Leveraging Visual Question Answering for Image-Caption Ranking

Xiao Lin  Devi Parikh

Virginia Tech

Abstract. Visual Question Answering (VQA) is the task of taking as input an image and a free-form natural language question about the image, and producing an accurate answer. In this work we view VQA as a “feature extraction” module to extract image and caption representations. We employ these representations for the task of image-caption ranking. Each feature dimension captures (imagines) whether a fact (question-answer pair) could plausibly be true for the image and caption. This allows the model to interpret images and captions from a wide variety of perspectives. We propose score-level and representation-level fusion models to incorporate VQA knowledge in an existing state-of-the-art VQA-agnostic image-caption ranking model. We find that incorporating and reasoning about consistency between images and captions significantly improves performance. Concretely, our model improves state-of-the-art on caption retrieval by 7.1% and on image retrieval by 4.4% on the MSCOCO dataset.

1 Introduction

Visual Question Answering (VQA) is an “AI-complete” problem that requires knowledge from multiple disciplines such as computer vision, natural language processing and knowledge base reasoning. A VQA system takes as input an image and a free-form open-ended question about the image and outputs the natural language answer to the question. A VQA system needs to not only recognize objects and scenes but also reason beyond low-level recognition about aspects such as intention, future, physics, material and commonsense knowledge. For example (Q: Who is the person in charge in this picture? A: Chef) reveals the most important person and occupation in the image. Moreover, answers to multiple questions about the same image can be correlated and may reveal more complex interactions. For example (Q: What is this person riding? A: Motorcycle) and (Q: What is the man wearing on his head? A: Helmet) might reveal correlations observable in the visual world due to safety regulations.

Today’s VQA models, while far from perfect, may already be picking up on these semantic correlations of the world. If so, they may serve as an implicit knowledge resource to help other tasks. Just like we do not need to fully understand the theory behind an equation to use it, can we already use VQA knowledge captured by existing VQA models to improve other tasks?

In this work we study the problem of using VQA knowledge to improve image-caption ranking. Consider the image and its caption in Figure 1. Aligning them not only requires recognizing the batter and that it is a baseball game (mentioned in the caption), but also realizing that a batter up at the plate would imply that a player is holding a bat,
Fig. 1. Aligning images and captions requires high-level reasoning e.g. “a batter up at the plate” would imply that a player is holding a bat, posing to hit the baseball and there might be another player nearby waiting to catch the ball. There is rich knowledge in Visual Question Answering (VQA) corpora containing human-provided answers to a variety of questions one could ask about images. We propose to leverage knowledge in VQA by using VQA models learned on images and captions as “feature extraction” modules for image-caption ranking.

posing to hit the baseball and there might be another player nearby waiting to catch the ball (seen in the image). Image captions tend to be generic. As a result, image captioning corpora may not capture sufficient details for models to infer this knowledge.

Fortunately VQA models try to explicitly learn such knowledge from a corpus of images, each with associated questions and answers. Questions about images tend to be much more specific and detailed than captions. The VQA dataset of [1] in particular has a collection of free-form open-ended questions and answers provided by humans. These images also have associated captions [2].

We propose to leverage VQA knowledge captured by such corpora for image-caption ranking by using VQA models learned on images and captions as “feature extraction” schemes to represent images and captions. Given an image and a caption, we choose a set of free-form open-ended questions and use VQA models learned on images and captions to assess probabilities of their answers. We use these probabilities as image and caption features respectively. In other words, we embed images and captions into the space of VQA questions and answers using VQA models. Such VQA-grounded representations interpret images and captions from a variety of different perspectives and imagine beyond low-level recognition to better understand images and captions.

We propose two approaches that incorporate these VQA-grounded representations into an existing state-of-the-art VQA-agnostic image-caption ranking model [3]: fusing their predictions and fusing their representations. We show that such VQA-aware models significantly outperform the VQA-agnostic model and set state-of-the-art per-

---

1 To the best of our knowledge, [3] has the state-of-the-art caption retrieval performance on MSCOCO [2]. [4] has the state-of-the-art image retrieval performance on MSCOCO.
formance on MSCOCO image-caption ranking. Specifically, we improve caption retrieval by 7.1% and image retrieval by 4.4%.

This paper is organized as follows: Section 2 introduces related works. We first introduce VQA and image-caption ranking tasks as our building blocks in Section 3 then detail our VQA-based image-caption ranking models in Section 4. Experiments and results are reported in Section 5. We conclude in Section 6.

