Pre-Training Multi-Modal Dense Retrievers for Outside-Knowledge Visual Question Answering

Alireza Salemi
University of Massachusetts Amherst
United States
asalemi@cs.umass.edu

Mahta Rafiee
University of Massachusetts Amherst
United States
mrafiee@cs.umass.edu

Hamed Zamani
University of Massachusetts Amherst
United States
zamani@cs.umass.edu

ABSTRACT

This paper studies a category of visual question answering tasks, in which accessing external knowledge is necessary for answering the questions. This category is called outside-knowledge visual question answering (OK-VQA). A major step in developing OK-VQA systems is to retrieve relevant documents for the given multi-modal query. Current state-of-the-art asymmetric dense retrieval model for this task uses an architecture with a multi-modal query encoder and a uni-modal document encoder. Such an architecture requires a large amount of training data for effective performance. We propose an automatic data generation pipeline for pre-training passage retrieval models for OK-VQA tasks. The proposed approach leads to 26.9% Precision@5 improvements compared to the current state-of-the-art asymmetric architecture. Additionally, the proposed pre-training approach exhibits a good ability in zero-shot retrieval scenarios.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Question answering**; **Multimedia and multimodal retrieval**; • **Computing methodologies** → **Computer vision**.

KEYWORDS

Dense Retrieval; Visual Question Answering; Multi-Modal Retrieval; Pre-training; Data Generation

ACM Reference Format:
Alireza Salemi, Mahta Rafiee, and Hamed Zamani. 2023. Pre-Training Multi-Modal Dense Retrievers for Outside-Knowledge Visual Question Answering. In *Proceedings of the 2023 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '23), July 23, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3578337.3605137

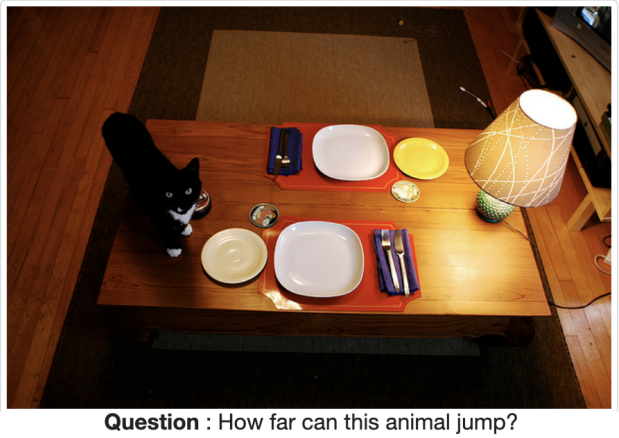

Figure 1: An example OK-VQA question. Answering this question requires external knowledge.
Image © nsfmc, https://www.flickr.com/photos/subliminal/589841807

1 INTRODUCTION

Outside-knowledge visual question answering (OK-VQA) [27] is a category of visual question answering tasks in which answering the given natural language question about an image requires access to external information. In OK-VQA retrieval tasks, queries are multi-modal (text and image) and the retrieval corpus is often uni-modal, consisting of text documents. To shed more light on this, Figure 1 presents an example of the queries in this task. It can be observed that answering the question 'how far can this animal jump?' requires an understanding of the entity (i.e., the cat) in the image, but this information may not be sufficient. In this case, accessing a knowledge source containing information about the animal in the picture and its abilities can facilitate answering the question. The extensive range of practical implementations for OK-VQA is worth noting. Consider the scenarios where individuals, who are patrons of e-commerce platforms, capture images of products or specific components and pose queries about them Salemi et al. [36]. Similarly, within the educational domain, students can interrogate an image from their textbook by asking questions Salemi et al. [36]. Moreover, users can leverage OK-VQA by photographing visual signs or artwork and inquiring about their significance or historical background. These instances merely scratch the surface of the diverse application potential of OK-VQA.

Lately, Salemi et al. [36] introduced a symmetric architecture for multi-modal retrieval and compared it with the previous state-of-the-art asymmetric architectures introduced by Qu et al. [29]. Despite the significantly better performance of the symmetric architecture, the fact that this architecture needs access to a caption generator at the inference time makes it costly to use in real-time. Consequently, the main emphasis of this research paper is placed

ICTIR '23, July 23, 2023, Taipei, Taiwan

ACM ISBN 979-8-4007-0073-6/23/07…$15.00
https://doi.org/10.1145/3578337.3605137

on asymmetric architecture, unveiling a novel methodology for enhancing the training of superior asymmetric retrievers. Importantly, this approach effectively curtails the necessity of caption generation solely to the training phase, sparing it from being required during inference time.

