Multimodal Inverse Cloze Task for Knowledge-based Visual Question Answering

Paul Lerner, Olivier Ferret, Camille Guinaudeau

Multimodal Inverse Cloze Task for Knowledge-based Visual Question Answering*

Paul Lerner[1], Olivier Ferret[2], and Camille Guinaudeau[1]

[1] Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
{paul.lerner,camille.guinaudeau}@lisn.upsaclay.fr
[2] Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
olivier.ferret@cea.fr

Abstract. We present a new pre-training method, Multimodal Inverse Cloze Task, for Knowledge-based Visual Question Answering about named Entities (KVQAE). KVQAE is a recently introduced task that consists in answering questions about named entities grounded in a visual context using a Knowledge Base. Therefore, the interaction between the modalities is paramount to retrieve information and must be captured with complex fusion models. As these models require a lot of training data, we design this pre-training task from existing work in textual Question Answering. It consists in considering a sentence as a pseudo-question and its context as a pseudo-relevant passage and is extended by considering images near texts in multimodal documents. Our method is applicable to different neural network architectures and leads to a 9% relative-MRR and 15% relative-F1 gain for retrieval and reading comprehension, respectively, over a no-pre-training baseline.

Keywords: Visual Question Answering · Pre-training · Multimodal Fusion.

1 Introduction

Knowledge-based Visual Question Answering about named Entities (KVQAE) is a challenging task recently introduced in [50]. It consists in answering questions about named entities grounded in a visual context using a Knowledge Base (KB). Figure 1 provides two examples of visual questions along with relevant visual passages from a KB. To address the task, one must thus *retrieve* relevant information from a KB. This contrasts with standard Visual Question Answering (VQA, [1]), where questions target the content of the image (e.g. the color of an object or the number of objects) or Knowledge-based VQA (about coarse-grained object categories) [40], where one can rely on off-the-shelf object detection [17]. In

* This work was supported by the ANR-19-CE23-0028 MEERQAT project. This work was granted access to the HPC resources of IDRIS under the allocation 2021-AD011012846 made by GENCI.

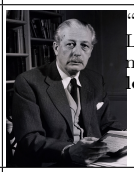
Visual Question (input)	Relevant visual passage in the Knowledge Base
 *"Which constituency did this man represent when he was Prime Minister?"*	 "Macmillan indeed lost Stockton in the landslide Labour victory of 1945, but returned to Parliament in the November 1945 by-election in **Bromley**."
 *"In which year did this ocean liner make her maiden voyage?"*	 "Queen Elizabeth 2, often referred to simply as QE2, is a floating hotel and retired ocean liner built for the Cunard Line which was operated by Cunard as both a transatlantic liner and a cruise ship from **1969** to 2008."

Fig. 1: Example of visual questions about named entities from the ViQuAE dataset along with relevant visual passages from its Knowledge Base [32].

KVQAE, both text and image modalities bring useful information that must be combined. Therefore, the task is more broadly related to *Multimodal Information Retrieval* (IR) and *Multimodal Fusion*.

There are two paradigms for multimodal IR and for multimodal learning more generally: early fusion (data- and feature-level) and late fusion (score- and decision-level) [30]. On the one hand, late fusion is more straightforward as both Natural Language Processing and Computer Vision techniques can be applied independently, but it neglects interaction between the modalities. On the other hand, the richness of early fusion often comes at the cost of increasing complexity and model parameters. This adds an extra challenge for KVQAE where the two existing datasets are either small (ViQuAE [32]) or generated automatically (KVQA [50]). To overcome this challenge, we propose to pre-train our fusion model using a multimodal Inverse Cloze Task (ICT). ICT has been introduced in [31] to pre-train a neural retriever for textual Question Answering (QA). It consists in considering a sentence as a pseudo-question and its context as a pseudo-relevant passage. It can be seen as a generalization of the skip-gram objective [41]. We extend it by considering images near texts in multimodal documents.

Our main contributions are: (i) Multimodal ICT, a new pre-training method that allows tackling small KVQAE datasets such as ViQuAE; (ii) a multimodal IR framework for KVQAE; (iii) experiments with different neural network architectures, including recently proposed multimodal BERTs.

2 Related Work

Dense Retrieval. Dense Retrieval is a rapidly evolving field, surveyed in [36,11], with new pre-training tasks, optimizing methods, and variants of the Transformer architecture emerging [47,23,15,14]. [31] were the first to outperform sparse bag-of-words representations such as BM25 with dense representations for QA. Their approach relies on three components: (i) pre-trained language models such as BERT [10], which allow to encode the semantic of a sentence in a

dense vector; (ii) a contrastive learning objective that optimizes the similarities between questions' and text passages' embeddings (see Section 3); (iii) an unsupervised training task, ICT (see Section 1). [27] criticize the latter for being computationally intensive[3] and argue that regular sentences are not good surrogates of questions. Instead, they propose DPR, which takes advantage of (i) the heuristic of whether the passage contains the answer to the question to deem it relevant; (ii) unsupervised IR methods such as BM25 to mine hard negatives examples, which proved to be the key of their method's success. We aim at taking advantage of both approaches by (i) pre-training our model on text QA datasets like DPR; (ii) incorporating multimodality into this hopefully-well-initialized model by adapting the ICT of [31] to multimodal documents.