2 Related Work

Visual Question Answering. Visual Question Answering (VQA) \cite{1} is the task of taking an image and a free-form open-ended question about the image and automatically predicting the natural language answer to the question. VQA may require fine-grained recognition, object detection, activity recognition, multi-modal and commonsense knowledge. Large datasets \cite{5,6,7,8,1} have been made available to cover the diversity of knowledge required for VQA. Most notably the VQA dataset \cite{1} contains 614,163 questions and ground truth answers on 204,721 images of the MSCOCO \cite{2} dataset.

Recent VQA models \cite{9,6,8,1,4} explore state-of-the-art deep learning techniques combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). \cite{1} also explores a slight variant of VQA that answers a question about the image by reading a caption describing the image instead of looking at the image itself. We call this variant VQA-Caption. Interestingly, VQA models using captions instead of images perform better than models that use images, perhaps because current vision systems are unable to understand images accurately enough, but AI systems are capable of understanding single-sentence captions reasonably well.

VQA is still a challenging task in its early stages. In this work we propose to use both VQA and VQA-Caption models as implicit knowledge resources. We show that current VQA models, while far from perfect, can already be used to improve other multi-modal AI tasks; specifically image-caption ranking.

Semantic mid-level visual representations. Previous works have explored the use of attributes \cite{10,11,12}, parts \cite{13,14}, poselets \cite{15,16}, objects \cite{17}, actions \cite{18} and contextual information \cite{19,20,21} as semantic mid-level representations for visual recognition. Benefits of using such semantic mid-level visual representations include improving fine-grained visual recognition, learning models of visual concepts without example images (zero-shot learning \cite{22,23}) and improving human-machine communication where a user can explain the target concept during image search \cite{24,25}, or give a classifier an explanation of labels \cite{26,27}. More recent works also explore using word embeddings \cite{28} and free-form text \cite{29} as representations that allow zero-shot learning of new object categories. \cite{30} proposes scene graphs for image retrieval. \cite{31} proposes using abstract scenes as an intermediate representation for zero-shot action recognition. Closest to our work is the use of objects, actions, scenes \cite{32}, attributes and object interactions \cite{33} for generating and ranking image captions. In this work we propose to use free-form open-ended questions and answers as mid-level representations and we show that they provide rich interpretations of images and captions.

Commonsense knowledge for visual reasoning. Recently there has been a surge of interest in visual reasoning tasks that require high-level reasoning such as physical reason-
Images and captions. Recent works \cite{45,46,3,47,48,49} have made significant progress on automatic image caption generation and ranking by applying deep learning techniques for image recognition \cite{50,51,52} and sequence modeling \cite{53,54} on large datasets \cite{55,2}. Algorithms can now often generate accurate, human-like natural-language captions for images. However, evaluating the quality of such automatically generated open-ended image captions is still an open research problem \cite{56,57}.

On the other hand, ranking images given captions and ranking captions given images require a similar level of image and language understanding, but are amenable to automatic evaluation metrics. Recent works on image-caption ranking mainly focus on improving model architectures. \cite{3,48} study different architectures for projecting CNN image representations and RNN caption representations into a common multi-modal space. \cite{49} uses multi-modal CNNs for image-caption ranking. \cite{45} aligns image and caption fragments using CNNs and RNNs. Our work takes an orthogonal approach to previous works. We propose to leverage knowledge in VQA corpora containing questions about images and associated answers for image-caption ranking. Our proposed VQA-based image and caption representations provide complementary information to those learned using previous approaches on a large image-caption ranking dataset.

3 Building Blocks: Image-Caption Ranking and VQA

In this section we present image-caption ranking and VQA modules that we build on top of.

3.1 Image-caption ranking

The image-caption ranking task is to retrieve relevant images given a query caption, and relevant captions given a query image.

