi-modal classification method, while ours is based on the sequence to sequence generation paradigm. To fairly compare them, we also implemented a t5-encoder base classification model. That is we just drop the decoder part of t5 and use the mean pooling of the encoder output with a linear layer as the model. The results is presented in Table 4 as well. The encoder classification model is with significant performance gap with the generation model on the "E" (extraction) type, which is the most challenging in this VTQA dataset as the open-end answers.

Vision feature source do not matter. We also compare different type of vision features, including the Faster-RCNN region (used in the main results) and grid feature, as well as the Vit [8] hidden states. The validation results is shown in Table 5. We can see that there is no major performance variance among these features. Hence, considering the efficiency, we employ the shortest region feature in the main experiments.

6 CONCLUSION

In this paper, we propose the use of a T5-based multi-modal generative network to address multi-modal question answering. This network structure can be unconstrained by a pre-defined answer space and adaptively extract answers or generate relevant answers according to the question. Extensive experiments conducted on the VQTA dataset validated the effectiveness of the approach. Our solution received first place in both English and Chinese tracks on this competition.

ACKNOWLEDGMENTS

To the people struggling long COVID.

[1] http://vtqa-challenge.fixtankwun.top:20010/index.html

[2] https://github.com/visual-text-QA/VTQA-Demo

REFERENCES

[1] 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*. Computer Vision Foundation / IEEE Computer Society, 6077–6086.

[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*. Computer Vision Foundation / IEEE Computer Society, 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*. IEEE Computer Society, 2425–2433.

[4] Kang Chen and Xiangqian Wu. 2023. VTQA: Visual Text Question Answering via Entity Alignment and Cross-Media Reasoning. *CoRR* abs/2303.02635 (2023). https://doi.org/10.48550/arXiv.2303.02635 arXiv:2303.02635

[5] Xiaolin Chen, Xuemeng Song, Ruiyang Ren, Lei Zhu, Zhiyong Cheng, and Liqiang Nie. 2020. Fine-Grained Privacy Detection with Graph-Regularized Hierarchical Attentive Representation Learning. *ACM Trans. Inf. Syst.* 38, 4 (2020), 37:1–37:26.

[6] Xiaolin Chen, Xuemeng Song, Yinwei Wei, Liqiang Nie, and Tat-Seng Chua. 2023. Dual Semantic Knowledge Composed Multimodal Dialog Systems. In *SIGIR*. ACM, 1518–1527.

[7] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 1724–1734.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[9] Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2019. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. *Int. J. Comput. Vis.* 127, 4 (2019), 398–414.

[10] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the Knowledge in a Neural Network. In *NeurIPS*.

[11] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[12] Gouthaman Kv and Anurag Mittal. 2020. Reducing language biases in visual question answering with visually-grounded question encoder. In *ECCV*. Springer, 18–34.

[13] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *NeurIPS*. 9694–9705.

[14] Ilya Loshchilov and Frank Hutter. [n. d.]. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

[15] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

[16] Panzhong Lu, Xin Zhang, Meishan Zhang, and Min Zhang. 2022. Extending Phrase Grounding with Pronouns in Visual Dialogues. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 7614–7625.

[17] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *CVPR*. Computer Vision Foundation / IEEE, 3195–3204.

[18] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[19] Jun Rao, Liang Ding, Shuhan Qi, Meng Fang, Yang Liu, Li Shen, and Dacheng Tao. 2023. Dynamic Contrastive Distillation for Image-Text Retrieval. *IEEE Transactions on Multimedia* (2023).

[20] Jun Rao, Xv Meng, Liang Ding, Shuhan Qi, and Dacheng Tao. 2022. Parameter-Efficient and Student-Friendly Knowledge Distillation. *CoRR* abs/2205.15308 (2022). https://doi.org/10.48550/arXiv.2205.15308 arXiv:2205.15308

[21] Jun Rao, Tao Qian, Shuhan Qi, Yulin Wu, Qing Liao, and Xuan Wang. 2021. Student Can Also be a Good Teacher: Extracting Knowledge from Vision-and-Language Model for Cross-Modal Retrieval. In *CIKM*.

[22] Jun Rao, Fei Wang, Liang Ding, Shuhan Qi, Yibing Zhan, Weifeng Liu, and Dacheng Tao. 2022. Where Does the Performance Improvement Come From? -A Reproducibility Concern about Image-Text Retrieval. In *SIGIR*. 2727–2737.

[23] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proceedings of the Conference on Neural Information Processing Systems*. 91–99.

[24] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proceedings of the International Conference on Learning Representationss*.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of the Advances in Neural Information Processing Systems*. MIT Press, 5998–6008.

[26] Haokun Wen, Xuemeng Song, Xin Yang, Yibing Zhan, and Liqiang Nie. 2021. Comprehensive Linguistic-Visual Composition Network for Image Retrieval. In *SIGIR*. ACM, 1369–1378.

[27] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Perric Cistac, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. Association for Computational Linguistics, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[28] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2016. Stacked Attention Networks for Image Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 21–29.

[29] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep Modular Co-Attention Networks for Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation / IEEE, 6281–6290.

[30] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From Recognition to Cognition: Visual Commonsense Reasoning. In *CVPR*. Computer Vision Foundation / IEEE, 6720–6731.

[31] Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. 2022. Fengshenbang 1.0: Being the Foundation of Chinese Cognitive Intelligence. *CoRR* abs/2209.02970 (2022).

[32] Zhuosheng Zhang, Hanqing Zhang, Keming Chen, Yuhang Guo, Jingyun Hua, Yulong Wang, and Ming Zhou. 2021. Mengzi: Towards Lightweight yet Ingenious Pre-trained Models for Chinese. arXiv:2110.06696 [cs.CL]