Multimodal Fusion and Pre-Training. The success of BERT in NLP [10], which relies on the easily-parallelizable Transformer architecture [58], an unsupervised training objective, and a task-agnostic architecture, has concurrently inspired many works in the VQA and cross-modal retrieval fields [57,38,35,56,34,7]. These models are unified under a single framework in [5] and partly reviewed in [28]. All of these models rely on the Transformer architecture, often initialized with a pre-trained BERT, in order to fuse image and text. The training is weakly supervised, based upon image caption datasets such as COCO [37] or Conceptual Captions [51], and pre-trained object detectors like Faster R-CNN [48]. [22] show that these models learn nontrivial interactions between the modalities for VQA. Multimodal BERTs can be broadly categorized into *single-stream* and *multi-stream*. Single-stream models feed both text tokens' embeddings and image regions' embeddings to the same Transformer model, relying on the *self-attention* mechanism to fuse them. Instead, in the multi-stream architecture, text and image are first processed by two independent Transformers before using *cross-attention* to fuse the modalities. Both architectures have been shown to perform equally well in [5]. In this work, we use a single-stream model to take advantage of pre-training on text-only (on QA datasets). Also note that, while inspired by these work, we do not use the same training objectives or data, which are arguably unsuited for named entities' representations, as explained in the next section.

Multimodal Information Retrieval and KVQAE. Multimodal IR has largely been addressed using late fusion techniques (see [9] for a survey) but we are mostly interested in early fusion techniques in this work.

[9] review first attempts at early fusion. It was then systematically done by concatenating the features of both modalities in a single vector, with a focus on the feature weighting scheme. Concatenation is confronted with the curse of dimensionality as the resulting feature space equals the sum of the dimensions of each modality's features.

[44] and [39] concurrently proposed an approach quite similar to ours for Knowledge-based VQA. They adapt DPR [27] by replacing the question encoder

[3] [31] use a batch size of over 4K questions.

with LXMERT [57], which allows to fuse the question and image. However, unlike us, they keep the passage encoder based on text-only and use the same pre-training objectives as [57], namely Masked Language Modeling, Masked Region Modeling, and Image-Text Matching. We expect that these objectives are suited to learn representations of coarse-grained object categories but not named entities. In other words, they are suited for standard VQA but not KVQAE. For example, Masked Region Modeling relies on an object detector, which is not applicable to KVQAE. While both [44] and [39] experiment on OK-VQA [40], their results are inconsistent: [44] show that their model is competitive with a BM25 baseline that takes as input the question and the *human-written* caption of the image while the model of [39] is outperformed by BM25 with an *automatically-generated* caption. The discrepancies between these works can be explained because they use neither the same KB nor the same evaluation metrics. [19] also experiment with different multimodal BERTs but dispense passage-level annotation for an end-to-end training of the retriever and answer classifier[4].

Although they experiment with KVQA [50], we do not consider the work of [52,16,21] as their systems take a *human-written* caption as input, which makes the role of the image content unclear. [50] follow a late fusion approach at the decision-level. First, they detect and disambiguate the named entity mentions in the question. Then, they rely on a face recognition step as their dataset, KVQA, is restricted to questions about person named entities. Facts from both textually- and visually-detected entities are retrieved from Wikidata[5] and processed by a memory network [60]. In contrast, our work is in line with [32], who use unstructured text from Wikipedia as KB. Unlike [32], who follow a late fusion approach, searching the question and the image independently, we aim at a unified representation of the text and image, both on the visual question and KB sides.

3 Methods

In this section, we first formalize our KVQAE framework, then describe the models before diving into the three training stages: (i) DPR for textual Question Answering; (ii) Multimodal Inverse Cloze Task, our main contribution; (iii) Fine-tuning for KVQAE. Finally, we discuss the inference mechanism and implementation details.

3.1 Information Retrieval Framework

In our multimodal setting, both visual questions (from the dataset) and visual passages (from the KB) consist of a text-image pair (t, i), as in Figure 1. Our goal is to find the optimal model E to encode adequate representations $\mathbf{q} = E(t_q, i_q)$ and $\mathbf{p} = E(t_p, i_p)$ such that they are close if (t_p, i_p) is relevant for (t_q, i_q) (denoted

[4] Standard (Knowledge-based) VQA is often treated as a classification task.
[5] https://www.wikidata.org/

with the superscripts ($^+$) and ($^-$)). Search then boils down to retrieving the K closest visual passages to the visual question. When computing the similarity between two vectors, here with the dot product, the objective used throughout all the training stages (§3.3) is to minimize the following negative log-likelihood loss for all visual questions in the dataset [31,27]: $-\log \frac{\exp(\mathbf{q}\cdot\mathbf{p}^+)}{\exp(\mathbf{q}\cdot\mathbf{p}^+)+\sum_j \exp(\mathbf{q}\cdot\mathbf{p}_j^-)}$. This contrastive objective allows to efficiently utilize passages relevant to other questions in the batch as *in-batch negatives*, since computing the similarity between two vectors is rather inexpensive compared to the forward pass of the whole model. We present two different models E in the next section according to their fusion mechanism.

3.2 Models

All of our models take advantage of BERT[6] [10] and CLIP[7] [46] as building blocks to represent text and image, respectively. BERT is trained for masked language modeling and next sentence prediction on Wikipedia and BooksCorpus [62]. CLIP has been trained with a contrastive objective in a weakly-supervised manner over 400M image and caption pairs. It has demonstrated better generalization capacities than fully-supervised models and is efficient for KVQAE, as empirically demonstrated in [32]. We experiment with two different fusion techniques: ECA and ILF.

Early Cross-Attention fusion (ECA) is carried out by a single-stream Transformer model like the multimodal BERTs described above (e.g. UNITER [7]). However, instead of relying on a fixed object detector such as Faster R-CNN, we take advantage of CLIP, as motivated above. To enable early fusion, the visual embedding produced by CLIP is projected in the same space as the text using a linear layer with $\mathbf{W}_c \in \mathbb{R}^{c\times d}$ parameters trained from scratch: $\mathbf{e}_c = \text{CLIP}(i)\cdot\mathbf{W}_c$. The resulting embedding is then concatenated with the word embeddings in the sequence dimension, acting as a "visual token". Those embeddings are then fed to the Transformer model, where the attention mechanism should enable interaction between the modalities. The final embedding corresponds to the special `[CLS]` token: $\text{ECA}(t,i) = \text{BERT}([t;\mathbf{e}_c])_{\texttt{[CLS]}}$. The Transformer model is first initialized from BERT.