During training we are given image-caption pairs \((I, C)\) that each corresponds to an image \(I\) and its caption \(C\). For each pair we sample \(K - 1\) other images in addition to \(I\) so the image retrieval task becomes retrieving \(I\) from \(K\) images \(I_i, i = 1, 2 \ldots K\) given caption \(C\). We also sample \(K - 1\) random captions in addition to \(C\) so the caption retrieval task becomes retrieving \(C\) from \(K\) captions \(C_i, i = 1, 2 \ldots K\) given image \(I\).
Our image-caption ranking models learn a scoring function $S(I, C)$ such that the corresponding retrieval probabilities:

$$P_{im}(I|C) = \frac{\exp(S(I, C))}{\sum_{i=1}^{K} \exp(S(I_i, C))} \quad P_{cap}(C|I) = \frac{\exp(S(I, C))}{\sum_{i=1}^{K} \exp(S(I, C_i))}$$

are maximized. Let $S(I, C)$ be parameterized by $\theta$ (to be learnt). We formulate an objective function $L(\theta)$ for $S(I, C)$ as the sum of expected negative log-likelihoods of image retrieval and caption retrieval over all image-caption pairs $(I, C)$:

$$L(\theta) = \mathbb{E}_{(I, C)}[-\log P_{im}(I|C)] + \mathbb{E}_{(I, C)}[-\log P_{cap}(C|I)]$$

At test time we rank images and captions using $S(I, C)$.

Recent works on image-caption ranking often construct $S(I, C)$ by combining a vectorized image representation which is usually hidden layer activations in a CNN pretrained for image classification, with a vectorized caption representation which is usually a sentence encoding computed using an RNN in a multi-modal space. Such scoring functions rely on large image-caption ranking datasets to learn knowledge necessary for image-caption ranking and do not leverage knowledge in VQA corpora. We call such models VQA-agnostic models.

In this work we use the publicly available state-of-the-art image-caption ranking model of [3] as our baseline VQA-agnostic model. [3] projects a $D_xI$-dimensional CNN activation $x_I$ for image $I$ and a $D_xC$-dimensional RNN latent encoding $x_C$ for caption $C$ to the same $D_xC$-dimensional common multi-modal embedding space as unit-norm vectors $t_I$ and $t_C$:

$$t_I = \frac{W_I x_I}{||W_I x_I||_2} \quad t_C = \frac{x_C}{||x_C||_2}$$

The multi-modal scoring function is defined as their dot product $S_t(I, C) = \langle t_I, t_C \rangle$.

The VQA-agnostic model of [3] uses the 19-layer VGGNet [51] ($D_xI = 4096$) for image encoding and an RNN with 1024 Gated Recurrent Units [53] ($D_xC = 1024$) for caption encoding. The RNN and parameters $W_I$ are jointly learned on the image-caption ranking training set using a margin-based objective function.

### 3.2 VQA

VQA is the task of given an image $I$ and a free-form open-ended question $Q$ about $I$, generating a natural language answer $A$ to that question. Similarly, VQA-Caption task proposed by [1] takes a caption $C$ of an image and a question $Q$ about the image, then generates an answer $A$. In [1] the generated answers are evaluated using $\min\left(\frac{\#\text{humans that provided } A}{3}, 1\right)$. That is, $A$ is 100% correct if at least 3 humans (out of 10) provide the answer $A$.

We closely follow [1] and formulate VQA as a classification task over top $M = 1000$ most frequent answers from the training set. The oracle accuracies of picking the best answer for each question within this set of answers are 89.37% on training and 88.83% on validation. During training, given triplets of image $I$, question $Q$ and ground truth answer $A$, we optimize the negative log-likelihood (NLL) loss to maximize the probability of the ground truth answer $P_I(A|Q, I)$ given by the VQA model. Similarly
given triplets of caption $C$, question $Q$ and ground truth answer $A$, we optimize the NLL loss to maximize the VQA-Caption model probability $P_C(A|Q, C)$.

Following [1], for a VQA question $(I, Q)$ we first encode the input image $I$ using the 19-layer VGGNet [51] as a 4,096-dimensional image encoding $x_I$, and encode the question $Q$ using a 2-layer RNN with 512 Long Short-Term Memory (LSTM) units [58] per layer as a 2,048-dimensional question encoding $x_Q$. We then project $x_I$ and $x_Q$ into a common 1,024-dimensional multi-modal space as $z_I$ and $z_Q$:

$$z_I = Tanh(W_I x_I + b_I) \quad z_Q = Tanh(W_Q x_Q + b_Q) \quad (4)$$

As in [1] we then compute the representation $z_{I+Q}$ for the image-question pair $(I, Q)$ by element-wise multiplying $z_I$ and $z_Q$: $z_{I+Q} = z_I \odot z_Q$. The scores $s_A$ for 1,000 answers are given by:

$$s_A = W_s z_{I+Q} + b_s \quad (5)$$

We jointly learn the question encoding RNN and parameters $\{W_I, b_I, W_Q, b_Q, W_s, b_s\}$ during training.

