University of Glasgow | School of Computing Science

More than a BLIP in the road

Ben Brown (2455263b)

April 2024

ABSTRACT

Pre-trained large multimodal models, such as BLIP-2, have achieved state-of-the-art results in several general vision-language and text-image retrieval benchmarks. These models improve text-image retrieval as they do not rely on unreliable user-generated captions and can better understand the images' content and style.

This paper investigates applying BLIP-2 for text-image retrieval on the AToMiC dataset, with a task to identify appropriate images for sections of Wikipedia articles. We first look at using BLIP-2 as part of a dense neural retrieval pipeline. As the queries are longer than the associated text the model is trained upon, we also look at summarising the queries with Llama 7b. As Wikipedia differs from most datasets in having reliable captions, we also investigate combining the dense neural approach with classical retrieval techniques and using BLIP-2 to generate captions for retrieval.

We found that the performance of BLIP-2 was similar to other large multitask multimodal models but significantly worse than classical textual retrieval methods (such as BM25) using the provided captions. None of summarising the queries, combining the dense retrieval scores with BM25, and using captions generated from BLIP-2 improve retrieval.

1. INTRODUCTION

Pre-trained large multimodal models such as BLIP-2 [1], FLAVA [2] and BEiT-3 [3] have achieved state-of-the-art results in several general vision-language benchmarks such as visual question answering [4] and image captioning [5].

Text-image retrieval tries to find relevant images for a given textual query. It has a wide variety of uses, including online shopping and product search, library and archive search, accessibility systems for low-visibility users, and multimedia document creation.

These models use a dense neural approach and work by creating embedding vectors for the textual query and all the images to search. They learn both modalities, mapping them into a shared space, allowing them to be compared to score the images for a query.

Previous approaches only used captions for the images and treated the images like text documents. People often do not take care when creating captions, so they often have errors or missing significant details, if present at all. Captions are also highly dependent on context. What is correct for one use of the image may not be suitable for another use. For example, "a cat sitting outside a black door" and "Larry Chief Mouser sitting outside Downing Street" could be captions for the same image but have different information. Images are in a larger information space than text, so a caption can never capture all the information in an image. Models that work with the actual images can be more accurate and draw on much more information in ranking images.

Some large models can also generate text based on the images and a text input query. This means that they can be used to generate captions for retrieval, which are more reliable as they are automatically produced and can overcome some of the context issues from the query. However, the quality of the information in the generated captions is still sometimes lacking, and good captions created by humans can still perform better.

We investigate the performance of using BLIP-2 on the AToMiC benchmark [6] for text-image retrieval. The AToMiC benchmark is a new benchmark designed to mimic the real scenario of working on a document and wanting to find an image for a section you have just written. It is based on data from Wikipedia. We chose BLIP-2 as it has shown excellent performance on other text-image retrieval tasks and can be used to produce embeddings and output text.

We conduct a study to answer the following research questions: **RQ1:** What is the performance of using BLIP-2 for dense retrieval on the AToMiC dataset? **RQ2:** Does summarising the articles (the queries) lead to better performance when using BLIP-2 for retrieval?

Most scenarios do not have reliable captions, and a strength of Large Multimodal Models is being able to operate without captions. However, Wikipedia generally has high-quality and reliable captions. Therefore, we investigate whether the provided captions can be used with BLIP-2 with the following research questions: **RQ3:** Does BLIP-2 improve retrieval performance compared with classical sparse retrieval techniques on the captions? **RQ4:** Does combining the scores from BLIP-2 retrieval with classical sparse retrieval techniques improve retrieval performance? **RQ5:** Do captions generated by BLIP-2 improve retrieval performance when used with text retrieval techniques?

2. BACKGROUND

We give an overview of the AToMiC dataset and task, how dense retrieval approaches with large multimodal models work, the structure of the BLIP-2 model and how it was trained, and an overview of using Large Language Models for the summarisation of articles.

2.1 AToMiC

TREC Authoring Tools for Multimedia Content (AToMiC) [6] is a large-scale image/text retrieval test collection based on Wikipedia data, which focuses on tasks to simulate use cases of creating multimedia documents. When creating documents for the web, users want to be able to find ap-

propriate images to illustrate and improve their work. The task we specifically look at is image suggestion: when given a section of a Wikipedia article, we are to find the best image to accompany it.

AToMiC differs from existing image/text retrieval collections such as MS-COCO [7] and Flickr30k [8] as they only have simplistic descriptions of the images, whereas the article sections tend to be longer. Wikipedia-based Image Texts (WIT) [9] is the basis of the AToMiC dataset; however, it focuses on the image captions rather than associations with articles. The articles also often do not specifically describe what is in the image, which is usually the primary purpose of captions. For example, the Wikipedia article about Marie Curie does not describe what she looked like, but the relevant image is a picture of her. This presents a challenge to learn visual details about a wide range of subjects without having prompts from the query. AToMiC has roughly 11 million images, which is significantly larger than existing datasets such as MS-COCO (roughly 330,000 images) and Flickr30k (approximately 30,000 images). The larger size means that approaches must be more efficient to be completed within a reasonable time.

The dataset is based upon the WIT dataset [9], with train-validation-test splits matching the underlying dataset. Images not associated with a specific passage or no longer available were removed from the dataset. Query relevance judgements were created, associating each article section with the image appearing in it. This means the judgements are sparse, and each article only has one relevant image. This adds difficulty as an article could have many plausible images, but only one is marked as relevant. The judgements also only exist for sections with images, and those sections could be affected by the structure of Wikipedia as sections without images are not considered.

Due to Wikipedia's permissive licensing and high-quality data, it is used in the pre-training phase for most large language and multimodal models. As any models used could have already seen the images or articles used, performance for the task may not be representative when used on unseen images. Wikipedia has strict guidelines for image captions meaning they are usually high quality, accurate, and descriptive. This differs from many real collections, where the captions may be absent, incorrect, or irrelevant.

2.2 Retrieval Approaches

2.2.1 Sparse

Traditional information retrieval approaches with text documents generally use a sparse approach where an inverted index is created, which maps each term to the documents where they occur [10, 11, 12]. This index is then used to score all the documents to rank them for display. This depends on having words to map; thus, this technique cannot be used for multimodal retrieval. However, if appropriate captions exist, they could be used. Traditional sparse retrieval also struggles when the terms do not exactly match what is in the documents.