Intermediate Linear Fusion (ILF) introduces an additional $\mathbf{W}_t \in \mathbb{R}^{d\times d}$ parameters trained from scratch used to simply project the representation of the `[CLS]` token in the same space as the CLIP embedding before summing the two[8]: $\text{ILF}(t,i) = \text{BERT}(t)_{\texttt{[CLS]}}\cdot\mathbf{W}_t + \mathbf{e}_c$.

Because both ECA and ILF produce multimodal representations $\mathbf{q}$ and $\mathbf{p}$, ranking is done directly using their similarity $\mathbf{q}\cdot\mathbf{p}$. As baseline, we follow [32] and linearly combine text and image similarities after zero-mean and unit-variance

[6] Uncased "base" 12-layers version available at https://huggingface.co.

[7] With a ResNet-50 backbone [20].

[8] Note that this is equivalent to concatenating both before projecting like $[\text{BERT}(t)_{\texttt{[CLS]}};\text{CLIP}(i)]\cdot[\mathbf{W}_t;\mathbf{W}_c]$.

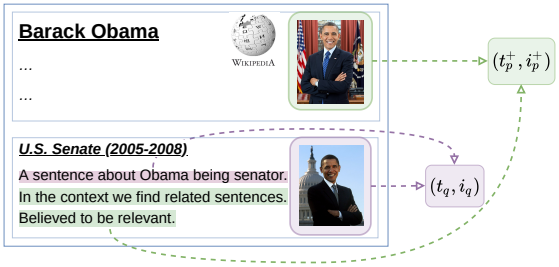


Fig. 2: Overview of Multimodal Inverse Cloze Task via Wikipedia/WIT.

normalization (omitted in the following equation):

$$\alpha \times \text{BERT}(t_q)_{\texttt{[CLS]}} \cdot \text{BERT}(t_p)_{\texttt{[CLS]}} + (1-\alpha) \times \cos(\text{CLIP}(i_q), \text{CLIP}(i_p)) \quad (1)$$

The interpolation hyperparameter α is optimized on the validation set using grid search to maximize Mean Reciprocal Rank. The left term (text similarity) is referred to as DPR in the rest of the paper.

3.3 Training stages

The models are trained sequentially in three stages:

Stage 1: DPR for textual Question Answering. Leaving visual representations aside, a DPR model is trained starting from the BERT initialization [27]. DPR consists of two BERT encoders: one for the question t_q and one for the text passage t_p. We use the model pre-trained by [32] on TriviaQA, filtered of all questions used in their dataset, ViQuAE. They use the KILT [43] version of TriviaQA and Wikipedia, which serves as KB in this stage. Each article is then split into disjoint passages of 100 words for text retrieval, while preserving sentence boundaries, and the title of the article is appended to the beginning of each passage. This yields 32M passages, that is ≈ 5.4 passages per article. Following [27], irrelevant passages (i.e. hard negatives) are mined using BM25 [49].

Stage 2: Multimodal Inverse Cloze Task. This is the main contribution of the paper. We propose to extend the ICT of [31] to multimodal documents. ICT consists in considering a sentence as a pseudo-question t_q and its context as a pseudo-relevant passage t_p^+. Note that the title of the article is appended to the beginning of each passage t_p (as in Stage 1). We extend it using the contextual images of Wikipedia paragraphs for the pseudo-question and the *infobox* image for the passage (see Figure 2). [31] empirically demonstrated that a key success of their approach was to leave the pseudo-question in the relevant passage in 10% of the training samples so that the model will learn to perform word matching, as lexical overlap is ultimately a very useful feature for retrieval. In our case,

however, we argue that it is neither necessary, as the model should be strongly initialized from Stage 1 training on TriviaQA, nor beneficial, as the model could then ignore the image modality. Question and passage encoders pre-trained in Stage 1 are used to initialize the visual question and visual passage encoders, respectively.

The process is eased thanks to the WIT dataset [54]. WIT consists of millions of images with associated text from Wikipedia and Wikimedia Commons in 108 different languages. We are, however, only interested in English for this work. While [54] have multiple strategies to find text related to a given Wikipedia image, such as its Commons' caption, we use only the contextual paragraph as text source in order to mimic the downstream KVQAE setting. The resulting English subset of WIT yields 400K *infobox* images/articles that correspond to 1.2M paragraphs/images. Those 1.2M paragraphs consist of 13.6M sentences, i.e. potential pseudo-questions, which are 26 words long on average. Therefore, to stick as close as possible to stages 1 and 3, where passages are up to 100 *words* long, passages consist of *four* sentences. This slightly differs from [31] who consider passages of up to 288 *wordpieces*, *prior* to the pseudo-question masking.

Because both ViQuAE and WIT images are taken from Wikimedia Commons[9], we can estimate from the image URLs that 14% of ViQuAE images overlap with WIT. This might lead to a bias that we analyze in Section 4.1.

Inspired by [2], to prevent *catastrophic forgetting* and enforce a *modality-invariant* representation of the entities, the last l layers of BERT are frozen during this stage. In this way, we tune only the first, modality-specific layers of ECA, the intuition being to "replace" the text-named entities learned during Stage 1 with the "visual" entities present in the images. ILF fully freezes BERT during this stage, relying only on the $\mathbf{W}_t$ parameters to tune the text representation. Furthermore, CLIP is systematically frozen throughout all stages.

