Cross-Modal Dense Passage Retrieval for Outside Knowledge Visual Question Answering

Benjamin Reichman, Larry Heck
Georgia Institute of Technology
bzr@gatech.edu, larryheck@gatech.edu

Abstract

In many language processing tasks including most notably Large Language Modeling (LLM), retrieval augmentation improves the performance of the models by adding information during inference that may not be present in the model's weights. This technique has been shown to be particularly useful in multimodal settings. For some tasks, like Outside Knowledge Visual Question Answering (OK-VQA), retrieval augmentation is required given the open nature of the knowledge. In many prior works for the OK-VQA task, the retriever is either a unimodal language retriever or an untrained cross-modal retriever. In this work, we present a weakly supervised training approach for cross-modal retrievers. Our method takes inspiration from the natural language modeling task of information retrieval and extends those methods to cross-modal retrieval. Since the OK-VQA task does not typically have consistent ground truth retrieval labels, we evaluate our model using lexical overlap between the ground truth and the retrieved passage. Our approach showed an average recall improvement of 28% across a large range of retrieval sizes compared to a baseline backbone network.

1. Introduction

There are many applications that require the use of multiple modalities (e.g. VQA, Visual Dialogue, etc.). Multimodal large language models like GPT-4, Flamingo, and others have recently shown success in performing these tasks [23, 1]. These models are fully parametric and thus can only recall knowledge that is trained into their weights. If they did not remember a particular fact or concept, they cannot access it.

There exists a separate category of models known as semi-parametric or retrieval-augmented models. These models add an extra external retrieval step to their processing. This allows them to recall facts beyond what is found in their weights and training data. In both the unimodal and multimodal cases, retrieval augmentation has been shown to improve model performance [7, 32]. For example, RA-CM3 showed that adding an extra retrieval step allowed it to understand the finer detail and context of an image and the prompts it received [32]. In conversational AI, it was found that adding retrieval decreased model hallucination [26].

In some cases, external knowledge retrieval is required. This is the case in the Outside Knowledge Visual Question Answering (OK-VQA) task where models combine language and vision to answer questions that are only partially related to an image. This expands the set of questions that are possible to ask as the questions no longer need to be directly related to the given image. This is an extension of the VQA task which only asks questions directly related to an image. VQA as a task became popular with the introduction of the 760K Visual QA dataset and baseline by [2] in 2015. However, the questions were generally easy enough that it was judged that an 8-year-old child can answer 55% of them [2]. This gave rise to the need for a more challenging task, hence the introduction of OK-VQA. The most recent dataset to address this task is OK-VQAv2 [25].

To be successful in the OK-VQA task, a model has to be skilled at multiple inference types: (1) visual understanding of the input picture, (2) language Understanding of the input question, (3) commonsense reasoning to understand how the picture and question fit together and into the world, and (4) factual and categorical knowledge to understand how different facts may relate to a given question.

In many prior works for OK-VQA, retrieval of external facts rely on unimodal or untrained multimodal retrieval networks. Furthermore, often the datasets that open-domain multimodal models are trained on do not provide ground truth retrieval sets. When this is the case in the language domain, using a self-supervised matching task like the inverse cloze task (ICT) is a common method to improve retriever performance [16, 11]. In this paper, we (a) extend multimodal inverse cloze pre-training to have additional contrastive tasks built into it and (b) adapt a weakly supervised unimodal retrieval method for multimodal and cross-modal retrieval. The weakly supervised fine-tuning with inverse cloze pre-training improves performance by 28% over the baseline backbone network.

2. Related Work

DPR Retrieval. Biencoder dense passage retrieval (DPR) is the common underlying structure for neural retrieval models. In this general class of methods, one language model computes embeddings for a knowledge base and a second language model computes embeddings for any given query. A semantic similarity scorer, commonly the inner product, is then computed between all of the embeddings in a knowledge base and the embedding of the query. The documents that are most similar to the query are then selected as the retrieval entities by a contrastively trained knowledge and query embedding model. Maximum inner product search is commonly used as it can be implemented in sublinear time, allowing for retrieval within large knowledge bases [15]. This general structure is taken advantage of in many retrieval-based works [3, 7, 8, 9, 10, 12, 13, 18, 28].