2.2.2 Dense

Modern dense retrieval techniques take a different approach and have produced state-of-the-art results on textual [13, 14] and multimodal retrieval tasks [1, 3, 15]. They use neural network models to produce dense vector representations (embeddings) of each document in the corpus, where the embeddings are combined to form a large matrix, forming the index for retrieval. The query is also encoded using the deep learning model to produce a vector, which is then compared against the index of all the vector embeddings for documents in the corpus. The comparison gives a score that is used to rank the documents; a standard vector similarity measure such as cosine similarity is often used. As we are just comparing vectors, a model can encode multiple different modalities into the same vector space, and we can query between modalities. However, these vectors are not explainable to humans; thus, it can be hard to reason about performance. These models can learn the meanings of words and concepts so the terms do not have to match exactly. They are also more compute-intensive, so they are often only used as a re-ranker for text retrieval after an initial sparse retrieval.

Dense retrieval approaches can use the models as dual or fusion encoders. Dual encoders encode the query and documents separately, with the vectors compared afterwards. A fusion encoder takes a query and a document together and encodes them together to get a score for similarity, this means that for every document you want to consider you have to encode it with the query. As each document is considered with each query, the accuracy of fusion encoders can be better as they are directly compared, but not all models support use as fusion encoders. However, the computational requirements for fusion encoder models are significantly larger as all pairs must be considered for each query. A dual encoder can pre-compute the representations of all the documents and only needs to compute the representation of the query once at retrieval time. Constructing the index for dual encoders can take a significant time to process all documents, but the task is embarrassingly parallel. Most models support being used as a dual-encoder. For the AToMiC dataset, running each article with over 11 million images through a fusion encoder would not be feasible, whereas vector similarity is significantly quicker.

2.3 BLIP-2

BLIP-2 [1] is a large multimodal model created from existing image encoders and language models. They train a "Querying Transformer" (Q-Former), which is based on the BERT [16] architecture, to sit between the frozen models to learn a shared representation between modalities. It has achieved state-of-the-art performance on various image-text tasks such as visual question answering, image retrieval and captioning.

Using frozen unimodal models, BLIP-2 has fewer trainable parameters and is more compute-efficient than other models with similar performance. As images and text are

Figure 1: Overview of the architecture of BLIP-2. Images are fed through a frozen image encoder before being passed to the Q-Former. Text can also be passed to the Q-former to be encoded. The output of the Q-Former can be used as a shared embedding of text and images or passed to a frozen language model for text generation.

not processed simultaneously, the model is simplified. However, it restricts some use cases as only one modality can be processed at once, meaning that BLIP-2 cannot be used as a fusion encoder. The use of frozen models means that work on unimodal tasks can be applied in more scenarios, and the training does not need to be repeated. They showed that the performance of BLIP-2 improved with better image encoders and large language models. Thus, if the unimodal models are improved, these improvements could be used to improve the performance of BLIP-2.

The frozen image encoder is a vision transformer [17], and the default model is taken from EVA-CLIP [18]. They look at using unsupervised-trained OPT models [19] and the instruction-trained FlanT5 model [20] for the frozen language model. We only consider the OPT models due to computational constraints.

BLIP-2 is pre-trained in two stages, first on vision-language representation learning, so the Q-Former learns the visual representation most relevant to the test. The first stage has three training tasks: contrastive learning, image-grounded text generation from the image, and image text matching. Contrastive learning is training with an image-text pair that is related and a pair that is not related and ensuring that the embeddings of the related pair are closer than the unrelated pair. The image-grounded text generation gets the Q-Former to generate texts for images and compares the output against the provided captions for the image. Image-text matching aims to create fine-grained alignment by using the model to predict if an image and a text are related.

The second stage of training focuses on aligning the Q-former with the frozen language model so the output of the Q-former is understandable by the language model. The output of the Q-former is fed through a fully connected layer to scale the embedding into the same dimension as the language model. As the Q-former is trained to match with the text, it discards information irrelevant to the text output, reducing the need for the LLM to learn alignment and reducing forgetting problems. The image embeddings are prepended to any provided text embeddings. Decoder-based LLMs like OPT are trained on language modelling loss, where the LLM generates text from the visual representation of the Q-former. Encoder-decoder LLMs like FlanT5 are trained with prefix language modelling loss, where the expected text is split into two. The prefix is provided with the generated image embedding to the LLM, and it has to generate the suffix.

BLIP-2 is pre-trained on a range of datasets including MS-COCO [7], CC3M [21], and LAION400M [22]. LAION400M and CC3M include images from Wikipedia. Thus, BLIP-2 could have already seen the images in the AToMiC dataset. Captions for the images are a combination of the original web captions of the images, with ten captions generated by BLIP [23]. A CLIP model then selects the best captions from the original caption and the generated captions. This ensures that all images have a caption and that the caption matches the image. However, important information about the image could be lost, or the caption could be incorrect if the CLIP does not understand the original caption. The style of the captions also differs from AToMiC, with the captions being used more focused on describing the content. Using a single 16-A100(40G) machine full pre-training requires less than nine days for the largest model they considered (ViT-G and FlanT5-XXL).

BLIP-2 struggles to use in-context learning to be applied to other tasks. As it has just been trained on singular image-text pairs, the model has not learned any correlations between sequences of images. This means intensive options such as full fine-tuning are needed to apply it to unseen tasks.

2.4 LLMs for Summarisation

Many information retrieval tasks focus on web search, where queries are provided directly by a user and are often only a couple of terms. The queries from AToMiC differ, being significant sections of Wikipedia pages with over 160 terms on average. In a typical retrieval pipeline, we want to expand the query to get more information, whereas, for our use case, we want to extract the useful parts and reduce the query.

Large language models have succeeded at a wide variety of language tasks, including generating summaries of documents. Models such as GPT-3 [24] and GPT-4 [25] have achieved state-of-the-art results on various textual tasks. However, their architecture and training data have not been published, and the only way to use the models is through OpenAI's paid AI, which is too restrictive for this project and the size of AToMiC.

GPT-2 [26] was state of the art for text generation when published; however, on document summarization, it performed worse than the state-of-the-art models. The model and weight are all open source. The generated text looks at first to be credible summaries, except they often get specific details from the articles wrong. Compared to current models, GPT-2 performs very poorly, and though its output is not very accurate, it could be enough to remove less useful information from the document. When prompting the model to summarise, instead of placing a prompt before the article, they append "TL;DR:". This means that the prompt is closer to the generated tokens, thus the probabilities have a higher weight thus the output should better match the prompt.