Qu et al. [29] demonstrates that supervised asymmetric dense retrieval models with multi-modal query encoder and uni-modal document encoder lead to state-of-the-art passage retrieval performance for OK-VQA tasks. However, it requires large-scale manually labeled training data which is expensive and time consuming to obtain. Inspired by the prior research on text retrieval based on weak supervision [7, 48, 49] and Inverse Cloze Task (ICT) pre-training [4], this paper introduces a novel pipeline for automatic generation of training data for OK-VQA tasks. This data generation pipeline requires no manually labeled OK-VQA data. It first obtains an image corpus (e.g., MS COCO [23]) and generates captions for the images. Each caption is then used as a query to retrieve text passages from Wikipedia. We then select some noun phrases from each passage as potential answers and generate a question for each of them using a fine-tuned language model. To reduce the noise introduced into the pre-training data, we design a question-answering model and filter out questions for which the model cannot produce a close enough answer. This process leads to a large-scale dataset with about 4.6 million question-image pairs for OK-VQA tasks. The generated data can then be used for pre-training dense retrieval models for OK-VQA tasks. To the best of out knowledge, this is the first attempt to automatic generation of data for OK-VQA tasks.

Our experiments on the OK-VQA passage retrieval dataset [27, 29] demonstrate that training dense retrieval models using the proposed data generation pipeline leads to 40.2% Precision@5 improvements in a zero-shot setting compared to competitive baselines. We also show that pre-training state-of-the-art supervised dense retrieval models improves state-of-the-art performance by 26.9% in terms of Precision@5. The obtained improvements are statistically significant in all cases. Further analysis suggests that the proposed pre-trained model that is fine-tuned only on 25% of the OK-VQA supervised data outperforms the model that is trained on 100% of the supervised data without pre-training. Moreover, the performance of the pre-trained model becomes relatively stable after observing 50% of the supervised training data. Therefore, the proposed pre-training procedure reduces the need to large-scale manually labeled training sets.

In summary, the major contributions of this work include:

(1) Introducing the first automatic data generation pipeline for outside-knowledge visual question answering tasks.
(2) Improving the current state-of-the-art asymmetric passage retrieval models in both zero-shot and supervised settings.
(3) Providing extensive result analysis to better understand the impact of pre-training on OK-VQA performance.

To foster research in this area, we release our generated dataset, our data creation pipeline, and our learned model parameters.[1]

[1]The data and code are available at https://github.com/alirezasalemi7/pretraining-multimodal-dense-retriever-for-okvqa

2 RELATED WORK

Multi-Modal Dense Passage Retrieval. Multi-modal dense retrieval can be defined in different categories based on where the multi-modality takes place. The multi-modality can be in the queries, with a corpus of uni-modal documents, which enables the underlying information need to be expressed through a multi-modal representation [27]. Our work fits into this category with queries comprised of images with corresponding questions, and uni-modal textual passages in the corpus. Another line of work has been focusing on multi-modal documents in the corpus, such as a mix of textual, tabular, or visual information, while the query is expressed in one modality [13, 25, 39]. In another setting, both queries and documents can be multi-modal, for example where the answer to a query about an image contains multiple modalities [38]. Cross-modal retrieval is also partly related to multi-modal retrieval, where both queries and documents are uni-modal but they come from different modalities [15, 30].

Outside-Knowledge Visual Question Answering. In standard visual question answering (VQA) [2], the answer lies in the image; however, in outside-knowledge visual question answering (OK-VQA) [27], the image and question are jointly used to find the answer to the question from an external knowledge source [29]. That being said, retrieving relevant passages to a query, which consists of an image and a question about it, plays an essential role in this task [29]. Previous work [10, 12, 26, 37, 46] mostly utilizes knowledge graphs as a source of external information; however, the lack of a complete and easily updatable knowledge source is challenging [1, 41]. Therefore, following Qu et al. [29] and Salemi et al. [36], we focus on retrieving passages from Wikipedia as the knowledge source.

Previous work mostly evaluates OK-VQA based on the answer generation quality [6, 10–12, 26, 37, 46, 47]; however, following Qu et al. [29], we only investigate the retrieval performance in the aforementioned task. In contrast with Salemi et al. [36], which focuses on designing a symmetric architecture for OK-VQA retrieval and answer generation, we investigate the data generation and augmentation methods to train the proposed asymmetric retriever architecture by Qu et al. [29] with no labeled training data.

Pre-Training Dense Passage Retrievers. In recent years, pre-training transformers [43] using semi- and self-supervised tasks has become a standard approach for achieving strong performance in natural language and vision tasks [8, 9, 24]. Moreover, retrieval-specific pre-training tasks, such as Inverse Cloze Task (ICT) [4], have been shown to be effective for uni-modal retrieval. Recently, a multi-modal variant of ICT has been proposed by Lerner et al. [21], in which queries are question-image pairs, and documents are passage-image pairs. However, our work focuses on the case that passages are only textual, while queries consist of question-image pairs.

The research by Changpinyo et al. [5] is perhaps the closest work to ours, in which the authors focus on pre-training models for VQA tasks, which is by nature different from OK-VQA. Changpinyo et al. [5] only generates questions from the image captions due to the nature of VQA, in which the answers lie in the image. In contrast, we use captions to retrieve a relevant passage to the image and

generate questions from that passage to ensure that answering them requires external knowledge.