For the VQA-Caption task given caption $C$ and question $Q$, we use the same network architecture and learning procedure as above, but using the most frequent 1,000 words in training captions as the dictionary to construct a 1,000 dimensional bag-of-words encoding for caption $C$ as $x_C$ to replace the image feature $x_I$ and compute $z_C$, $z_{C+Q}$ respectively.

The VQA and VQA-Caption models are learned on the train split of the VQA dataset [1] using 82,783 images, 413,915 captions and 248,349 questions. These models achieve VQA validation set accuracies of 54.42% (VQA) and 56.28% (VQA-Caption), respectively. Next, they are used as sub-modules in our image-caption ranking approach.

### 4 Approach

To leverage knowledge in VQA for image-caption ranking, we propose to represent the images and the captions in the VQA space using VQA and VQA-Caption models. We call such representations VQA-grounded representations.

#### 4.1 VQA-grounded representations

Let’s say we have a VQA model $P_I(A|Q, I)$, a VQA-Caption model $P_C(A|Q, C)$ and a set of $N$ questions $Q_i$ and their plausible answers (one for each question) $A_i$, $i = 1, 2, \ldots N$. Then given an image $I$ and a caption $C$, we first extract the $N$ dimensional VQA-grounded activation vectors $u_I$ for $I$ and $u_C$ for $C$ such that each dimension $i$ of $u_I$ and $u_C$ is the log probability of the ground truth answer $A_i$ given a question $Q_i$.

$$u^{(i)}_I = \log P_I(A_i|Q_i, I) \quad u^{(i)}_C = \log P_C(A_i|Q_i, C), i = 1, 2, \ldots, N \quad (6)$$

For example if the $(Q, A)$ pairs are $(Q_1: \text{What is the person riding?}, A_1: \text{Motorcycle})$ and $(Q_2: \text{What is the man wearing on his head?}, A_2: \text{Helmet})$, $u^{(1)}_I$ and $u^{(1)}_C$ verify
Fig. 2. Images and captions sorted by $P_I(A|Q, I)$ and $P_C(A|Q, C)$ assessed by our VQA (top) and VQA-Caption (bottom) models respectively. Indeed, images and captions that are more plausible for the $(Q, A)$ pair are scored higher by our VQA and VQA-Caption models.

if the person in image $I$ and caption $C$ respectively is riding a motorcycle. At the same time $u_I^{(2)}$ and $u_C^{(2)}$ verify whether the man in $I$ and $C$ is wearing a helmet.

In cases where there is not a man in the image or the caption, i.e. the assumption of $Q_i$ is not met, $P_I(A_i|Q_i, I)$ and $P_C(A_i|Q_i, C)$ may still reflect if there were a man or if the assumption of $Q_i$ were fulfilled, could he be wearing a helmet. In other words, even if there is no person present in the image or mentioned in the caption, the model may assess the plausibility of a man wearing a helmet or a motorcycle being present in the scene. This imagination beyond what is depicted in the image or caption can be helpful in providing additional information when reasoning about the compatibility between an image and a caption. We show qualitative examples of this imagination or plausibility assessment for selected $(Q, A)$ pairs in Figure 2, where we sort images and captions based on $P_I(A|Q, I)$ and $P_C(A|Q, C)$. Indeed, scenes where the corresponding fact $(Q, A)$ (e.g., man is wearing a helmet) is more likely to be plausible are scored higher.

Based on the activation vectors $u_I$ and $u_C$, we then compute the VQA-grounded vector representations $v_I$ and $v_C$ for $I$ and $C$ by projecting $u_I$ and $u_C$ to a $D_u$-dimensional vector embedding space:

$$v_I = \sigma(W_{u_I}u_I + b_{v_I})$$
$$v_C = \sigma(W_{u_C}u_C + b_{v_C})$$

(7)

Here $\sigma$ is a non-linear activation function.

By verifying question-answer pairs on image $I$ and caption $C$ and computing vector representations on top of them, the VQA-grounded representations $v_I$ and $v_C$ explicitly
Fig. 3. We propose two approaches: score-level fusion (left) and representation-level fusion (right) to utilize VQA for image-caption ranking. We propose to use VQA and VQA-Caption models as “feature extraction” schemes for images and captions. We use those features to construct VQA-grounded representations for images and captions. The score-level fusion approach combines the scoring functions of a VQA-grounded model and a baseline VQA-agnostic model. The representation-level fusion approach combines VQA-grounded representations and VQA-agnostic representations to produce a VQA-aware scoring function.