Llama-2 [27] has shown the most success at textual tasks of any open-source language model. There are three variants: 7B, 13B, and 70B, indicating the number of parameters for each model. Llama-2 70B's performance at summarisation is generally seen as slightly worse than close-sourced alternatives but more efficient [28], so it has been deployed for meeting summaries. When asked to summarize legal text, Llama-2 70B human evaluators preferred it slightly to the results of a legal intern [29], and Llama-2 70B has similar factual accuracy to GPT-4 [30].

3. RELATED WORK

AToMiC is a new benchmark, and when we started our research, the only available results were the baselines published in the AToMiC paper [6]. Some additional results and approaches have recently been published; we detail them below. Our methodology and approach were unchanged by the insights from these papers, as they were published after we had completed the majority of our work.

3.1 Overview Results from TREC

At TREC 2023, an overview of results was presented [31], where annotators judged the top 25 candidates from each submitted approach for 70 chosen queries to give richer an-

notations to compare against. These judgements are not currently available.

The best-performing model at the image suggestion task is a learned sparse model, which adapts a CLIP model by training Multilayer Perceptron and Masked Language Model heads. Adapting the model took 8 hours, and we do not have any more details on their approach. The dense approaches, which only consider the image and not the captions, perform worse than approaches that use the captions. All the approaches using only the image perform worse than traditional sparse retrieval with BM25. This could arise from the annotators who may have favoured images with English captions due to the difficulty of judging relevancy or may be from the images not containing sufficient information for the task. The results show that captions will need to be used as part of the retrieval, with some hybrid models performing better than just using captions alone.

3.2 Learned Sparse Retrieval

The team from the University of Amsterdam explored using a multimodal learned sparse retrieval approach using a DistilBERT model [32]. Both the queries and the documents are encoded using a transformer architecture, and they consider both multi-layer perception (MLP) and a masked language model (MLM) sparse projection heads on the top of the model.

Due to resource limitations, they only considered images with English captions. This will improve the performance on the AToMiC test collection, as all the articles are in English and only have one relevant image with English captions. It may have reduced the performance on the judgements at TREC, as additional query relevance judgements are made with some images that do not have English captions.

They only considered the image suggestion task and used the "Page Title", "Section Title", and "Section Description" fields as the query. They tried all the combinations of the image pixel values and the "caption reference description" field for the captions to search. They dismissed using the MLM encoder for the query as it produced terms humans could not interpret, as the vocabulary was not constrained. They also found it could lead to the output being highly coactivated, which could harm the retrieval efficiency.

They found that encoding the caption with the MLP encoder and the image with the MLM encoder and combining the outputs produced the best results. However, using only the caption encoded with the MLP encoder is only marginally worse. Using just the MLM encoder on the image produced very poor results, as it could not match the specific facts required for the queries. They do not report if their results are statistically significant against each other.

3.3 Text2Pic Swift

The team from the University of Glasgow used a two-stage retrieval pipeline. The first stage embeds each entity in the query using a multimodal model and compares the embedding against an index of image embeddings to produce the top candidates for each entity. For the second stage, they summarise the query with BART [33] and then embed the summary and compare it against all the previously found candidates to find the best approach.

They use a modified version of the BEiT-3 model to encode queries. The original BEiT-3 model is based on multiway transformers [3] with a shared self-attention module and then a different feed-forward network for each different modality (which are known as modality experts). They split the vision-language expert from the model, freeze its weights and the shared self-attention weights, and only finetune the language expert. This allows them to use the model as a dual encoder rather than a fusion encoder, which has to consider every query image pair individually. BEiT-3 is optimised for captions, which are very different from the long article queries from AToMiC, and dealing with text and images separately can better suit the task. It also reduces the compute needed as the new model has 30% fewer parameters. Freezing the image encoder means the precomputed index of image embeddings does not need to be updated between finetuning steps.

Finding candidates for all entities in the query ensures that all possible images should be considered, even if the query is not specific about what the image would expect to include.

They do not explicitly state which fields of the query they use; however, as they are using summarisation, they must use one of the description fields to have enough text to summarise. They also do not specify the performance of comparing only the summarised query embedding without the first stage. Nor do they mention how long it took to generate the image embeddings.

They achieve a 99% reduction in retrieval duration by encoding the images and queries separately and using vector similarity comparisons. They achieve an 11% increase over other dense approaches using multimodal models. However, their results are still significantly worse than those of classical textual retrieval methods.

4. METHOD

We introduce our retrieval pipeline, the metrics we are looking at, and how we handle captions.

4.1 Base Retrieval Pipeline

We use a dense retrieval approach, using BLIP-2 to convert the articles and queries into dense vectors and then compare them. We chose BLIP-2 as a model which has shown state-of-the-art performance on other multimodal tasks and is less resource-intensive than other models. This approach focuses on the Image Suggestion task from AToMiC.

First, we create an index for all the images in the AToMiC dataset. The indexing code is based upon the PyTerrier-DR library [34]. We pass the raw image pixel values to BLIP-2 to create an embedding for each image, using batches of 50 to utilise the GPU fully. We load the BLIP-2 model and weights using the LAVIS library [35]. Our approach does not use any of the image captions or other fields. The indexing for the full AToMiC images dataset took approximately 75 hours, and the index is approximately 11.5GB.

For retrieval, we use PyTerrier [36] to manage the process and calculate the retrieval metrics. For each article, we feed the text from the article we want to use for the query into BLIP-2 to produce an embedding. The query embedding is then compared with cosine similarity against all the image embeddings in the index to give each image a score. That score is then used to rank each image.

4.2 Metrics

In a "real" scenario, an article creator who wants to add an image to their article would be presented with relevant images. Ideally, the best image should be the first image presented to them, but if the image appears in the first panel, that is the next best thing. Thus, the primary metric we want to consider should weigh the top options higher. To match the AToMiC paper [6], we use mean reciprocal rank (MRR) [37], with a cutoff of 10 to match a sensible number of images which could be displayed to the user. We also consider recall with the same cutoff of 10 to see if the "correct" image would be displayed to the user.

In AToMiC, each article only has one relevant image. However, there may be many other relevant images for the article. Thus, the top 10 images might be too tight to see if the retrieval system is working effectively, as the results may be very relevant but have no relevance labels. We then also look at recall with a cutoff of 1000 to see if the system can identify a collection of relevant images, even if the ranking is not correct. AToMiC has 11 million images; thus, if it is returned in the top thousand results, it is ranked in the top 0.009% images.

4.3 Captions

For approaches that consider captions, we construct text from all three available caption fields (Reference Description, Alt. Text, and Alt. Text Description). We concatenate the fields together, removing any duplicates where the fields contain the same content. It is inconsistent which field has critical information for identifying the images, so we combine the fields and use all languages to ensure that almost no images do not have a caption.