3 PROBLEM STATEMENT

While multi-modal retrieval can be defined in different ways as mentioned in section 2, this paper only focuses on multi-modal scenarios where the query (Q, I) consists of the question Q about the image I, and the corpus C from which relevant passages should be selected is only textual.

Suppose $T = \{(Q_1, I_1, A_1, R_1), ..., (Q_N, I_N, A_N, R_N)\}$ represents the training set for multi-modal retrieval in this paper. Each training sample in T consists of a question Q_i written in natural language, an image I_i, a set of answers A_i to the question Q_i, and a set of relevant passages R_i that contains the answer to Q_i. In more detail, the answer set A_i might contain more than one answer to the question, which are syntactically different but semantically the same ($|A_i| \geq 1$). Additionally, each question and image might have more than one related passage ($|R_i| \geq 1$ and $R_i \subseteq C$).

The main task in this paper is to use training set T to train a dense retriever that takes query (Q, I) as input and retrieves K passages that are relevant to the query from the corpus C ($|C| \gg K$). In this paper, we introduce a pipeline for generating weakly supervised data, similar to the proposed problem definition, to first pre-train the model on the weakly supervised generated data and then fine-tune the pre-trained on the task's data. The following sections explain our proposed pipeline for this purpose.

4 THE PROPOSED PRE-TRAINING PIPELINE

Automatic data generation for (pre-)training neural models for text retrieval and question answering has proven to be effective. For instance, Dehghani et al. [7] introduced weak supervision in information retrieval by utilizing an existing unsupervised retrieval model as a weak labeler. Zamani and Croft [49] provided theoretical justification on when and why weak supervision lead to strong and robust improvements. Wang et al. [44] used a similar approach for adapting well-trained retrieval models to an unseen target domain. More recently, Chang et al. [4] used Inverse Cloze Task for pre-training text retrieval models and Bonifacio et al. [3] used large-scale language models, such as GPT [31], for data generation. All these approaches are developed for text retrieval tasks.

For multi-modal tasks, Changpinyo et al. [5] focused on pre-training models for visual question answering (VQA) tasks, which is fundamentally different from OK-VQA. Changpinyo et al. [5] only generates questions from the image's caption due to the nature of VQA, in which the answers lie in the image (e.g., asking about the color of an object in the image). VQA is not an information-seeking task; thus, this approach cannot be applied to OK-VQA.

This section introduces our data generation pipeline for pre-training dense passage retrieval models for OK-VQA tasks.

4.1 Automatic Data Generation for Pre-training

Figure 2 depicts an overview of our automatic data generation pipeline for pre-training multi-modal dense passage retrieval models. We start with an image and use an automatic image captioning model to produce a textual description of the image. We then retrieve M passages from a large collection, such as Wikipedia, given the image caption as the query. We then extract a set of potential short answers from the retrieved passages. For each potential answer, we generate a question using a sequence-to-sequence model. We later filter out low quality questions. A negative selection component is also developed to produce data for optimizing retrieval models. The outcome of our pipeline is a set of data instances, each represented as $(Q_i, I_i, A_i, R_i, N_i)$, where Q_i is a question about the image I_i, A_i is the answer to the question Q_i, R_i is a relevant passage to the question, and N_i is a hard negative passage for the question Q_i. In the following sections, we explain the procedure of generating each component in detail.

Matching Images and Passages using Captions. For each image in the MS COCO [23] training set ($\tilde{8}$2K images), we aim at retrieving M passages. Therefore, designing retriever $R_{img2text}$, which takes an image as input and retrieves a set of related passages from corpus C), is required. We use a Wikipedia dump with 11M passages as the corpus C.[2] We retrieve $M = 5$ passages for each image.

While some models, such as CLIP [30] and ALIGN [15], are designed to act as R_{img2text}, for simplicity and without losing generality, we use BM25 [35] as R_{img2text}, in which we use a textual description of the image to retrieve a set of passages. To calculate the similarity score between the image I and the passage P, we use the following formula: $S_R(I, P) = S_{BM25}(\phi_{I \to T}(I), P)$, where $\phi_{I \to T}$ is a modality converting module that takes an image and generates a textual description for it.

Generating a description of an image can happen in several ways. For instance, the textual label of objects in an image can be used to describe the image using text. This approach suffers from two issues: 1) object labels are limited to a pre-defined set, and 2) labeling objects in images in large scale is costly. Conversely, using captions as the image description resolves the mentioned issues by generating an open-ended textual description of an image and using the large-scale available image-caption data on the web. That being said, we use ViT-GPT [19], a transformer-based [43] image-to-text model, to generate a caption for each image. ViT-GPT is trained on the images and captions provided by MS COCO [23] dataset using a cross-entropy loss function. Once the model is trained, we freeze the model's parameters and use it in inference mode.