4.2 Score-level fusion

A simple strategy to combine our VQA-grounded model with a VQA-agnostic image-ranking model is to combine them at the score level. Given image $I$ and caption $C$, we first compute the VQA-grounded score as the dot product between the VQA-grounded representations of image and caption $S_v(I, C) = \langle v_I, v_C \rangle$. We then combine it with the VQA-agnostic scoring function $S_t(I, C)$ to get the final scoring function $S(I, C)$:

$$S(I, C) = \alpha S_t(I, C) + \beta S_v(I, C) + \gamma$$

In experiments, we first learn $\{W_{u_I}, b_{u_I}, W_{u_C}, b_{u_C}\}$ to project $u_I$ and $u_C$ to the VQA-grounded representations $v_I$ and $v_C$ on the image-caption ranking training set, and then learn parameters $\alpha$, $\beta$ and $\gamma$ on a held out validation set to avoid overfitting.

4.3 Representation-level fusion

An alternative to combining the VQA-agnostic and VQA-grounded representations at the score level is to inject the VQA-grounding at the representation level. Given the VQA-agnostic $D_t$-dimensional image and caption representations $t_I$ and $t_C$ used by
the baseline model, we first compute the VQA-grounded representation $v_I$ for image and $v_C$ for caption introduced in Section 4.1, and then combined them with VQA-agnostic representations to produce VQA-aware representations $r_I$ for image $I$ and $r_C$ for caption $C$ by projecting them to a $D_r$-dimensional multi-modal embedding space as follows:

$$r_I = \sigma(W_{tI}t_I + W_{vI}v_I + b_r) \quad r_C = \sigma(W_{tC}t_C + W_{vC}v_C + b_r)$$

(9)

The final image-caption ranking score is then

$$S(I,C) = \langle r_I, r_C \rangle$$

(10)

In experiments, we jointly learn parameters $\{W_{uI}, b_{uI}, W_{uC}, b_{uC}\}$ for projecting $u_I$ and $u_C$ to the VQA-grounded representations $v_I, v_C$ with parameters $\{W_{tI}, W_{vI}, b_{rI}, W_{tC}, W_{vC}, b_{rC}\}$ for computing the combined VQA-aware representations $r_I$ and $r_C$ on the image-caption ranking training set by optimizing Eq. 2.

Score-level fusion and representation-level fusion models are implemented as multi-layer neural networks. All activation functions $\sigma$ are ReLU($x$) = max($x$, 0) (for speed) and dropout layers [59] are inserted after all ReLU layers to avoid overfitting. We set the dimensions of the multi-modal embedding spaces $D_v$ and $D_r$ to 4,096 so they are large enough to capture necessary concepts for image-caption ranking. Optimization hyperparameters are selected on the validation set. We optimize both models using RMSProp with batch size 1,000 at learning rate 1e-5 for score-level fusion and 1e-4 for representation-level fusion. Optimization runs for 100,000 iterations with learning rate decay every 50,000 iterations.

Our main results in Section 5.1 use $N = 3000$ question-answer pairs, sampled 3 questions per image with their ground truth answers with respect to their original images from 1,000 random VQA training images. We discuss using different numbers of question-answer pairs $N$ and different strategies for selecting the question-answer pairs in Section 5.5.

5 Experiments and Results

We report results on MSCOCO [2] which is the largest available image-caption ranking dataset. Following the splits of [45,3] we use all 82,783 MSCOCO train images with 5 captions per image as our train set, 413,915 image-caption pairs in total. Note that this is the same split as the train split in the VQA dataset [1] we used to train our VQA and VQA-Caption models. The validation set consists of 1,000 images sampled from the original MSCOCO validation images. The test set consists of 5,000 images sampled from the original MSCOCO validation images that were not in the image-caption ranking validation set. Same as the train set, there are 5 captions available for each validation and test image.

We follow the evaluation metric of [45] and report caption and image retrieval performances on the first 1,000 test images following [45,60,48,43]. Given a test image, the caption retrieval task is to find any 1 out of its 5 captions from all 5,000 test captions. Given a test caption, the image retrieval task is to find its original image from all 1,000 test images. We report recall@$(1, 5, 10)$, which measures the fraction of times a correct item was found among the top $(1, 5, 10)$ predictions.
Table 1 shows our main results on MSCOCO. Our score-level fusion VQA-aware model using \( N = 3000 \) question-answer pairs (“\( N = 3000 \) score-level fusion VQA-aware”) achieves 46.9% caption retrieval recall@1 and 35.8% image retrieval recall@1. This model shows an improvement of 3.5% caption retrieval recall@1 and 4.8% image retrieval recall@1 over the state-of-the-art VQA-agnostic model of [3].