Our approach of using all languages differs from that used in existing work [6, 38], where only English captions are considered. When dealing with proper nouns (like many of the articles have), they will be the same in most languages, allowing more images to be considered, so the best image is more likely to be found. However, all the queries are in English, thus all images with query scores will have English captions so overall metrics could be lower.

4.4 Experimental Environment

All experiments were run on the University of Glasgow IDA cluster on a Nvidia 3090 GPU with 32GB of RAM. Experiments were run through a Jupyter Notebook running Python 3.8.13. We use PyTerrier 0.9.2 and LAVIS 1.0.2.

5. RESULTS & ANALYSIS

We present empirical results for using BLIP-2 for retrieval on the AToMiC dataset to answer our research questions. The questions fall into two categories: those which consider approaches that only use the images provided and those which also use the captions provided.

5.1 Without Captions

We first consider approaches that only use images for retrieval, not captions, making the approach more applicable to other scenarios where high-quality captions are unavailable.

RQ1: What is the performance of using BLIP-2 for dense retrieval on the AToMiC dataset?

The LAVIS library offers different variants of BLIP-2. There is the base pre-trained model, a model finetuned on the MS-COCO dataset [7], and a model using the smaller ViT-L image encoder from CLIP [15] rather than the default ViT-G model from EVA-CLIP [18].

We evaluated each model variant on the AToMiC test collection. Instead of using an index of all the images in the dataset, we only used the images with relevancy judgements for any query. This significantly reduces the time to create the index from roughly three days to under an hour. However, it is still a representative task, as there are still nearly 10,000 images to search. We used the main text fields of the articles for the queries, concatenated with a single space in the order they appear (Page Title, Page Description, Section Title, Section Description).

Table 1 shows that the pre-trained model performs better than the other variants across all metrics examined. As expected, the smaller ViT-L image model is less capable; thus, the results are lower. The MS-COCO dataset is focused on image captions, and as such, the text related to each image is very short, whereas the AToMiC dataset has much longer text. As finetuning the model improves its performance for the exact task it is trained on against the detriment of general performance, it makes sense that the results are worse after finetuning on a differing dataset.

Table 1: Mean Reciprocal Rank at 10, and Recall at 10 and 1000 for each model variant on the AToMiC test collection, considering only images with a relevancy judgement. Using a paired t-test ($p < 0.05$), all results are statistically significant with the base pre-trained model as a baseline.

Model Variant	MRR @ 10	R @ 10	R @ 1000
Pre-trained	0.313177	0.891826	0.160017
Finetuned MS-COCO	0.213106	0.793680	0.108798
ViT-L	0.250785	0.852223	0.128897

When referring to the BLIP-2 model in all future experiments, we refer to the BLIP-2 model loaded from the LAVIS library with the base pre-trained weights, as it has shown the best baseline results.

Article Fields

The articles provided in the AToMiC dataset have a variety of different fields which can be used. AToMiC article queries are not full Wikipedia articles. They only contain the overall title (Page Title), description (Page Description), and section of the articles from which the image came (Section Title and Section Description). The categories of the Wikipedia page are also given. A page can be part of many categories, but the categories do not form a tree structure. The categories are quite specific; for example, there is an "Albert Einstein" category.

We hypothesise that a subset of fields will yield the best performance, as there is a large amount of data if all fields are used, which could overwhelm the valuable parts of the article. We consider all combinations of fields to see which produces the best retrieval results. Fields were concatenated together, separated with a single space, and are in the order they would appear on a Wikipedia page. The list of

Table 2: Retrieval results of BLIP-2 retrieval on the AToMiC test collection for each combination of the article fields used as a query.

Page Title	Page Description	Section Title	Section Description	Category	MRR @ 10	R @ 10	R @ 1000
1					**0.00998**	**0.01661**	0.08731
	1				0.00152	0.00375	0.05277
		1			0.00028	0.00071	0.01388
			1		0.00595	0.01185	0.12570
				1	0.00200	0.00598	0.08235
1	1				0.00981	**0.01661**	0.09855
1		1			0.00844	0.01540	0.09207
1			1		0.00659	0.01378	**0.13046**
1				1	0.00524	0.01256	0.11050
	1	1			0.00172	0.00395	0.05338
	1		1		0.00555	0.01165	0.12813
	1			1	0.00220	0.00669	0.09754
		1	1		0.00544	0.01205	0.11942
		1		1	0.00241	0.00679	0.08478
			1	1	0.00530	0.01074	0.12114
1	1	1			0.00792	0.01428	0.09146
1	1		1		0.00630	0.01337	0.12539
1	1			1	0.00545	0.01297	0.11445
1		1	1		0.00650	0.01418	0.12428
1		1		1	0.00552	0.01307	0.11293
1			1	1	0.00610	0.01246	0.12529
	1	1	1		0.00509	0.01134	0.12296
	1	1		1	0.00213	0.00608	0.09308
	1		1	1	0.00486	0.01033	0.12327
		1	1	1	0.00448	0.01033	0.11547
1	1	1	1		0.00619	0.01367	0.11911
1	1	1		1	0.00534	0.01195	0.11111
1	1		1	1	0.00580	0.01236	0.12104
1		1	1	1	0.00576	0.01307	0.12114
	1	1	1	1	0.00412	0.00952	0.11932
1	1	1	1	1	0.00542	0.01256	0.11597

categories is concatenated with a single space and then considered like any other field. We evaluate using the AToMiC test collection against an index of all images in the dataset.

Table 2 shows that only using the "Page Title" field performs best when looking at a cutoff of 10. The best results occur when the text describes what you expect to see in an image. Most of the images are of what is described in the title, and the short nature of the page title matches the short captions BLIP-2 is trained on. However, the overall page title sometimes does not match the image, as the image is related to a specific part of the section. When combined with the "Page description", the recall does not drop, but the mean reciprocal rank decreases, indicating that the description does not help differentiate between the top-ranked images.

The performance for the "Section Title" field is poor when considered on its own and when combined with other fields. This is likely due to this often being a very generic title such as "Life" or "Introduction".

When considering recall at 1000 table 2 shows "Page Title" combined with "Section Description" performs the best. The recall at 1000 is better for all the combinations which include the "Section Description" than without it. This indicates that the field has essential information for selecting the correct image but contains irrelevant terms, as performance is reduced when considering a cutoff of 10.

Table 3 shows when the top performing field combinations are used as the queries for BLIP-2 show similar performance against the baseline results for other multi-task multimodal models such as BLIP [23] and FLAVA [2] published in the AToMiC paper [6]. The published baselines use all the available fields concatenated together.