We do not have a straightforward way of mining irrelevant visual passages in this stage. In early experiments, we tried to synthesize them by permuting images in the batch: $(t_p^+, i_p^+) \leftarrow (t_p^+, i_p^-)$, but it did not improve the results.

After filtering corrupted images or images with inappropriate image formats (e.g. .svg) and paragraphs with a single sentence, we end up with 975K paragraphs/images. We refer to it as WIT in the rest of the paper. It is split into train (878K), validation (48K, to tune hyperparameters), and test (48K, as a sanity check) subsets such that there is no overlap between articles.

Stage 3: Knowledge-based Visual Question Answering about Named Entities. This stage consists in fine-tuning the model on a downstream KVQAE dataset, which provides visual questions (t_q, i_q) and relevant visual passages (t_p^+, i_p^+). Following [2], all layers of the model are tuned during this stage.

A subtlety of this stage is the selection of irrelevant visual passages (t_p^-, i_p^-). As mentioned in Section 2, it was shown to be essential to DPR [27], and it is more generally important for contrastive learning [26]. In [32], irrelevant passages

[9] https://commons.wikimedia.org/

are mined with BM25 to train DPR. However, we suppose that this is suboptimal for ECA and ILF as BM25 will only mine textually-plausible passages but not visually-plausible ones. Therefore, we use the system provided by [32] to mine irrelevant passages. It is a late-fusion of DPR, ArcFace [8], CLIP, and ImageNet-ResNet [20]. This leads to different training setups between DPR (used as a baseline) and our models. However, we have experimented both for DPR and found no significant differences[10].

We use the same KB as [32], which is based upon KILT's Wikipedia and Wikidata images of the corresponding entities. It consists of 1.5M articles (thus images/entities) split into 12M passages of at most 100 words as in Stage 1.

Visual questions in ViQuAE are split into train (1,190), validation (1,250), and test (1,257) without overlap between images' URLs [32]. We do not experiment with KVQA [50] for the following reasons: (i) it is generated automatically from Wikidata so our text-based KB has a poor coverage of the answers; (ii) it comprises yes/no questions for which passage relevance cannot be assessed automatically.

3.4 Inference

For efficient retrieval, every passage in the KB is embedded along with its corresponding image by the visual passage encoder beforehand. Given a question grounded in an image, both are embedded by the visual question encoder. Search is then carried out with maximum inner product search using Faiss [25].

3.5 Implementation Details

Our code is built upon PyTorch [42], Hugging Face's `transformers` [61] and `datasets` [33] (itself wrapping Faiss). It is freely available along with the data and trained models[11].

To train ECA, we use the same hyperparameters as [32] for DPR, themselves based upon [27]. In particular, we use a learning rate of 2×10^{-5} along with the Adam optimizer [29]. It is scheduled linearly with 100 and 4 warm-up steps for stages 2 and 3, respectively. However, for ILF, we found, based on the validation set, that it converged faster with a learning rate of 2×10^{-3} and a constant scheduler during Stage 2. We believe this is because ILF fully freezes BERT in Stage 2, so it does not require careful scheduling or a small learning rate. Dropout [55] is applied in BERT and after projecting embedding with $\mathbf{W}_c$ and $\mathbf{W}_t$ with a probability of 0.1 (as in the standard BERT configuration). Likewise, layer normalization [3] is applied in BERT and after summing the two embeddings in ILF. Gradients' norms are clipped at 2.

Models in stages 2 and 3 are trained with a batch size of 512 and 298 visual questions, respectively. The success of contrastive learning partly relies on a large number of *in-batch negatives* and, therefore, a large batch size [45]. We found

[10] Evaluation methods are detailed in Section 4

[11] https://github.com/PaulLerner/ViQuAE

Table 1: IR evaluation on ViQuAE. l: Number of frozen layers during Multimodal ICT. Superscripts denote significant differences in Fisher's randomization test with $p \leq 0.01$. Hits@1 is omitted as it is equivalent to P@1.

#	Model	Multimodal ICT	MRR@100	P@1	P@20	Hits@20
a	DPR	NA	32.8	22.8	16.4	61.2
b	DPR + CLIP	NA	34.5^{a}	24.8^{a}	15.8	61.8
c	ECA	✗	34.6	25.9^{a}	17.2^{ab}	61.6
d	ECA ($l = 6$)	✓	$\mathbf{37.8}^{abce}$	26.7^{a}	$\mathbf{19.5}^{abce}$	$\mathbf{67.6}^{abce}$
e	ECA ($l = 0$)	✓	35.1	24.7	17.6^{b}	63.7
f	ILF ($l = 12$)	✓	37.3^{a}	$\mathbf{26.8}^{a}$	19.1^{abce}	66.9^{abc}

that gradient checkpointing [6] enables the use of much larger batch sizes for ECA[12]. Instead of [32] who use *four* NVIDIA V100 GPUs with 32GB of RAM each for a total batch size of 128 questions, we are able to fit a batch of 298 questions (as stated above) in a *single* V100 GPU. Stage 2 takes most of the compute budget, with most models converging after $\approx$ 8K steps, which takes around three days[13]. Checkpoint selection is made based on the validation *in-batch* Mean Reciprocal Rank, for all stages. In-batch means that only the other visual passages in the batch are ranked and that each visual question is paired with only one relevant visual passage (as during training).

4 Results

The retrieval models are evaluated in two different ways: (i) by computing standard IR metrics on visual passage retrieval; (ii) by feeding retrieved visual passages to a reader module that is tasked with extracting the concise answer to the question, thus achieving KVQAE. Put differently, either evaluate whether the system is able to retrieve a *relevant passage* for the question or whether it is able to *answer* the question. We find both metrics to correlate. Ablation studies are carried out with IR metrics.