OK-VQA Model Components. OK-VQA models usually have at least two of the three components: an implicit retrieval model, an explicit retrieval model, and a reader model [6, 20, 4, 22, 5]. The implicit retrieval system usually starts with converting an image into text. Typically this is done by an image captioning and/or object detection model. This description paired with the question is inputted into a LLM, like GPT-3, which is used as a knowledge base to yield an answer proposal. This process alone can have a top-1 recall of 58.43% and a top-5 recall of 71.31% [20].

Explicit retrieval uses a legible knowledge base such as Wikidata or Wikipedia to provide information to the retrieval model. To find knowledge within these knowledge bases, DPR-based methods are used. The novel method implemented in this paper trains a DPR-based explicit retriever with a weakly supervised, multimodal approach. This method is detailed in Section 3.

The final component is the reader model. This model takes the outputs of the previous two components along with the question and sometimes the image and produces a final answer. Reader models come in two forms: decoder-only and encoder-decoder models. An encoder-decoder model takes language inputs, has the model encode them into latent vectors, and then decodes those latent vectors back into language. Decoder-only models, like GPT, do not have this feature [24]. To augment decoder-only models with retrieval, the text of the retrieved passages is added to the input of the model. An example of this is [31]. One way to process retrieved evidence in the encoder-decoder paradigm is to individually encode each entity with the question, concatenate all encodings, and then jointly decode [12]. This is called fusion-in-decoder [12].

Multimodal Retrieval. Previous works in the OK-VQA literature attempted to address multimodal retrieval. One common solution is to use a non-fine-tuned pre-trained network. KAT and Revive both use the CLIP model for retrieval. KAT extracts image patches using a sliding window over the image [6]. Revive uses object detection to select patches [20]. In each method, patches are inputted into CLIP to retrieve entities from the knowledge base.

Another common approach is to perform DPR on a textual representation of an image. MAVEx, for example, answers the question with a VQA system and then validates the answer proposal using unimodal DPR [30]. TRiG and RA-VQA convert the image into a caption and use OCR [19, 4]. The text is then inputted into a DPR system. Additionally, RA-VQA fine-tunes the DPR system.

A rarer approach in the literature is the use of a trained cross-modal DPR system. One paper did so using LXMERT for their cross-modal system but did not use multimodal ICT pertaining [21]. This approach also used a different type of signal for weakly supervised training.

3. Methods

In this section, we present our training framework for weakly supervised multimodal DPR. The first part of this framework is an extension of the inverse cloze task to be both multimodal and cross-modal. The second part performs weakly supervised fine-tuning of the retriever network by using signals from the reader model. Previously, cross-attention weights were distilled from a unimodal reader to a unimodal retriever. In this work, we distill from a unimodal reader to a multimodal retriever. Fine-tuning the network in a weakly supervised manner is necessary as OK-VQA datasets do not commonly have ground truth labels for retrieval.

3.1. Multimodal Inverse Cloze Pre-training

The cloze task optimizes a model to fill in missing tokens in a string based on the surrounding context [29]. This pre-training task is used to train LLMs to develop a sense of language. CT works in the opposite direction. The model receives a sentence and has to match it to its greater context [16]. This trains a network to understand which pieces of text are closely related and which are more distantly related. This task is closely related to the DPR task, which is why unimodal DPR systems are often pre-trained with ICT.

The OK-VQA task is multimodal, with each modality giving information about what needs to be retrieved The first step to extending DPR systems to be multimodal is to extend the inverse cloze task to be multimodal.

The multimodal inverse cloze task shares the structure of its unimodal counterpart where the model matches a query to a within-batch set [17]. However, in the case of multimodal inverse cloze, both the query and index are multimodal which allows for more internal variation for constructing the query and index.