Table 3: Retrieval metrics for the top performing field combinations against results from AToMiC paper [6].

Model Variant	MRR @ 10	R @ 10	R @ 1000
Page Title	0.010	0.0166	0.0873
Page Title and Section Description	0.007	0.0138	0.1305
All Fields	0.005	0.0126	0.1160
BLIP ViT-B-32	0.008	0.0142	0.1246
FLAVA ViT-B-32	0.004	0.0114	0.1973

RQ2: Does summarising the articles (the queries) result in better performance when using BLIP-2 for retrieval?

The "Section Description" field has valuable information for retrieval, as performance improves at a cutoff of 1000; however, it reduces the performance at a cutoff of 10. This indicates that the field is often needed to identify images successfully; however, there are unhelpful extra terms. We hypothesise that we can use a large language model to summarise the field so it contains only essential information. We use Llama-2 7B [27] as a leading open-source large language model, with no limitations on how many articles we could use it on as we could run it ourselves.

We concatenate all relevant article fields (Page Title, Page Description, Section Title, Section Description, Category) and pass to the Llama-2 model with the prompt "Summarise the following article:". Summarising all the prompts took a little over a day to process.

Large language models are very sensitive to their prompts, and changing a single word can change their performance [39]. Thus, we also try an alternate prompt, and as we are focussed on selecting an image, we prompt the model with "Describe what you would expect to see for an image of the following article:". This should also allow the model to add details from its knowledge about the appearance of items which might not be in the article whilst removing anything irrelevant to an image.

Table 4 shows the results for the summarised and described articles are worse than the best-performing raw field combinations from Table 2 on the AToMiC test collection. We also look at combining the summaries with the page title. All the summaries perform worse than the baselines using the standard unaltered fields using a cutoff of 10. Only when the summaries are combined with the page titles is recall at 1000 improved over using only the page titles. They are

Table 4: Retrieval metrics for articles after they have been summarised or described by Llama-2 and combining the summary/ description with the page title against the best combinations of unchanged fields on the AToMiC test collection. * and ˆ denote non-statistically significant results against the "Page Title" and the "Page Title and Section Description" baselines, respectively, both using a paired t-test ($p < 0.05$).

Model Variant	MRR @ 10	R @ 10	R @ 1000
Page Title	0.00998	0.01661	0.08731
Page Title and Section Description	0.00659	0.01378	0.13046
Summarised	0.00261	0.00618	0.07799
Titles and Summarised	0.00420	0.00922	*0.09339
Described	0.00245	0.00608	0.07819
Titles and Described	ˆ0.00431	0.01033	0.09754

well below the combined page title and section description performance. It should be expected that the titles, combined with the summaries, perform better at recall at 1000, as all combinations, including the section description field, perform better at this metric.

The summaries generated are generally of poor quality and typically remove information about what would be in a picture, and the descriptions do not seem to differ from the summaries substantially. Sometimes, the model continues with the Wikipedia article rather than summarising it. As the model is trained as a word prediction task with Wikipedia in the training data, it has just memorised the article and added unhelpful information. These findings match the poor results when used for retrieval. We are considering the smallest variant of Llama-2; if a larger, more capable model was used, results may be more successful. Most of the successful retrieval results with Llama-2 have used the larger 70B variant rather than the 7B variant; however, we can only use the smallest variant due to limited resources.

Further Analysis

We analysed the results of our baseline dense retrieval pipeline using various techniques to find where it performs well or poorly and identify possible improvements.

We used K-Means from FAISS [40] to produce clusters to see if there were any specific trends with the embeddings produced. We produced clusters across a wide range of K values (100, 250, 500, 750, 1000) based on the image embeddings and then based on the query embeddings. We found no correlation between clusters and queries ranking successfully in the top 10 or 1000.

We looked at the length of the query to see if there was an optimum length for the query. We considered using all the relevant article fields concatenated together as the query, using our baseline dense retrieval pipeline on the AToMiC test collection. We placed each query into one of 25 bins based on their lengths and calculated the average retrieval statistics for each bin. Figure 2 shows retrieval performance shows no real trends against query length. No specific lengths seem to perform better than others. Overall, the shortest queries and the longest queries do slightly worse. However, the results are dependent on the individual queries in each bin. The overall trend is likely from short queries missing some information and longer queries containing too many irrelevant terms.

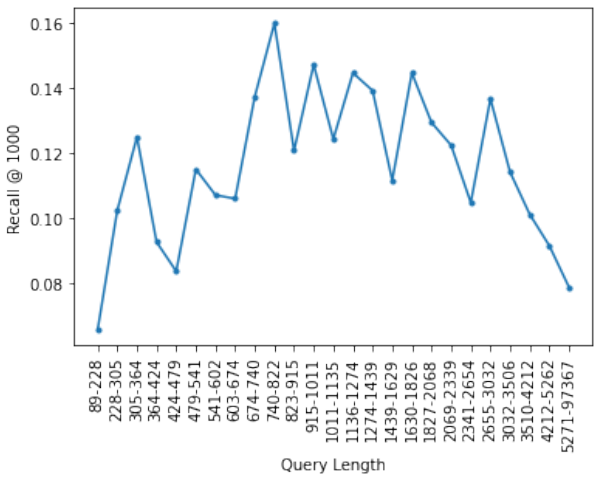

Figure 2: Plot of Recall @ 1000 against query length, grouped into 25 bins. Ran with the baseline pipeline, using all relevant article fields concatenated as the query on the AToMiC test collection.

5.2 With Captions

As AToMiC is based on Wikipedia and, as such, has reliable and accurate captions, we will next consider approaches that use captions.

RQ3: Does BLIP-2 improve retrieval performance compared with classical sparse retrieval techniques on the captions?

Table 5 shows that classical textual retrieval techniques such as TF-IDF and BM25 perform significantly better against BLIP-2 on all metrics considered. Showing poor results against simple retrieval techniques indicates that BLIP-2 alone is unsuitable for use on the AToMiC task and that captions should be considered. This matches others' findings [31] that the most successful approaches all use the captions provided.

Table 6 shows the number of queries where each system performs best. As indicated by the Recall and Mean Reciprocal Rank, using BM25 on captions performs much better, but BLIP-2 outperforms it in a few cases. Most cases where

Table 5: Retrieval metrics for BM25 and TF-IDF retrieval on the captions against BLIP-2 retrieval using the images. T is a system using the "Page Title" field as the query, and T+D uses both the "Page Title" and "Content Section Description" fields. All results are statistically significant results against the matching BLIP-2 results, using a paired t-test ($p < 0.05$).