ViQuAE is based upon TriviaQA, so it is only distantly supervised: the answer is considered correct if it string-matches the ground truth and, likewise, a passage is deemed relevant if it contains the ground truth[14]. Moreover, Wikipedia aliases of the ground truth are considered to be valid answers.

4.1 Information Retrieval

Because of the setting of ViQuAE, it is impossible to get complete coverage of relevant passages. Therefore we do not use any metric based on recall (e.g. R-

[12] It is not necessary for ILF that fully freezes BERT.

[13] Jean Zay GPUs consume 0.482kW (or 0.259kW after heat recovery) in France, which has an average grid emission factor of 0.0569 kgCO2e/kWh according to https://bilans-ges.ademe.fr/en.

[14] After standard preprocessing (lowercasing, stripping articles, and punctuation).

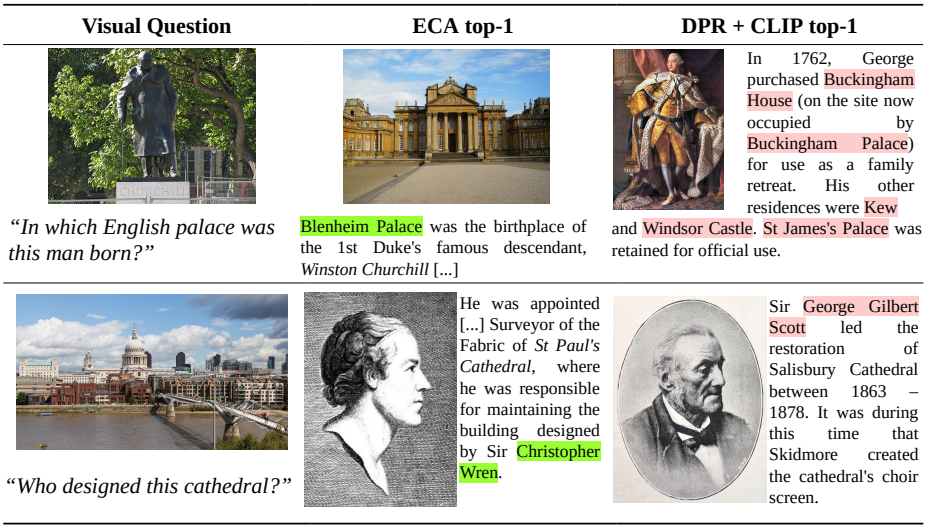


Fig. 3: Qualitative examples where ECA ($l = 6$) finds a relevant visual passage in top-1 but late fusion falls behind.

Precision, mAP, etc.). Instead, we evaluate the models with Precision@K (P@K), Mean Reciprocal Rank (MRR), and Hits@K. Hits@K is the proportion of questions for which IR retrieves *at least one* relevant passage in top-K. Statistical significance tests are conducted using Fisher's randomization test [12,53]. Metrics and statistical tests are computed with `ranx` [4] and are reported in Table 1.

The best models pre-trained with Multimodal ICT (d and f) outperform the text-only (a) and late-fusion (b) baselines on all metrics. Some qualitative examples are shown in Figure 3. In the first row, we can see evidence of cross-modal interaction between the image depicting Winston Churchill and the passage that mentions him (while being illustrated by a totally different image). In contrast, the late fusion baseline exhibits textual bias by returning a passage that mentions several English palaces (highlighted in red). The same observation can be made for the second row, where St Paul's Cathedral is only mentioned in the relevant passage but not depicted in the contextual image. Cross-modal interactions prove useful in this case because of the heterogeneity of visual depictions: Winston Churchill is depicted by a statue in the visual question but by a photograph in the KB.

We can see that Multimodal ICT is essential to ECA (c vs. d). Without it, it performs on par with late fusion. We believe this is because of overfitting on the small training set of ViQuAE. However, we find that fine-tuning on ViQuAE is also essential to ECA, which exhibits catastrophic forgetting because of the sequential learning setup: indeed, after Stage 2, it falls behind DPR. We see that the freezing technique of [2] helps to prevent catastrophic forgetting to some extent (d vs. e). It is also visible in the upstream WIT pre-training where

Table 2: Reading Comprehension evaluation on ViQuAE, averaged over 5 runs of the *reader*. l: Number of frozen layers during Multimodal ICT.

#	IR Model	Multimodal ICT	Exact Match	F1
a	DPR	NA	16.9 ± 0.4	20.1 ± 0.5
b	DPR + CLIP	NA	19.0 ± 0.4	22.3 ± 0.4
c	ECA	✗	17.7 ± 0.6	21.2 ± 0.8
d	ECA ($l = 6$)	✓	20.6 ± 0.3	24.4 ± 0.2
e	ECA ($l = 0$)	✓	20.8 ± 0.8	24.3 ± 0.9
f	ILF ($l = 12$)	✓	**21.3** ± 0.6	**25.4** ± 0.3

ECA achieves 91.6 and 92.9 in-batch MRR on WIT's test set with $l = 6$ and $l = 0$, respectively: fitting WIT better leads to further forgetting.

Unlike what is suggested by related work (§2), we find that the linear fusion model performs on par with the more early, cross-attention based, fusion model (f vs. d). This suggests that the improvement over the late fusion baseline indeed comes from the Multimodal ICT pre-training, which is not very sensitive to the model's architecture. Moreover, the architecture of ILF allows to fully freeze BERT during Stage 2, which circumvents catastrophic forgetting[15]. We leave other training strategies (e.g. multi-tasking, using adapters [24]) for future work.