Our multimodal ICT has four different matching subtasks within it, visually depicted in Figure 1:

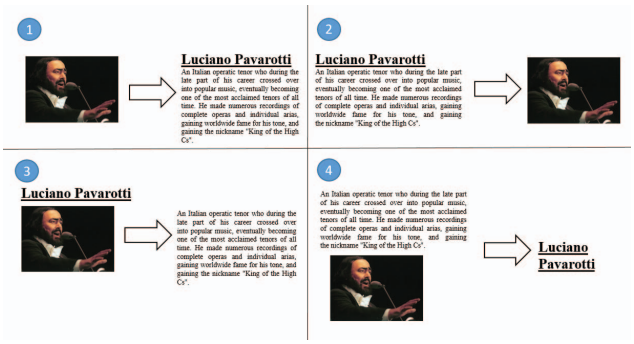

Figure 1. Four variations of multimodal ICT subtasks.

1. Match image to correct NL document
2. Match NL document to corresponding image
3. Match title and image to corresponding passage
4. Match passage and image to corresponding title

The first two tasks are cross-modal in nature. The query is in one modality and the index is in another modality. The last two tasks are both multimodal and cross-modal where both the image and the language are used as part of the query to find a textual match. To prevent the task from being too simple, references to the title in the text passage are removed. In each of the sub-tasks, the cross-entropy loss is used to update the network.

3.2. Cross-Attention Distillation

In-domain training can further improve retrieval. The dataset used has little ground-truth retrieval data. Therefore, the reader model was used as a weak supervision signal.

An encoder-decoder network with fusion-in-decoder was used as the reader model. The fusion-in-decoder method teaches networks to weigh different pieces of evidence during decoding. This is done by having the model individually encode each piece of evidence, but jointly decode the answer from the concatenation of encodings. As these attention weights contain the relative importance of each piece of evidence, they can be used as a learning signal for the retrieval network. The KL-Divergence between the cross-attention weights and the inner-product score given by the retriever can be used to optimize a network in a weakly supervised way to make better retrievals, as depicted in Figure 2. For more in-depth details on this approach, see [11].

As the network performance improves, using the same retrieval passages and cross-attention weights over time becomes less informative. This limitation can be addressed by having the retriever retrieve new passages every set number of steps and then having the reader model rescore the passages. Iteratively training in this fashion gives the retriever more data from which to learn how to retrieve [11].

Previous applications of this method distilled information from a unimodal reader to a unimodal retriever. In this paper, we use the scores from a unimodal reader to supervise a cross-modal retriever, showing the versatility of the cross-attention score.

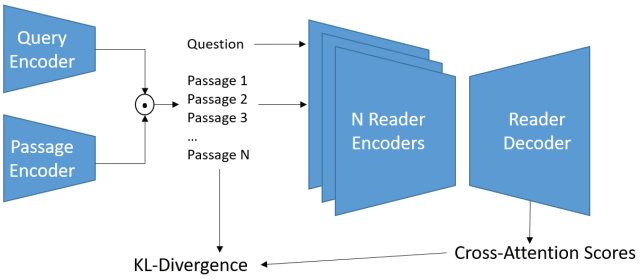

Figure 2. Cross-attention Distillation Framework.

4. Experiments

4.1. Implementation

The method described in the previous section necessitated a model that was trained contrastively to match text and images. This led to the choice of the ALIGN model as the backbone network [14]. Though CLIP was commonly used in previous OK-VQA papers for retrieval, the ALIGN model was chosen as it allows for a longer sequence length. It was also chosen as its shallow cross-encoding allows for text and image to be encoded in the absence of the other.

The Wikipedia-based Image Text (WIT) Dataset was used for the multimodal ICT pre-training. This dataset is a multimodal, multilingual extraction of Wikipedia pages providing millions of language and image pairs [27]. The language and image pairs used were ones where the text was in English and derived from the page title and section text. To get the best performance from contrastive pre-training, large batch sizes must be used. We used gradient checkpointing which allowed for a batch size of 245.

We used the OK-VQAv2 dataset [25] for the downstream retrieval task with image and question pairs. Similar to previous work, the lexical overlap between the retrieved entity and the ground truth answer was used to assess the performance of the retrieval step.