Model Variant	MRR @ 10	R @ 10	R @ 1000
BLIP-2 (T)	0.0100	0.0166	0.0873
BLIP-2 (T+D)	0.0066	0.0138	0.1305
BM25 (T)	0.3518	0.4991	0.6906
TF-IDF (T)	0.4992	0.6906	0.6906
BM25 (T+D)	0.3052	0.4232	0.5965
TF-IDF (T+D)	0.3065	0.4232	0.5933

Table 6: Comparison of per-query performance of BM25 retrieval using captions against BLIP-2 dense retrieval on the AToMiC test collection using the "Page Title" field as the query.

Scoring Method	BM25 Better	BLIP-2 Better	Equal
MRR @ 10	4927	77	4869
R @ 10	4885	60	4928
R @ 1000	6467	176	3230

they are equal happen when both systems fail to provide any results.

Figure 3 shows some examples of expected images and queries where BLIP-2 does better. Most of the time, when BLIP-2 does better, the expected image contains text. Of the 77 queries where MRR is improved, only nine do not include text in the image. The cases where there is text can be signs, logos, or magazines and can be difficult to read or presented strangely. The majority of the non-text cases which do well contain something very obviously referencable by the text, for example "Burrito", "Roll of Tape", or "Cow Chocolate" (a chocolate bar with a cow stamped into it). There are a couple of cases in which the model must have already seen the images, as it can correctly identify specific people within the images. For example, it identifies the actors "Kit Harrington" and "Iran-Venezuela Relations". BLIP-2 generally has a weakness for these kinds of queries, and for similar images, if you asked it to generate a caption, it would describe the person rather than their name or name someone wrong.

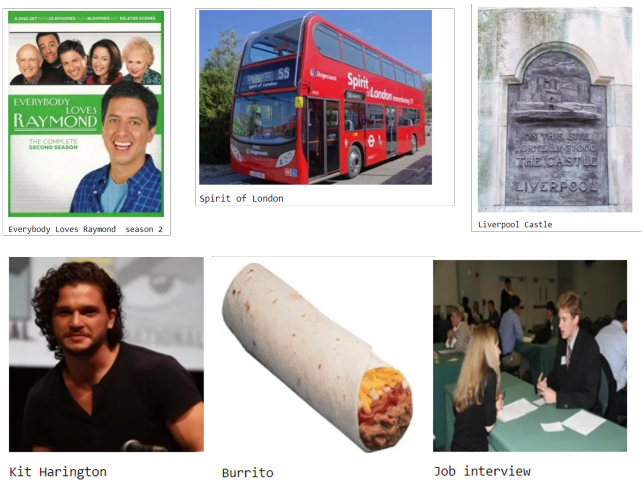

Figure 3: A sample of images which BLIP-2 can identify from the given query better than BM25

RQ4: Does combining the scores from BLIP-2 retrieval with classical sparse retrieval techniques improve retrieval performance?

Some of the contents of the captions are incorrect or don't provide any helpful information, or in a very small number of cases (0.00003%), there is no caption. Some images also only have captions in a foreign language, but all the queries are in English.

We look at whether a linear combination of our dense approach using BLIP-2 with classical retrieval using BM25 on the captions can improve performance. The BLIP-2 scores due to using cosine similarity are always between 0 and 1; thus, to combine with BM25, we perform MinMax normalisation for the scores of each query.

Figure 4 shows there is an incredibly small performance improvement (net around four queries improved) when a small amount of the BLIP-2 score is included; however, the improvement is not statistically significant (using a paired t-test, $p < 0.05$). When most of the score comes from BLIP-2, the performance drops quite steeply; the cliff in performance is around 70% as the BM25 scores, despite being normalised, tend to be larger than the BLIP-2 scores.

Figure 5 shows that including any of the scores from BLIP-2 reduces the system's recall. The reduction is statistically significant for all values.

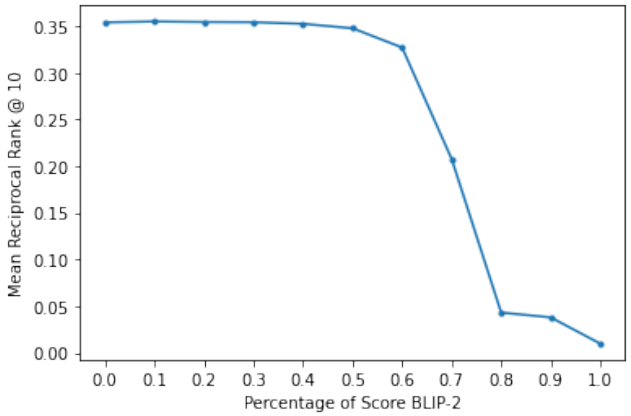

Figure 4: Mean Reciprocal Rank @ 10 against how much of the score came from the BLIP-2 model. "Page Title" was used as the query on the AToMiC test collection.

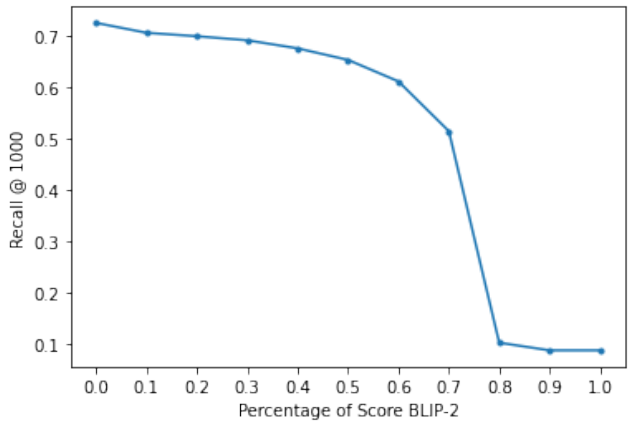

Figure 5: Recall @ 1000 against how much the score came from the BLIP-2 model. "Page Title" was used as the query on the AToMiC test collection.

RQ5: Do captions generated by BLIP-2 improve retrieval performance when used with text retrieval techniques?

One of the primary training tasks for BLIP-2 was generating captions for images. Thus, we investigated if captions produced by BLIP-2 can improve the retrieval performance when combined with the given captions. We use the Opt 2.7 model [19] as the frozen text model, with the "pre-trained" weights for BLIP taken from LAVIS. This model and weights appeared to provide the best captions whilst balancing the compute required for the options available.