Selecting Potential Answer Phrases from Retrieved Passages. Investigating the OK-VQA dataset shows that approximately 80% of the answers in this dataset are noun phrases. Following this observation, we use noun phrases in the retrieved passages as potential answers. This approach has been previously used by Lee et al. [20]. In more detail, we use spaCy[3] to extract noun phrases from passages. We consider all noun phrases as potential answers, except those that have a pronoun or determiner (e.g., "a", "an", and "the") in their subtrees. This is because pronouns and determiners usually refer to a specific word in the passage (i.e., co-references), and we would like to select "standalone" answer phrases.

Question Generation and Filtering. The next step in the pre-training data generation pipeline is generating a question for each selected answer phrase. Suppose $M_{QG}(A, P)$ is a question generator

[2]This Wikipedia dump is available at: https://ciir.cs.umass.edu/downloads/ORConvQA/all_blocks.txt.gz
[3]https://spacy.io/

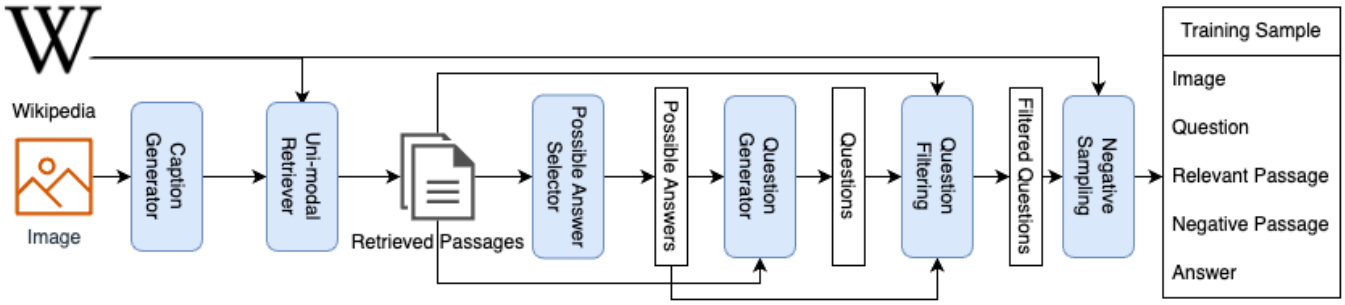

Figure 2: The proposed data generation pipeline for pre-training OK-VQA models.

that takes passage P and the answer phrase A as input and generates a question Q whose answer is A. To implement M_{QG}, following Ushio et al. [42], we feed the passage P and the potential answer phrase A to T5-large [32] and instruct the model to generate a question. To this aim, we utilize SQuAD v1.1 [34] dataset for fine-tuning this question generation model. For each training sample in the SQuAD v1.1 dataset, the answer is surrounded with <hl> token, and the passage with the surrounded answer is fed to T5. The cross-entropy loss is used for training the model:

$$L_{QG} = -\sum_{i}^{|Q|} \log P(y_i | y_{k<i}; P') \tag{1}$$

where y_i is the i^{th} token in the question Q, and P' is the passage P with <hl> surrounding tokens[4].

As a reference on the quality of the question generation model, we evaluate it on the test set of SQuAD [34] and it achieves a BLEU-4 [28] score of 27.21 and rouge-L [22] of 54.13. To further reduce the amount of noise in the generated pre-training data, we filter out the questions that a question-answering model cannot answer. Suppose $M_{QA}(Q, P)$ is a question-answering model, which takes the question Q and the passage P as inputs and generates or selects an answer phrase. Finally, we only select the generated questions that satisfy the following condition: $\text{ROUGE-1}(A, M_{QA}(M_{QG}(A, P), P)) > T$, where ROUGE-1 is the rouge-1 score [22], A is the potential answer from the passage P, and T is a threshold for the similarity of the potential answer and the answer selected by the question-answering model. We use $T = 0.5$ in our experiments.

To implement M_{QA}, we use a RoBERTa-base [24] that is fine-tuned for answer span selection trained on the SQuAD dataset. The model is trained based on the log-likelihood of predicting the correct start and end tokens. For selecting the answer span, the span with the highest $P(S_i|P;Q) + P(E_j|P;Q)$ is selected where $P(S_i|P;Q)$ shows the probability of the i^{th} token being the start of the span and $P(E_j|P;Q)$ shows the probability of the j^{th} token being the end of the span. As a reference, this question-answering model achieves a F1 score of 82.91% and exact match of 79.87% on the test set of SQuAD v2 dataset [33].[5]

[4]The checkpoint for this question-generation model is available at: https://huggingface.co/lmqg/t5-large-squad-qg

[5]The checkpoint for this question-answering model is available at: https://huggingface.co/deepset/roberta-base-squad2

Negative Passage Sampling. Using hard negatives and their quality plays an essential role in the final performance of dense passage retrieval models [17]. For each generated question, we retrieve passages using BM25. We choose the highest scored passage that does not contain the answer A as the negative passage.