Our representation-level fusion approach adds an additional layer on top of the VQA-grounded and VQA-agnostic representations, resulting in a deeper model, so we experiment with adding an additional layer to the VQA-agnostic model for a fair comparison. That is equivalent to representation-level fusion using \( N = 0 \) question-answer pair (“\( N = 0 \) representation-level fusion”, i.e. deeper VQA-agnostic). Comparing with the VQA-agnostic model of [3], adding this additional layer improves performance by 2.4% caption retrieval recall@1 and 2.6% image retrieval recall@1.

By leveraging VQA knowledge our “\( N = 3000 \) representation-level fusion VQA-aware” model achieves 50.5% caption retrieval recall@1 and 37.0% image retrieval recall@1, which further improves 4.7% and 3.4% over the \( N = 0 \) VQA-agnostic representation-level fusion model. These improvements are consistent with our score-level fusion approach so this shows that the VQA corpora consistently provide complementary information to image-caption ranking.

To the best of our knowledge, the \( N = 3000 \) representation-level fusion VQA-aware result is the best result on MSCOCO image-caption ranking and significantly surpasses previous best results by as much as 7.1% in caption retrieval recall@1 and 4.4% image retrieval recall@1.

Our VQA-grounded model alone (“\( N = 3000 \) score-level fusion VQA-grounded only”) achieves 37.0% caption retrieval recall@1 and 26.2% image retrieval recall@1.
This indicates that the VQA activations $u_I$ and $u_C$ which evaluate the plausibility of facts (question-answer pairs) in images and captions are informative representations.

Figure 4 shows qualitative results on image retrieval comparing our approach ($N = 3000$ score-level fusion) with the VQA-agnostic model. By looking at several top retrieved images from our model for the failure case (last column), we find that our model seems to have picked up on a correlation between bats and helmets. It seems to be looking for helmets in retrieved images, while the ground truth image does not have one.

We also experiment with using the hidden activations available in the VQA and VQA-Caption models ($z_I$ for images and $z_C$ for captions in Section 3.2) as image and caption encodings in place of the image and caption VQA activations ($u_I$ for images and $u_C$ for captions in Section 4.1). Using these hidden activations of the VQA models is conceptually similar to using the hidden activations of CNNs pretrained on ImageNet as features [61]. These features achieve 46.8% caption retrieval recall@1 and 35.2% image retrieval recall@1 for score-level fusion, and 49.3% caption retrieval recall@1 and 37.9% image retrieval recall@1 for representation-level fusion which are as good as our semantic features $u_I$ and $u_C$. This shows that our semantically meaningful features, $u_I$ and $u_C$, performs as well as their corresponding non-semantic representations $z_I$ and $z_C$ using both score-level fusion and representation-level fusion. Note that such hidden activations may not always be available in different VQA models and the semantic features have the added benefit of being interpretable (e.g., Figure 2).

5.2 Ablation study

As an ablation study, we compare the following four models: 1) full representation-level fusion: our full $N = 3000$ representation-level fusion model that includes both image and caption VQA representations; 2) caption-only representation-level fusion: the same representation-level fusion model but using the VQA representation only for the caption, $t_C$, and not for the image; 3) image-only representation-level fusion: the same
model but using the VQA representation only for the image, \( t_I \), and not for the caption; 4) deeper VQA-agnostic: The \( N = 0 \) representation-level fusion model described earlier that does not use VQA representations for neither the image nor the caption.

Table 2 summarizes the results. We see that incrementally adding more VQA-knowledge improves performance. Both caption-only and image-only models outperform the \( N = 0 \) deeper VQA-agnostic baseline. The full representation-level fusion model which combines both representations yields the best performance.

### 5.3 The role of VQA and caption annotations

In this work we transfer knowledge from one vision-language task (i.e. VQA) to another (i.e. image-caption ranking). However, VQA annotations and caption annotations serve different purposes.

The target language to be retrieved is caption language, and not VQA language. [1] showed qualitatively and quantitatively that the two languages are statistically quite different (in terms of information contained, and in terms of nouns, adjectives, verbs, etc. used). As a result, VQA can not be thought of as providing additional “annotations” for the captioning task. Instead, VQA provides different perspectives/views of the images (and captions). It provides an additional feature representation. To better utilize this representation for an image-caption ranking task, one would still require sufficient ground truth caption annotations for images. In fact, with varying amounts of ground truth (caption) annotations, the VQA-aware representations show improvements in performance across the board. See Figure 5 (left).