Multiple knowledge bases were used to test if the method generalizes across knowledge bases. In this paper, we used the Wikidata-187K and Wikipedia-21M knowledge bases as retrieval corpora. The Wikidata-187K is a subset of a much larger knowledge base, used and made available by prior works [6, 20]. The Wikidata knowledge base contains 187K entities each with a title and a one-sentence description. The Wikipedia knowledge base contains 21M entities with a title and paragraph description. The downstream QA model receives both the title and paragraph/description.

4.2. Results

To determine if cross-modal modeling of sub-tasks improved the overall task performance, the ALIGN model un-

Tasks	Top N				
	1	5	10	20	40
Baseline (tasks 3-4)	**0.72**	0.87	0.92	0.95	0.97
Ours (all tasks)	0.70	**0.88**	**0.93**	**0.96**	**0.99**

Table 1. The table presents the average performance of the ALIGN model across the various sub-tasks of ICT. The models are trained on all tasks and multimodal tasks (tasks 3 and 4). Accuracy is the fraction that the correct match appears within the top-N.

Inverse Cloze Task	Cross-Attn Distillation	# of Retrieved Passages				
		1	10	50	80	100
✓		13.49	39.22	60.14	66.04	68.84
✓	✓	17.67	47.29	66.77	72.05	74.13

Table 2. This table displays the performance of our multimodal DPR system on the Wikipedia-21M retrieval corpus.

derwent ICT pre-training in two different ways as shown in Table 1. The first (Baseline) only included the last two subtasks while the second (Ours) included all sub-tasks discussed in Section 3.1. The table displays the accuracy of the models on a randomly selected subset of 2,500 data points from the WIT validation set. The accuracy metric is computed by determining the frequency the model predicts the correct match in its top-N predicted matches. This subset was created for faster evaluation. The table shows that incorporating cross-modal tasks improved the accuracy of the multimodal tasks in almost all cases.

Table 3 summarizes the overall results of the methods described in the previous section. The table describes the retrieval performance in terms of its recall of knowledge across different numbers of passages retrieved. Cross-attention fine-tuning for Wikidata improves retriever recall over the Baseline by an average of 10% (relative) and the combination of iterative cross-attention fine-tuning and ICT improves retriever recall by an average of 28%. The results show that the pre-trained model (Baseline) outperforms all but the final model in top-1 retrieval recall. However, the pre-trained model performs worse than any model fine-tuned with cross-attention distillation. The model fine-tuned only with cross-attention distillation outperforms other models not trained with cross-attention distillation. Interestingly, the inverse cloze task does not stand on its own in these results and performs worse than the pre-trained model. However, when paired with cross-attention distillation it outperforms the backbone network that was fine-tuned only with cross-attention distillation. The two tasks are very similar, with the difference being that the inverse cloze task is self-supervised whereas cross-attention distillation is weakly supervised. This seems to suggest that the inverse cloze task changes the model features to be more tunable for retrieval tasks. Cross-attention distillation then takes advantage of this to better fine-tune the model to perform retrieval.

Lastly, Table 2 displays results on a larger and more complete retrieval corpus. While the addition of cross-attention distillation still improves performance over fine-tuning with ICT, the difference with and without cross-attention distillation is smaller. The improved performance is likely a factor of both the model and how the evaluation is carried out. The evaluation measures the lexical overlap between the retrieved passage and the ground truth answers. The Wikipedia retrieval corpus has longer passages, which increases the likelihood that the correct word appears in a retrieved passage. It also has more passages, increasing the chances that a passage with the correct word is surfaced. On the modeling side, the ALIGN model is designed to have a sequence length of 512 tokens which is much longer than the average Wikidata entity description. The longer passages in Wikipedia are more descriptive and thus better able to take advantage of ALIGN's latent space which likely helps the model find the correct passage.

Inverse Cloze Task	Cross-Attn Distillation	# of Retrieved Passages				
		1	10	50	80	100
Baseline		5.99	20.83	38.07	43.02	45.35
	✓	5.47	23.61	42.12	47.57	50.33
✓		4.67	18.51	36.56	42.62	45.35
✓	✓	6.65	26.37	46.98	52.59	55.19
✓	Iterative	7.67	28.54	48.87	53.94	56.72

Table 3. Performance of our multimodal DPR system compared to the previous approach (Baseline) using the ALIGN backbone at different stages of training for different numbers of retrieved passages on the Wikidata-187K retrieval corpus.