Nothing suggests that ECA is better on the 14% of ViQuAE images that overlap with WIT. ECA is better on the out-of-WIT subset (38.0 vs. 36.5 MRR), but it is the other way around for DPR and late fusion.

4.2 Reading Comprehension

To extract the answers from the retrieved passages, we keep the same model as [32]. It uses the Multi-passage BERT architecture [59] and is thus based on text only because *once the relevant passage has been retrieved*, the question may be answered without looking at the image. To limit the variations due to training and the number of experiments, we use the model trained by [32] off-the-shelf and simply change its input passages. It takes the top-24 passages as input. The model was first trained on TriviaQA (filtered of all questions used in ViQuAE), then fine-tuned on ViQuAE, much like stages 1 and 3. The authors provide *five* different versions of the model that correspond to different random seeds.

We use Exact Match and F1-score (at the bag-of-words level) to evaluate the extracted answers. In Table 2 we can verify that more relevant passages indeed lead to better downstream answers. The only difference with the IR evaluation is the role of the freezing technique of [2] (d vs. e), which is less clear here.

[15] ILF only achieves 87.1 in-batch MRR on WIT's test set because of the freezing.

5 Generic vs. Specialized Image Representations

Numbers reported in the previous section are actually on par with the best results of [32]. This is because the latter is based on ArcFace and ImageNet-ResNet, in addition to DPR and CLIP. In particular, [32] have a heuristic for taking advantage of the face representations provided by ArcFace: they use ArcFace if faces are detected and a combination of CLIP and ImageNet-ResNet otherwise. They show that this method improves retrieval precision for questions about persons (for which face representations are relevant). However, this approach is not scalable for two reasons: (i) there are near 1,000 different entity types in ViQuAE (according to Wikidata's ontology), and not all can benefit from specialized representations; (ii) combining several representations (e.g. CLIP and ImageNet-ResNet) for the same entity type is computationally expensive and quickly saturates. To provide a comparable system to the late fusion of [32], we have tried integrating ArcFace and ImageNet-ResNet in ECA. However, we have failed to outperform the CLIP-only version of ECA. Intuitively, we think that ECA dilutes the specialized representations of ArcFace and is unable to preserve them throughout all twelve layers of BERT. Therefore, in this setting, ECA is overall on par with late fusion (37.7 vs. 37.9 MRR, not significant) but better on questions about non-persons (39.3 vs. 35.7 MRR), which again suggests that it is unable to exploit ArcFace's representations.

6 Conclusion and Perspectives

We have presented a new pre-training method, Multimodal Inverse Cloze Task, for Knowledge-based Visual Question Answering about Named Entities. Multimodal ICT leverages contextual images in multimodal documents to generate pseudo-visual questions. It enables the use of more complex multimodal fusion models than previously proposed late fusion methods. Consequently, our method improves retrieval accuracy over the latter by 10% relative-MRR, leading to a 9% relative-F1 improvement in downstream reading comprehension (i.e. answer extraction), on the recently introduced ViQuAE dataset. We believe it is thanks to cross-modal interactions, which are prohibited by late fusion. More precisely, we qualitatively observed that these interactions occurred between the image of the visual question and the text of the KB, which counteracts the heterogeneity of visual depictions.

We have experimented our pre-training method with two different neural networks architectures: (i) ECA, which follows recently proposed Multimodal BERTs by fusing modalities Early via Cross-Attention; (ii) ILF, a more standard model that fuses modalities through a linear projection. We found that both perform equally well, unlike in standard VQA and cross-modal retrieval. We argue that it might be because of their difference in training settings, which leads ECA to catastrophic forgetting. However, further investigations are required.

While aiming for generic multimodal representations of named entities, we found that integrating specialized representations in our models, such as ArcFace

for faces, was not beneficial. We hypothesize that the studied architectures may be inappropriate but we leave this issue for future studies.

For future work, we think that generalizing Multimodal ICT for re-ranking (processing (t_q, i_q) and (t_p, i_p) simultaneously) and reading comprehension (generating or extracting the answer from (t_p, i_p)) is an exciting research lead. Indeed, there is evidence that sharing the same model for IR and reading comprehension, or IR and re-ranking, is beneficial for textual QA [13] and cross-modal retrieval [18], respectively: two tasks that closely relate to KVQAE.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual Question Answering. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 2425–2433. IEEE, Santiago, Chile (Dec 2015). https://doi.org/10.1109/ICCV.2015.279, http://ieeexplore.ieee.org/document/7410636/
2. Aytar, Y., Castrejon, L., Vondrick, C., Pirsiavash, H., Torralba, A.: Cross-modal scene networks. IEEE transactions on pattern analysis and machine intelligence **40**(10), 2303–2314 (2017), publisher: IEEE
3. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer Normalization. arXiv:1607.06450 [cs, stat] (Jul 2016), http://arxiv.org/abs/1607.06450, arXiv: 1607.06450
4. Bassani, E.: ranx: A Blazing-Fast Python Library for Ranking Evaluation and Comparison. In: Hagen, M., Verberne, S., Macdonald, C., Seifert, C., Balog, K., Nørvåg, K., Setty, V. (eds.) Advances in Information Retrieval. pp. 259–264. Lecture Notes in Computer Science, Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-030-99739-7_30
5. Bugliarello, E., Cotterell, R., Okazaki, N., Elliott, D.: Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs. Transactions of the Association for Computational Linguistics **9**, 978–994 (Sep 2021). https://doi.org/10.1162/tacl_a_00408, https://doi.org/10.1162/tacl_a_00408
6. Chen, T., Xu, B., Zhang, C., Guestrin, C.: Training Deep Nets with Sublinear Memory Cost (Apr 2016). https://doi.org/10.48550/arXiv.1604.06174, http://arxiv.org/abs/1604.06174, number: arXiv:1604.06174 arXiv:1604.06174 [cs]
7. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: European Conference on Computer Vision. pp. 104–120. Springer (2020)
8. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019), https://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html
9. Depeursinge, A., Müller, H.: Fusion Techniques for Combining Textual and Visual Information Retrieval. In: Müller, H., Clough, P., Deselaers, T., Caputo, B. (eds.) ImageCLEF: Experimental Evaluation in Visual Information Retrieval, pp. 95–114. The Information Retrieval Series, Springer, Berlin, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15181-1_6, https://doi.org/10.1007/978-3-642-15181-1_6