A better analogy of our VQA representation is hidden activations (e.g., fc7) from a CNN trained on ImageNet. Having additional ImageNet annotations would improve the fc7 feature. But to map this fc7 feature to captions, one would still require sufficient caption annotations. Conceptually, caption annotations and category labels in ImageNet play two different roles. The former provides ground truth for the target task at hand (image-caption ranking), and having additional annotations for the target application typically helps. The latter helps learn a better image representation (which may provide improvements in a variety of tasks). Evaluating if the VQA representation can be transferred to other tasks beyond image-caption ranking is part of future work.

---

**Table 2.** Ablation study evaluating the gain in performance as more VQA-knowledge is incorporated in the model.

| Approach                      | Caption Retrieval | Image Retrieval |
|-------------------------------|-------------------|-----------------|
|                               | \( R@1 \) \( R@5 \) \( R@10 \) | \( R@1 \) \( R@5 \) \( R@10 \) |
| Deeper VQA-agnostic           | 45.8 76.8 86.1    | 33.6 67.8 81.0  |
| Caption-only representation-level fusion | 47.3 77.3 86.6    | 35.5 69.3 81.9  |
| Image-only representation-level fusion | 47.0 80.0 89.6    | 36.4 70.1 82.3  |
| Full representation-level fusion | **50.5 80.1 89.7** | **37.0 70.9 82.9** |
5.4 Number of question-answer pairs

Our VQA-grounded representations extract image and caption features based on question-answer pairs. It is important for there to be enough question-answer pairs to cover necessary aspects for image-caption ranking. We experiment with using $N = 30, 90, 300, 900, 3000$ $(Q, A)$ pairs (or facts) for both score-level and representation-level fusion. Figure 5 (right) shows caption and image retrieval performances of our approaches with varying $N$. Performance improves quickly from $N = 30$ to $N = 300$, and then starts to level off after $N = 300$.

An alternative to sampling 3 question-answer pairs per image on 1,000 images to get $N = 3000$ questions is to sample 1 question-answer pair per image from 3,000 images. Sampling multiple $(Q, A)$ pairs from the same image provides correlated $(Q, A)$ pairs. For example $(Q$: What are these animals? $A$: Giraffes) and $(Q$: Would this animal fit in a house? $A$: No). Using such correlated $(Q, A)$ pairs, the model could potentially better predict if there is a giraffe in the image by jointly reasoning if the animal looks like a giraffe and the if the animal would fit in a house, if the VQA and VQA-Caption models have not already picked up such correlations. Our results show that sampling 3 question-answer pairs per image for correlated $(Q, A)$ pairs does not significantly outperform sampling 1 question-answer pair per image, so we hypothesize that our VQA and Caption-QA models have already captured such correlations.

5.5 On the interpretability of the VQA-aware model

Deep models are well known to be difficult to interpret. Using image-captioning models as an example, it is difficult to tell based on which facts from the image the model generates the caption, or why it fails when it does. Lacking understanding of the model, common practices often resign to adding more training data and using more complex architectures and hoping the performance improves.
By using VQA as a submodule, the VQA-aware model presents opportunities for us to probe the model: “which fact do you want verified for this image for better caption retrieval?” That could help make the model more transparent and interpretable, allowing us to potentially improve the model more effectively, or strategically (e.g., via active learning).

We ran a proof-of-concept qualitative experiment. Recall that in our VQA-aware model, each \((Q, A)\) pair represents a fact about the image. We compute the mutual information between a fact’s validity for an image and the relevance of a caption for the image. Computing this mutual information requires an estimate of the joint distribution over the fact and the caption. We assume that the fact’s validity and the caption’s relevance are independent conditioned on model parameters (a VQA model for fact validity, and an image-caption model for caption relevance). To marginalize the models out, we use ideas from [62] which showed that turning on dropout at test time allows unbiased sampling of model architectures. We identify the most “informative” facts or \((Q, A)\) pairs for an image whose validity has the highest mutual information with captions.

Figure 6 shows such most informative \((Q, A)\) pairs for caption retrieval selected using our \(N = 3000\) representation-level fusion VQA-aware model. Exploring active learning and other human-in-the-loop applications is part of future work.