Summary. The proposed pipeline leads to 4,621,973 question-image pairs from 82,783 unique images of MS COCO [23]. The average question, passage, and answer length in the created dataset are 9.6 ± 3.0, 187.2 ± 105.7 and 2.3 ± 1.2 words, respectively.

4.2 Dense Retrieval Model

The nature of the multi-modal retrieval task that we attempt to solve in this paper requires the bi-encoder dense passage retriever to encode queries in multi-modal semantic space and to encode passages in textual semantic space. We use an asymmetric state-of-the-art dense passage retrieval for OK-VQA tasks proposed by Qu et al. [29]. It uses an asymmetric dense passage retriever with the multi-modal query encoder E_{MM} and the textual passage encoder E_T. Then, the relevance score is calculated as follows: $S((Q, I), P) = E_{MM}(Q, I) \cdot E_T(P)$, where $\cdot$ denotes the inner product. Following Qu et al. [29], we implement E_T using the representation of the [CLS] token provided by a BERT-base [8] model. Similarly, we utilize the representation of the [CLS] token generated by LXMERT [40], a vision-language model pre-trained with various vision-language tasks.

To train the retriever, we use a contrastive loss as follows:

$$L_{DR} = -\log \frac{e^{S((Q,I),P_{pos})}}{e^{S((Q,I),P_{pos})} + \sum_{P' \in \mathbf{P_{neg}}} e^{S((Q,I),P')}} \tag{2}$$

where P_{pos} is a positive (relevant) passage and $\mathbf{P_{neg}}$ is a set of negative passages for the question-image pair (Q, I). In addition to the selected negative passages, we use in-batch negatives, in which all the positive and negative passages of other queries in the same training batch are considered as negative passages to the query. We use the Faiss library [16] for indexing and efficient dense retrieval.

5 EXPERIMENTS

This section discusses the datasets, experiments, and results obtained in this paper.

Table 1: Passage retrieval performance on the OK-VQA dataset [27]. The superscript * denotes statistically significant improvement compared to all baselines based on two-tailed paired t-test with Bonferroni correction ($p < 0.05$).

Model	Zero-Shot Performance				Supervised Performance			
	Validation		Test		Validation		Test	
	MRR@5	P@5	MRR@5	P@5	MRR@5	P@5	MRR@5	P@5
BM25	0.2450	0.1668	0.2528	0.1642	0.2565	0.1772	0.2637	0.1755
Dense-BERT	0.0709	0.0382	0.0726	0.0375	0.4555	0.3155	0.4325	0.3058
BERT-LXMERT	0.0744	0.0376	0.0665	0.0345	0.4704	0.3364	0.4526	0.3329
Pre-trained BERT-LXMERT	**0.3716***	**0.2629***	**0.3364***	**0.2303***	**0.5557***	**0.4195***	**0.5603***	**0.4274***
% rel. imp. w.r.t. the best baseline	51.6% ↑	57.6% ↑	33.0% ↑	40.2% ↑	17.5% ↑	20.4% ↑	21.2% ↑	26.9% ↑

5.1 Experimental Setup

Dataset. In our experiments, we use the OK-VQA passage retrieval dataset [29], an extension to the OK-VQA dataset [27]. This dataset aims at evaluating passage retrieval tasks for outside-knowledge visual question answering tasks. This dataset contains 9009 questions for training, 2523 questions for validation, and 2523 for testing. As the retrieval collection, it uses the same Wikipedia dump that we use during pre-training (11M passages).

Pre-training and Fine-tuning setups. In order to pre-train the multi-modal dense passage retriever, we use a batch size of 32 on four RTX8000 GPUs, each with 49GB of GPU memory and a total of 256GB of RAM, which results in an effective batch size of 128. We utilize the Adam optimizer [18] with a learning rate of 10^{-5}. A linear learning rate scheduler with 10% of total training steps as warmup steps is used for pre-training. Additionally, gradient clipping with a clipping value of 1.0 is used in the training procedure. The maximum length of passages and queries for each encoder is 384 and 20 tokens, respectively. We only train the model for one epoch on the pre-training data to avoid overfitting.

For fine-tuning on the OK-VQA training set, we follow the same training setup, but we use two epochs and a batch size of 4 on each GPU for a fair comparison with previous work [29], which results in an effective batch size of 16.

Baselines and Terms of Comparison. We compare our models with the following baselines. (1) **BM25**: a baseline that only uses the question as the query and retrieves passages using BM25. (2) **Dense-BERT**: a dense retrieval baselines similar to DPR [17] that uses questions as queries and is trained using the same training objective as ours. (3) **BERT-LXMERT**: a state-of-the-art asymmetric dense retrieval model [29] that uses the exact same architecture as we introduced in Section 4.2. This baseline is basically our model but without being pre-trained using the generated data.

Evaluation. Following Qu et al. [29], we use mean reciprocal rank (MRR) and precision with ranking cut-off of 5 as evaluation metrics. We use the two-tailed paired t-test with Bonferroni correction as the statistical significance test ($p < 0.05$). Since the OK-VQA dataset does not provide relevant judgment for passages, we assume a passage is identified to be positive if it contains an exact match (case insensitive) of a ground truth answer [29].