As previously identified, the BLIP-2 model does well at identifying text within images and performs better than classical retrieval on the captions for this case. When passing the image to BLIP-2 to generate captions, we prompt the model with: "Question: what does the text in this image say? Answer:". This format matches the format used when training the model, and when testing, if the image doesn't contain any text, it falls back to describing what is in the image. The prompt encourages the model to focus on text if that is available, as it is usually very relevant for AToMiC.

The generation of captions took a significant amount of time, around 600ms per image. This is significantly longer than it took to generate the embeddings, which is expected as the frozen language model is added to create a much larger model. If we were to generate captions for all the images, it would have taken around 70 days. Thus, we only generated captions for the top 100 images returned for each query by BM25, which took around four days to process.

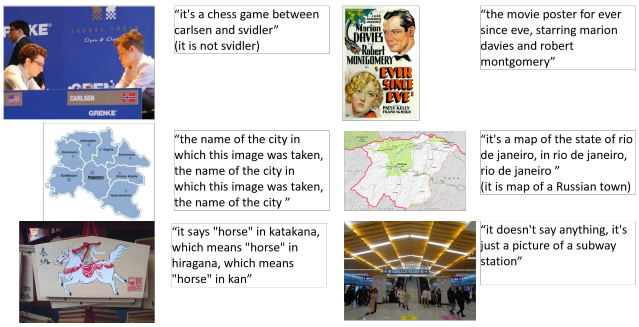

Figure 6: A selection of captions generated by BLIP-2 for images in AToMiC, showing varying levels of success.

The quality of the captions is quite mixed, with some providing good descriptions of the images given, others seemingly good captions but with important facts incorrect, and some incoherent. If the model is unsure, it often repeats a word or phrase several times. Figure 6 shows some examples of the captions generated for images.

Table 7 shows when the generated captions are used with BM25 to rerank the top 100 results from the standard captions, the results are significantly worse than when using the standard captions. All of these documents have captions generated. As the author uploads the images, they will likely use similar language for the image captions as they have used in the article, whereas the generated captions could use different phrasing, which BM25 cannot account for.

When the generated captions are appended to the existing captions to give an augmented index, the performance decreases only slightly (yet statistically significant).

Table 7: Retrieval metrics for the different approaches using generated captions, all methods use BM25. All results are statistically significant results against BM25 on the original captions, using a paired t-test ($p < 0.05$).

Indexed	MRR @ 10	R @ 10
Original Captions	0.371773	0.526790
Reranking Generated Captions	0.055660	0.129343
Augmented Captions	0.365388	0.520916

Neural Rerankers

As we have only considered simple textual retrieval methods, we examine more modern neural text retrieval techniques, which should not suffer from vocabulary mismatch and other fundamental issues with simple classical retrieval techniques like BM25.

We use MonoT5 [41] to rerank the results from BM25. MonoT5 scores each document query pair to score how closely they match each other. We use the default model from the PyTerrier plugin trained on MS-MARCO, and we rerank only the top 30 results due to the amount of available memory. Table 8 shows the performance of the reranked results is better than the standard results from BM25 for the standard and augmented captions. The augmented captions perform slightly worse than the standard captions. The model has a limit of 512 input tokens, and the generated captions are appended to the existing captions, so some are being cropped due to length, reducing performance. The mean response time is also significantly increased, at 783ms for the MonoT5 pipeline and 20ms for the standard BM25.

We also looked at using EPIC [42], an expansion-based neural retrieval method, to rerank the results. We rerank the top 30 results for BM25 using EPIC with the default model loaded by ONIR [43], which is trained on MS-MARCO. Again, this threshold was chosen due to the amount of available memory. Table 8 shows the performance of the reranked results is worse than the standard results from BM25. The difference between the standard and augmented captions being reranked by EPIC is small and not statistically significant.

Using models fine-tuned on the AToMiC dataset could likely improve the results, as the length of the text is longer than what they are trained upon.

Table 8: Retrieval metrics for the different neural reranking methods. All results are statistically significant results against the "BM25 (Captions)" baseline, using a paired t-test ($p < 0.05$).

Indexed	MRR @ 10	R @ 10
BM25 (Captions)	0.371773	0.526790
BM25 (Augmented)	0.365990	0.521118
Mono T5 Reranking (Captions)	0.403900	0.555859
Mono T5 Reranking (Augmented)	0.402180	0.555150
EPIC Reranking (Captions)	0.134394	0.288970
EPIC Reranking (Augmented)	0.134061	0.288869

6. DISCUSSION

Our dense pipeline using BLIP-2 performs very poorly, even simple classical retrieval techniques using captions perform significantly better. BLIP-2 has similar performance on the task to other similar dense models. Most of these models (including BLIP-2) are trained on tasks where the text associated with the images describes what is in the image. This differs from the AToMiC task, where the captions are more related to what is factually in the image or related background on the object. There could be future work to fine-tune BLIP-2 for AToMiC; however, the past work [6] has shown fine-tuning has not been successful for similar models.

The time taken to index all the images in the collection is significant. However, it is embarrassingly parallel, so the time could be reduced significantly if more GPUs were available. The indexing stage also only needs to be done once and then can be reused for all users, and if users have private collections of images, they are likely to be significantly smaller.

AToMiC is a very particular task, which to do well without using the captions requires the models to have a large amount of knowledge about the world. An article about a celebrity is unlikely to describe their physical appearance, thus to be able to find them the model must have been trained on images of that celebrity. BLIP-2 has success on a few of these kinds of queries, which indicates that this type of model could be successful but would require a larger model with significantly more training to learn all these details. This level of knowledge is likely impossible for a human to learn across such a wide range of subjects but should be possible for a large model.

Wikipedia forms the basis of training data for most models being high-quality and open data, and thus, their performance on AToMiC might not be representative of other scenarios or when acting with private collections of images. Finding other sources of information to train the model with reliable and accurate captions could be difficult or costly. The compute and effort to train a larger model specifically for the task may also not be worth it. Compared to classical retrieval techniques, which only use the captions, BLIP-2 uses orders of magnitude more compute, takes significantly longer, and produces significantly worse results.

A small increase in retrieval results could likely be found by following an approach similar to Text2Pic Swift, with candidates being identified for each entity and then compared against a summary. Even only using a summary could improve performance, as there would be less ambiguity in what was being looked for, as what refers to the image in the article differs between queries. For some, the title is enough to identify the image, whereas for other queries, the full description is needed to choose the correct image. The summaries, however, need to be accurate, and the small Llama model we considered showed that they need to be produced by a larger model, which requires more compute than we had available.