10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
11. Fan, Y., Xie, X., Cai, Y., Chen, J., Ma, X., Li, X., Zhang, R., Guo, J., Liu, Y.: Pre-training Methods in Information Retrieval. arXiv:2111.13853 [cs] (Nov 2021), http://arxiv.org/abs/2111.13853, arXiv: 2111.13853
12. Fisher, R.A.: The design of experiments. The design of experiments. (2nd Ed) (1937), https://www.cabdirect.org/cabdirect/abstract/19371601600, publisher: Oliver & Boyd, Edinburgh & London.
13. Fun, H., Gandhi, S., Ravi, S.: Efficient Retrieval Optimized Multi-task Learning. arXiv:2104.10129 [cs] (Apr 2021), http://arxiv.org/abs/2104.10129, arXiv: 2104.10129
14. Gao, L., Callan, J.: Condenser: a Pre-training Architecture for Dense Retrieval. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 981–993. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021), https://aclanthology.org/2021.emnlp-main.75
15. Gao, L., Callan, J.: Unsupervised corpus aware language model pre-training for dense passage retrieval. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2843–2853. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.203, https://aclanthology.org/2022.acl-long.203
16. Garcia-Olano, D., Onoe, Y., Ghosh, J.: Improving and diagnosing knowledge-based visual question answering via entity enhanced knowledge injection. In: Companion Proceedings of the Web Conference 2022. p. 705–715. WWW '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3487553.3524648, https://doi.org/10.1145/3487553.3524648
17. Gardères, F., Ziaeefard, M.: ConceptBert: Concept-Aware Representation for Visual Question Answering. Findings of the Association for Computational Linguistics: EMNLP 2020 p. 10 (2020), https://aclanthology.org/2020.findings-emnlp.44/
18. Geigle, G., Pfeiffer, J., Reimers, N., Vulić, I., Gurevych, I.: Retrieve Fast, Rerank Smart: Cooperative and Joint Approaches for Improved Cross-Modal Retrieval. Transactions of the Association for Computational Linguistics **10**, 503–521 (05 2022). https://doi.org/10.1162/tacl_a_00473, https://doi.org/10.1162/tacl_a_00473
19. Guo, Y., Nie, L., Wong, Y., Liu, Y., Cheng, Z., Kankanhalli, M.: A Unified End-to-End Retriever-Reader Framework for Knowledge-based VQA (Jun 2022), http://arxiv.org/abs/2206.14989, number: arXiv:2206.14989 arXiv:2206.14989 [cs]
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016), https://openaccess.thecvf.com/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf
21. Heo, Y.J., Kim, E.S., Choi, W.S., Zhang, B.T.: Hypergraph Transformer: Weakly-supervised multi-hop reasoning for knowledge-based visual question answering. In: Proceedings of the 60th Annual Meeting of the Association for Computational

Linguistics (Volume 1: Long Papers). pp. 373–390. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.29, https://aclanthology.org/2022.acl-long.29
22. Hessel, J., Lee, L.: Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 861–877. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.62, https://aclanthology.org/2020.emnlp-main.62
23. Hofstätter, S., Lin, S.C., Yang, J.H., Lin, J., Hanbury, A.: Efficiently teaching an effective dense retriever with balanced topic aware sampling. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 113–122. SIGIR '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3404835.3462891, https://doi.org/10.1145/3404835.3462891
24. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., Laroussilhe, Q.D., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-Efficient Transfer Learning for NLP. In: Proceedings of the 36th International Conference on Machine Learning. pp. 2790–2799. PMLR (May 2019), https://proceedings.mlr.press/v97/houlsby19a.html, iSSN: 2640-3498
25. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Transactions on Big Data **7**(3), 535–547 (2019). https://doi.org/10.1109/TBDATA.2019.2921572
26. Kalantidis, Y., Sariyildiz, M.B., Pion, N., Weinzaepfel, P., Larlus, D.: Hard Negative Mixing for Contrastive Learning. In: Advances in Neural Information Processing Systems. vol. 33, pp. 21798–21809. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper/2020/hash/f7cade80b7cc92b991cf4d2806d6bd78-Abstract.html
27. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.t.: Dense passage retrieval for open-domain question answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6769–6781. Association for Computational Linguistics, Online (Nov 2020), https://www.aclweb.org/anthology/2020.emnlp-main.550
28. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. ACM Comput. Surv. (dec 2021). https://doi.org/10.1145/3505244, https://doi.org/10.1145/3505244, just Accepted
29. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: Proceedings of the 3rd International Conference for Learning Representations (Jan 2015), http://arxiv.org/abs/1412.6980
30. Kludas, J., Bruno, E., Marchand-Maillet, S.: Information fusion in multimedia information retrieval. In: International Workshop on Adaptive Multimedia Retrieval. pp. 147–159. Springer (2007)
31. Lee, K., Chang, M.W., Toutanova, K.: Latent Retrieval for Weakly Supervised Open Domain Question Answering. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 6086–6096. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1612, https://aclanthology.org/P19-1612
32. Lerner, P., Ferret, O., Guinaudeau, C., Le Borgne, H., Besançon, R., Moreno, J.G., Lovón Melgarejo, J.: ViQuAE, a dataset for knowledge-based visual question answering about named entities. In: Proceedings of The 45th Interna-

tional ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR'22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3477495.3531753, https://hal.archives-ouvertes.fr/hal-03650618