5. Conclusion

In this paper, we focused on improving the performance of cross-modal retrieval for the OK-VQA task. We achieved this by extending the ICT to be both cross-modal and multimodal. We show that adding sub-tasks that are purely cross-modal helps with the multimodal tasks that retrieve across modalities. We then show that models trained with cross-attention outperform models that are not trained with cross-attention. It was found that adding ICT improves the performance of networks on retrieval only when coupled with cross-attention, however, on its own, it did not improve performance for the Wikidata retrieval corpus.

Future work should focus primarily on two factors limiting cross-attention distillation: (1) the performance of the reader model and (2) the relative usefulness of the initial retrievals. Retrieval models can only give an accurate signal of what passages are useful if it itself is accurate. Cross-attention scores from inaccurate answers are not useful. The relative usefulness of the initial retrieval is a limitation because if the retrieval did not provide any "good" passages or all of the passages provided are on the same level of relative "badness", the cross-attention score provides no signal. Training in an iterative manner partially addresses this problem, but other methods should be explored in future work.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[3] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR, 2022.

[4] Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5067–5077, 2022.

[5] François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. ConceptBert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 489–498, Online, Nov. 2020. Association for Computational Linguistics.

[6] Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language. In *NAACL*, 2022.

[7] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020.

[8] Dilek Hakkani-Tür, Malcolm Slaney, Asli Celikyilmaz, and Larry Heck. Eye gaze for spoken language understanding in multi-modal conversational interactions. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 263–266, 2014.

[9] Larry Heck, Dilek Hakkani-Tür, Madhu Chinthakunta, Gokhan Tur, Rukmini Iyer, Partha Parthasacarthy, Lisa Stifelman, Elizabeth Shriberg, and Ashley Fidler. Multi-modal conversational search and browse. In *Proceedings of the First Workshop on Speech, Langauge and Audio in Multimedia (SLAM), Marseille, France*, 2013.

[10] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338, 2013.

[11] Gautier Izacard and Edouard Grave. Distilling knowledge from reader to retriever for question answering. *arXiv preprint arXiv:2012.04584*, 2020.

[12] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering, 2020.

[13] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 2022.

[14] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[15] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[16] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*, 2019.

[17] Paul Lerner, Olivier Ferret, and Camille Guinaudeau. Multimodal inverse cloze task for knowledge-based visual question answering. In *European Conference on Information Retrieval*, pages 569–587. Springer, 2023.

[18] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[19] Weizhe Lin and Bill Byrne. Retrieval augmented visual question answering with outside knowledge. *arXiv preprint arXiv:2210.03809*, 2022.

[20] Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. Revive: Regional visual representation matters in knowledge-based visual question answering. *arXiv preprint arXiv:2206.01201*, 2022.

[21] Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. Weakly-supervised visual-retriever-reader for knowledge-based question answering. *arXiv preprint arXiv:2109.04014*, 2021.

[22] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14121, 2021.

[23] OpenAI. Gpt-4 technical report, 2023.

[24] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[25] Benjamin Z. Reichman, Anirudh Sundar, Christopher Richardson, Tamara Zubatiy, Prithwijit Chowdhury, Aaryan Shah, Jack Truxal, Micah Grimes, Dristi Shah, Woo Ju Chee, Saif Punjwani, Atishay Jain, and Larry Heck. Outside knowledge visual question answering version 2.0. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[26] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation, 2021.

[27] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2443–2449, New York, NY, USA, 2021. Association for Computing Machinery.

[28] Anirudh Sundar and Larry Heck. Multimodal conversational ai: A survey of datasets and approaches. *Proceedings of the 4th ACL Workshop on NLP for Conversational AI*, 2022.

[29] Wilson L Taylor. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433, 1953.

[30] Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. Multi-modal answer validation for knowledge-based vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2712–2721, 2022.

[31] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089, 2022.

[32] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. Retrieval-augmented multimodal language modeling, 2023.