How AToMiC is constructed makes it a very hard task to do well. Each query only has one relevant image marked, and with 11 million images, many queries could have more than one image that could be correct. From the data alone, a human can't discern the "correct" choice between two images for some queries. Thus, a pipeline could return many relevant images but get a very low score. The full relevancy judgements from TREC 2023 [31] would present a fairer overall judgement; however, they are not publicly available.

The task also has much greater factual requirements than most multimedia document creation tasks. A creator may only require a stock image to illustrate a general concept or fill some space when creating a document. For AToMiC, the exact right image (for example, a specific tree) must be chosen. This correctness is important for Wikipedia but might not be as important for other tasks.

As AToMiC may not be the most representative task, alternative datasets with the same aim could be used to simulate text-image retrieval during document creation. To increase the performance on the AToMiC benchmark, other large models that consider both captions and images simultaneously (unlike BLIP-2) should be used to make the best use of all available information.

7. CONCLUSION

In this paper, we investigated the performance of using BLIP-2 for text-image retrieval on the AToMiC dataset. We found that using BLIP-2 as a dense neural retriever on this dataset had very poor performance, falling far behind classical textual retrieval methods, like BM25, using the image captions provided (RQ3). However, high-quality captions are unavailable for most datasets, and BLIP-2 has similar performance to other multi-task multimodal models (RQ1).

Summarizing or describing the articles using a large language model did not improve retrieval performance with BLIP-2 (RQ2). Combining BLIP-2's scores with BM25 retrieval on the captions did not lead to meaningful gains (RQ4). Finally, using captions generated by BLIP-2 failed to improve retrieval performance over using the provided human-written captions (RQ5).

Our findings suggest that while BLIP-2 achieves state-of-the-art results on some multimodal benchmarks, it is not well-suited to the AToMiC dataset. Classical text retrieval methods remain highly competitive, likely because the captions already capture much of the critical information needed to match articles to relevant images. For this task, complex multimodal models like BLIP-2 provide little additional value over simpler text-based approaches when reliable captions exist.

8. REFERENCES

[1] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, June 2023. arXiv:2301.12597 [cs].

[2] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A Foundational Language And Vision Alignment Model, March 2022. arXiv:2112.04482 [cs].

[3] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks, August 2022. arXiv:2208.10442 [cs].

[4] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in

Visual Question Answering, May 2017. arXiv:1612.00837 [cs].

[5] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8947–8956, October 2019. arXiv:1812.08658 [cs].

[6] Jheng-Hong Yang, Carlos Lassance, Rafael Sampaio de Rezende, Krishna Srinivasan, Miriam Redi, Stéphane Clinchant, and Jimmy Lin. AToMiC: An Image/Text Retrieval Test Collection to Support Multimedia Content Creation, April 2023. arXiv:2304.01961 [cs].

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context, February 2015. arXiv:1405.0312 [cs].

[8] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, December 2014.

[9] 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*, pages 2443–2449, July 2021. arXiv:2103.01913 [cs].

[10] Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems*, 20(4):357–389, October 2002.

[11] Stephen Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *TREC*, pages 109–126, January 1995.

[12] Stephen Robertson and Hugo Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, April 2009.

[13] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval, October 2020. arXiv:2007.00808 [cs].

[14] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. Distilling Dense Representations for Ranking using Tightly-Coupled Teachers, October 2020. arXiv:2010.11386 [cs].

[15] 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. Learning Transferable Visual Models From Natural Language Supervision, February 2021. arXiv:2103.00020 [cs].

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. arXiv:1810.04805 [cs].

[17] 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. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. arXiv:2010.11929 [cs].

[18] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA: Exploring the Limits of Masked Visual Representation Learning at Scale, December 2022. arXiv:2211.07636 [cs].

[19] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open Pre-trained Transformer Language Models, June 2022. arXiv:2205.01068 [cs].

[20] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling Instruction-Finetuned Language Models, December 2022. arXiv:2210.11416 [cs].

[21] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[22] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs, November 2021. arXiv:2111.02114 [cs].

[23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, February 2022. arXiv:2201.12086 [cs].

[24] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess,

Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, July 2020. arXiv:2005.14165 [cs].

[25] OpenAI. GPT-4 Technical Report, March 2023. arXiv:2303.08774 [cs].

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

[27] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. arXiv:2307.09288 [cs].

[28] Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan TN. Building Real-World Meeting Summarization Systems using Large Language Models: A Practical Perspective, November 2023. arXiv:2310.19233 [cs].

[29] Justin Olsson and Waleed Kadous. LLM-based summarization: A case study of human, Llama 2 70b and GPT-4 summarization quality, November 2023.

[30] Waleed Kadous. Llama 2 vs. GPT-4: Nearly As Accurate and 30X Cheaper, August 2023.

[31] Jheng-Hong Yang, Carlos Lassance, Rafael Sampaio de Rexende, Krishna Srinivasan, Miriam Redi, Stephane Clinchant, and Jimmy Lin. TREC2023 AToMiC Overview, 2023.

[32] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, February 2020. arXiv:1910.01108 [cs].

[33] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, October 2019. arXiv:1910.13461 [cs, stat].

[34] Xiao Wang, Sean MacAvaney, Craig Macdonald, and Iadh Ounis. An Inspection of the Reproducibility and Replicability of TCT-ColBERT. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pages 2790–2800, New York, NY, USA, July 2022. Association for Computing Machinery.

[35] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C.H. Hoi. LAVIS: A One-stop Library for Language-Vision Intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–41, Toronto, Canada, July 2023. Association for Computational Linguistics.

[36] Craig Macdonald and Nicola Tonellotto. Declarative Experimentation in Information Retrieval using PyTerrier. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, pages 161–168, September 2020. arXiv:2007.14271 [cs].

[37] Paul B. Kantor and Ellen M. Voorhees. The TREC-5 Confusion Track: Comparing Retrieval Methods for Scanned Text. *Information Retrieval*, 2(2):165–176, May 2000.

[38] Thong Nguyen, Mariya Hendriksen, and Andrew Yates. Multimodal Learned Sparse Retrieval for Image Suggestion, February 2024. arXiv:2402.07736 [cs].

[39] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. GPT understands, too. *AI Open*, August 2023.

[40] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs, February 2017. arXiv:1702.08734 [cs].

[41] Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models, January 2021. arXiv:2101.05667 [cs].

[42] Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder. Expansion via Prediction of Importance with Contextualization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1573–1576, July 2020. arXiv:2004.14245 [cs].

[43] Sean MacAvaney. OpenNIR: A Complete Neural Ad-Hoc Ranking Pipeline. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, pages 845–848, New York, NY, USA, January 2020. Association for Computing Machinery.