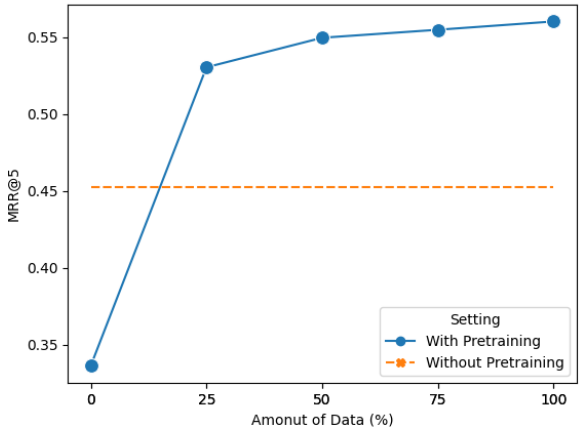

Figure 3: Learning curve for 'Pre-trained BERT-LXMERT' on the OK-VQA test set. The orange line shows the performance of the BERT-LXMERT model without pre-training that is fine-tuned on 100% of the supervised OK-VQA training data.

5.2 Results

This section presents our experimental results and analyzes the model performance to better demonstrate the impact of pre-training on OK-VQA performance.

Zero-Shot Performance. In the first set of experiments, we evaluate the zero-shot capabilities of the models. In this setting, the BM25 baseline uses the default parameters ($k_1 = 1.2, b = 0.75$), and the baselines with BERT and LXMERT use the parameters learned through their (vision-) language model pre-training. The 'Pre-trained BERT-LXMERT' model is trained on the data that we automatically generated. The results are reported in Table 1. BM25 demonstrates the strongest zero-shot performance. This suggests that the initialized parameters of BERT and LXMERT are not suitable for retrieval tasks. This is inline with findings by previous work on text retrieval [14, 50]. The proposed pre-training pipeline significantly outperforms all the baselines and leads to 33% and 40.2% MRR@5 and P@5 improvements compared to BM25, respectively.

Supervised Performance. In the second set of experiments, we fine-tune the same models on the OK-VQA training set. All neural models use the same training procedure. The BM25 parameters are tuned through exhaustive grid search where $k_1 \in [0.5, 1.5]$ and $b \in [0.2, 0.8]$ with a step size of 0.2. The model with the best MRR@5

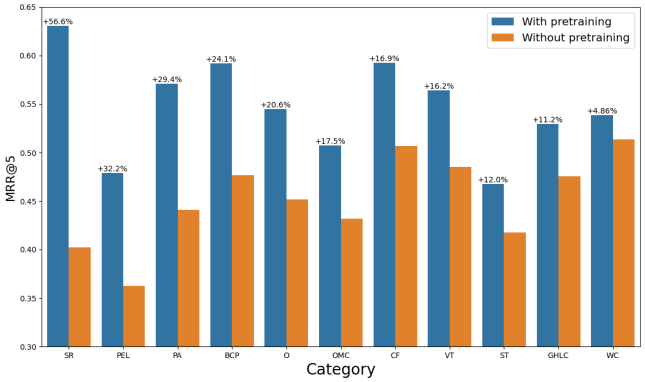

Figure 4: MRR@5 for the fine-tuned BERT-LXMERT model with and without pre-training for different question categories. The percentages on top of the bars indicate relative improvements compared to the model without pre-training.

on the validation set is selected. The selected parameters are $k_1 = 1.1, b = 0.4$. In Table 1, we observe that, as expected, all neural models largely benefit from fine-tuning on the OK-VQA training set and substantially outperform BM25. Fine-tuning BERT-LXMERT that is pre-trained using the proposed data generation pipeline leads to 21.2% MRR@5 and 26.9% P@5 improvements compared to BERT-LXMERT without pre-training (i.e., the current SOTA model on passage retrieval for OK-VQA [29]).

Learning Curve. We hypothesize that the proposed pre-training pipeline reduces the need for large-scale supervised training data, which is often difficult or expensive to obtain. To validate this hypothesis, we fine-tuned our pre-trained model using 25%, 50%, 75%, and 100% of the supervised data randomly sampled from the OK-VQA training set. The results are plotted in Figure 3. For the sake of space, the performance based on MRR@5 on the OK-VQA test set is reported. Other curves follow a similar behavior. The dashed orange line in the figure shows the performance of the BERT-LXMERT model without pre-training that is trained on 100% of the OK-VQA training set. The curve demonstrates that our pre-trained model outperforms the model without pre-training by only observing 25% of supervised training data. Moreover, the performance of the pre-trained model becomes relatively stable after observing 50% of the supervised data, which shows that pre-training retrieval models for OK-VQA reduces the need for supervised data.