33. Lhoest, Q., Villanova del Moral, A., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., Davison, J., Šaško, M., Chhablani, G., Malik, B., Brandeis, S., Le Scao, T., Sanh, V., Xu, C., Patry, N., McMillan-Major, A., Schmid, P., Gugger, S., Delangue, C., Matussière, T., Debut, L., Bekman, S., Cistac, P., Goehringer, T., Mustar, V., Lagunas, F., Rush, A., Wolf, T.: Datasets: A Community Library for Natural Language Processing. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 175–184. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021), https://aclanthology.org/2021.emnlp-demo.21
34. Li, G., Duan, N., Fang, Y., Gong, M., Jiang, D.: Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training. Proceedings of the AAAI Conference on Artificial Intelligence **34**(07), 11336–11344 (Apr 2020). https://doi.org/10.1609/aaai.v34i07.6795, https://ojs.aaai.org/index.php/AAAI/article/view/6795, number: 07
35. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: VisualBERT: A Simple and Performant Baseline for Vision and Language (Jan 2019), https://openreview.net/forum?id=OiixWIb7JpAg
36. Lin, J., Nogueira, R., Yates, A.: Pretrained transformers for text ranking: Bert and beyond. Synthesis Lectures on Human Language Technologies **14**(4), 1–325 (2021). https://doi.org/10.2200/S01123ED1V01Y202108HLT053
37. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 740–755. Lecture Notes in Computer Science, Springer International Publishing, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
38. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. Advances in Neural Information Processing Systems **32**, 13–23 (2019), https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html
39. Luo, M., Zeng, Y., Banerjee, P., Baral, C.: Weakly-supervised visual-retriever-reader for knowledge-based question answering. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 6417–6431. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.emnlp-main.517, https://aclanthology.org/2021.emnlp-main.517
40. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: OK-VQA: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3195–3204 (2019), https://ieeexplore.ieee.org/document/8953725/
41. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems **26** (2013), https://papers.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html
42. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z.,

Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems **32** (2019), https://papers.nips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

43. Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., Plachouras, V., Rocktäschel, T., Riedel, S.: KILT: a benchmark for knowledge intensive language tasks. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2523–2544. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.naacl-main.200, https://aclanthology.org/2021.naacl-main.200
44. Qu, C., Zamani, H., Yang, L., Croft, W.B., Learned-Miller, E.: Passage retrieval for outside-knowledge visual question answering. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1753–1757. SIGIR '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3404835.3462987, https://doi.org/10.1145/3404835.3462987
45. Qu, Y., Ding, Y., Liu, J., Liu, K., Ren, R., Zhao, W.X., Dong, D., Wu, H., Wang, H.: RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5835–5847. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.naacl-main.466, https://aclanthology.org/2021.naacl-main.466
46. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
47. Ram, O., Shachaf, G., Levy, O., Berant, J., Globerson, A.: Learning to Retrieve Passages without Supervision. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2687–2700. Association for Computational Linguistics, Seattle, United States (Jul 2022), https://aclanthology.org/2022.naacl-main.193
48. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Advances in Neural Information Processing Systems **28**, 91–99 (2015), https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html
49. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M.: Okapi at TREC-3. In: Harman, D.K. (ed.) Third Text REtrieval Conference (TREC-3). NIST Special Publication, vol. 500-225, pp. 109–126. National Institute of Standards and Technology (NIST) (1995), https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.32.9922&rep=rep1&type=pdf
50. Shah, S., Mishra, A., Yadati, N., Talukdar, P.P.: KVQA: Knowledge-Aware Visual Question Answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8876–8884 (2019), https://144.208.67.177/ojs/index.php/AAAI/article/view/4915
51. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceed-

ings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565 (2018)
52. Shevchenko, V., Teney, D., Dick, A., van den Hengel, A.: 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). pp. 1–18. Association for Computational Linguistics, Kyiv, Ukraine (Apr 2021), https://aclanthology.org/2021.lantern-1.1
53. Smucker, M.D., Allan, J., Carterette, B.: A comparison of statistical significance tests for information retrieval evaluation. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. pp. 623–632. CIKM '07, Association for Computing Machinery, New York, NY, USA (Nov 2007). https://doi.org/10.1145/1321440.1321528, https://doi.org/10.1145/1321440.1321528
54. Srinivasan, K., Raman, K., Chen, J., Bendersky, M., Najork, M.: 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. p. 2443–2449. SIGIR '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3404835.3463257, https://doi.org/10.1145/3404835.3463257
55. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014), publisher: JMLR. org
56. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In: Proceedings of ICLR 2020 (Feb 2020), http://arxiv.org/abs/1908.08530, arXiv: 1908.08530
57. Tan, H., Bansal, M.: 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). pp. 5100–5111. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1514, https://www.aclweb.org/anthology/D19-1514
58. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
59. Wang, Z., Li, L., Li, Q., Zeng, D.: Multimodal Data Enhanced Representation Learning for Knowledge Graphs. In: 2019 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (Jul 2019). https://doi.org/10.1109/IJCNN.2019.8852079, iSSN: 2161-4407
60. Weston, J., Chopra, S., Bordes, A.: Memory networks (2015)
61. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771 [cs] (Jul 2020), http://arxiv.org/abs/1910.03771
62. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Decem-

ber 2015), https://www.cv-foundation.org/openaccess/content_iccv_2015/html/Zhu_Aligning_Books_and_ICCV_2015_paper.html