Result Analysis. To have a deeper understanding of the proposed pre-training impact on OK-VQA tasks, Figure 4 presents MRR@5 obtained by the fine-tuned BERT-LXMERT model with and without pre-training for each question category. The categories are borrowed from the OK-VQA [27] dataset.[6] We observe that pre-training improves the OK-VQA performance on all question categories, however, the improvements are not the similar across categories. It can be seen that the highest improvement is achieved for the "sports & recreation" category (56.6%), while the lowest improvement is observed for the "weather & climate" category (4.86%). The reason is that we use the MS COCO dataset [23] as the image collection for automatic creation of our pret-training data and MS COCO does not include any category related to "weather & climate," "science & Tech," and "Geography & History & Language & Culture". As a result, the extent of improvement is smaller for these categories in the OK-VQA dataset. On the other hand, a considerable proportion of images in the MS COCO dataset are related to the categories such as "Sport & Recreation," "People & Everyday Life," and "Plants & Animals" that observe the highest improvements. This analysis demonstrates that including the nature of data included in the automatic data creation pipeline directly impact the downstream OK-VQA performance, and including images from underrepresented categories is likely to further improve the performance.

[6]The categories include "Plants & Animals (PA)," "Science & Tech (ST)," "Sport & Recreation (SR)," "Geography & History & Language & Culture, (GHLC)" "Brands & Companies & Products (BCP)," "Vehicles & Transportation (VT)," "Cooking & Food (CF)," "Weather & Climate (WC)," "People & Everyday Life (PEL)," "Objects & Material & Clothing (OMC)," and "Other (O)."

6 CONCLUSIONS AND FUTURE WORK

This paper introduced a pipeline for pre-training dense retrievers for OK-VQA tasks. The proposed pipeline started from an image collection and paired each image with a passage from a knowledge source. Then, a question generation model was used to generate questions for all possible answers to the questions about the image and the passage. Finally, low-quality questions were filtered out, and negative samples for the remaining questions were selected. Our experiments suggest statistically significant improvements compared to state-of-the-art asymmetric dense retrieval performance for OK-VQA tasks.

Even though our results show consistent improvement in the OK-VQA dataset, there might be some other kinds of knowledge-intensive VQA datasets, such as FVQA [45], that this pre-training approach needs to be revised. In the future, we intend to extend our data generation pipeline to other knowledge-intensive vision-language tasks. This paper also limits multi-modality to multi-modal queries and textual passages. Providing a solution for removing the mentioned limitations can be investigated in future work.

ACKNOWLEDGMENT

This work was supported in part by the Center for Intelligent Information Retrieval, in part by Lowes, and in part by NSF grant #2106282. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

REFERENCES

[1] Bilal Abu-Salih. 2021. Domain-specific knowledge graphs: A survey. *Journal of Network and Computer Applications* 185 (2021), 103076. https://doi.org/10.1016/j.jnca.2021.103076

[2] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2017. VQA: Visual Question Answering. *Int. J. Comput. Vision* 123, 1 (may 2017), 4–31. https://doi.org/10.1007/s11263-016-0966-6

[3] Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. InPars: Unsupervised Dataset Generation for Information Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2387–2392. https://doi.org/10.1145/3477495.3531863

[4] Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training Tasks for Embedding-based Large-scale Retrieval. *ArXiv*

abs/2002.03932 (2020).

[5] Soravit Changpinyo, Doron Kukliansy, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. 2022. All You May Need for VQA are Image Captions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 1947–1963. https://doi.org/10.18653/v1/2022.naacl-main.142

[6] Zhuo Chen, Yufeng Huang, Jiaoyan Chen, Yuxia Geng, Yin Fang, Jeff Z. Pan, Ningyu Zhang, and Wen Zhang. 2022. LaKo: Knowledge-driven Visual Question Answering via Late Knowledge-to-Text Injection. *CoRR* abs/2207.12888 (2022).

[7] Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W. Bruce Croft. 2017. Neural Ranking Models with Weak Supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 65–74. https://doi.org/10.1145/3077136.3080832

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*. https://openreview.net/forum?id=YicbFdNTTy

[10] François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. 2020. ConceptBert: Concept-Aware Representation for Visual Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 489–498. https://doi.org/10.18653/v1/2020.findings-emnlp.44

[11] Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. KAT: A Knowledge Augmented Transformer for Vision-and-Language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 956–968. https://doi.org/10.18653/v1/2022.naacl-main.70

[12] Yangyang Guo, Liqiang Nie, Yongkang Wong, Yibing Liu, Zhiyong Cheng, and Mohan Kankanhalli. 2022. A Unified End-to-End Retriever-Reader Framework for Knowledge-Based VQA. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 2061–2069. https://doi.org/10.1145/3503161.3547870

[13] Darryl Hannan, Akshay Jain, and Mohit Bansal. 2020. ManyModalQA: Modality Disambiguation and QA over Diverse Inputs. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (Apr. 2020), 7879–7886. https://doi.org/10.1609/aaai.v34i05.6294

[14] Gautier Izacard and Edouard Grave. 2021. Distilling Knowledge from Reader to Retriever for Question Answering. In *International Conference on Learning Representations*. https://openreview.net/forum?id=NTEz-6wysdb

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *International Conference on Machine Learning*.

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

[17] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

[18] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. https://doi.org/10.48550/ARXIV.1412.6980

[19] Ankur Kumar. 2022. The Illustrated Image Captioning using transformers. *ankur3107.github.io* (2022). https://ankur3107.github.io/blogs/the-illustrated-image-captioning-using-transformers/

[20] Hwanhee Lee, Thomas Scialom, Seunghyun Yoon, Franck Dernoncourt, and Kyomin Jung. 2021. QACE: Asking Questions to Evaluate an Image Caption. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 4631–4638. https://doi.org/10.18653/v1/2021.findings-emnlp.395

[21] Paul Lerner, Olivier Ferret, and Camille Guinaudeau. 2023. Multimodal Inverse Cloze Task for Knowledge-Based Visual Question Answering. In *Advances in Information Retrieval*, Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo (Eds.). Springer Nature Switzerland, Cham, 569–587.

[22] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755.

[24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://doi.org/10.48550/ARXIV.1907.11692

[25] Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. 2022. Universal Multi-Modality Retrieval with One Unified Embedding Space. https://doi.org/10.48550/ARXIV.2209.00179

[26] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Kumar Gupta, and Marcus Rohrbach. 2020. KRISP: Integrating Implicit and Symbolic Knowledge for Open-Domain Knowledge-Based VQA. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 14106–14116.

[27] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[29] Chen Qu, Hamed Zamani, Liu Yang, W. Bruce Croft, and Erik G. Learned-Miller. 2021. Passage Retrieval for Outside-Knowledge Visual Question Answering. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2021).

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[31] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

[32] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2022. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21, 1, Article 140 (jun 2022), 67 pages.

[33] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, 784–789. https://doi.org/10.18653/v1/P18-2124

[34] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2383–2392. https://doi.org/10.18653/v1/D16-1264

[35] Stephen Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. 1995. Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*. Gaithersburg, MD: NIST, 109–126.

[36] Alireza Salemi, Juan Altmayer Pizzorno, and Hamed Zamani. 2023. A Symmetric Dual Encoding Dense Retrieval Framework for Knowledge-Intensive Visual Question Answering. arXiv:2304.13649 [cs.CV]

[37] Violetta Shevchenko, Damien Teney, Anthony Dick, and Anton van den Hengel. 2021. Reasoning over Vision and Language: Exploring the Benefits of Supplemental Knowledge. In *Proceedings of the Third Workshop on Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*. Association for Computational Linguistics, Kyiv, Ukraine, 1–18. https://aclanthology.org/2021.lantern-1.1

[38] Hrituraj Singh, Anshul Nasery, Denil Mehta, Aishwarya Agarwal, Jatin Lamba, and Balaji Vasan Srinivasan. 2021. MIMOQA: Multimodal Input Multimodal Output Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 5317–5332. https://doi.org/10.18653/v1/2021.naacl-main.418

[39] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. MultiModal{QA}: complex question answering over text, tables and images. In *International Conference on Learning Representations*. https://openreview.net/forum?id=ee6W5UgQLa

[40] 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). Association for Computational Linguistics, Hong Kong, China, 5100–5111. https://doi.org/10.18653/v1/D19-1514

[41] Jizhi Tang, Yansong Feng, and Dongyan Zhao. 2019. Learning to Update Knowledge Graphs by Reading News. 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)*. Association for Computational Linguistics, Hong Kong, China, 2632–2641. https://doi.org/10.18653/v1/D19-1265

[42] Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2022. Generative Language Models for Paragraph-Level Question Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 670–688. https://aclanthology.org/2022.emnlp-main.42

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[44] Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022. GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 2345–2360. https://doi.org/10.18653/v1/2022.naacl-main.168

[45] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2018. FVQA: Fact-Based Visual Question Answering. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 10 (oct 2018), 2413–2427. https://doi.org/10.1109/TPAMI.2017.2754246

[46] Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2022. Multi-Modal Answer Validation for Knowledge-Based VQA. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 3 (Jun. 2022), 2712–2721. https://doi.org/10.1609/aaai.v36i3.20174

[47] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2021. An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA. In *AAAI Conference on Artificial Intelligence*.

[48] Hamed Zamani and W. Bruce Croft. 2017. Relevance-Based Word Embedding. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 505–514. https://doi.org/10.1145/3077136.3080831

[49] Hamed Zamani and W. Bruce Croft. 2018. On the Theory of Weak Supervision for Information Retrieval. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval* (Tianjin, China) *(ICTIR '18)*. Association for Computing Machinery, New York, NY, USA, 147–154. https://doi.org/10.1145/3234944.3234968

[50] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. An Analysis of BERT in Document Ranking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 1941–1944. https://doi.org/10.1145/3397271.3401325