6 Conclusion

VQA corpora provide rich multi-modal information that is complementary to knowledge stored in image captioning corpora. In this work we take the novel perspective of viewing VQA as a “feature extraction” module that captures VQA knowledge. We propose two approaches – score-level and representation-level fusion – to integrate this knowledge into an existing image-caption ranking model. We set new state-of-the-art by improving caption retrieval by 7.1% and image retrieval by 4.4% on MSCOCO.

Improved individual modules, i.e., VQA models and VQA-agnostic image-caption ranking models, end-to-end training, and an attention mechanism that selects question-answer pairs (facts) in an image-specific manner may further improve the performance of our approach.
References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual question answering. In: ICCV. (2015)
2. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollr, P.: Microsoft coco: Common objects in context. In: ECCV. (2014)
3. Kiros, R., Salakhutdinov, R., Zemel, R.: Unifying visual-semantic embeddings with multimodal neural language models. In: TACL. (2015)
4. Ma, L., Lu, Z., Li, H.: Learning to answer questions from image using convolutional neural network. arXiv preprint arXiv:1506.00333 (2015)
5. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. In: NIPS. (2014)
6. Ren, M., Kiros, R., Zemel, R.: Exploring models and data for image question answering. In: NIPS. (2015)
7. Yu, L., Park, E., Berg, A.C., Berg, T.L.: Visual madlibs: Fill in the blank image generation and question answering. arXiv preprint arXiv:1506.00278 (2015)
8. Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., Xu, W.: Are you talking to a machine? dataset and methods for multilingual image question answering. In: NIPS. (2015)
9. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: A neural-based approach to answering questions about images. In: ICCV. (2015)
10. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR. (2009)
11. Branson, S., Wah, C., Schroff, F., Babenko, B., Welinder, P., Perona, P., Belongie, S.: Visual recognition with humans in the loop. In: ECCV. (2010)
12. Wang, Y., Mori, G.: A discriminative latent model of object classes and attributes. In: ECCV. (2010)
13. Berg, T., Belhumeur, P.N.: Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In: CVPR. (2013)
14. Zhang, N., Farrell, R., Iandola, F., Darrell, T.: Deformable part descriptors for fine-grained recognition and attribute prediction. In: ICCV. (2013)
15. Bourdev, L., Maji, S., Brox, T., Malik, J.: Detecting people using mutually consistent poselet activations. In: ECCV. (2010)
16. Zhang, N., Paluri, M., Ranzato, M., Darrell, T., Bourdev, L.: Panda: Pose aligned networks for deep attribute modeling. In: CVPR. (2014)
17. Li, L.J., Su, H., Fei-Fei, L., Xing, E.P.: Object bank: A high-level image representation for scene classification & semantic feature sparsification. In: NIPS. (2010)
18. Sadanand, S., Corso, J.J.: Action bank: A high-level representation of activity in video. In: CVPR. (2012)
19. Gupta, A., Davis, L.S.: Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In: ECCV. (2008)
20. Tang, K., Paluri, M., Fei-fei, L., Fergus, R., Bourdev, L.: Improving image classification with location context. In: ICCV. (2015)
21. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: ICCV. (2015)
22. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR. (2009)
23. Parikh, D., Grauman, K.: Relative attributes. In: ICCV. (2011)
24. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Describable visual attributes for face verification and image search. In: IEEE TPAMI. (2011)
25. Kovashka, A., Parikh, D., Grauman, K.: Whittlesearch: Image Search with Relative Attribute Feedback. In: CVPR. (2012)
26. Donahue, J., Grauman, K.: Annotator rationales for visual recognition. In: ICCV. (2011)
27. Parkash, A., Parikh, D.: Attributes for classifier feedback. In: ECCV. (2012)
28. Socher, R., Ganjoo, M., Manning, C.D., Ng, A.: Zero-shot learning through cross-modal transfer. In: NIPS. (2013)
29. Elhoseiny, M., Saleh, B., Elgammal, A.: Write a classifier: Zero-shot learning using purely textual descriptions. In: ICCV. (2013)
30. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D., Bernstein, M., Fei-Fei, L.: Image retrieval using scene graphs. In: CVPR. (2015)
31. Antol, S., Zitnick, C.L., Parikh, D.: Zero-shot learning via visual abstraction. In: ECCV. (2014)
32. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: Generating sentences from images. In: ECCV. (2010)
33. Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Baby talk: Understanding and generating simple image descriptions. In: CVPR. (2011)
34. Hamrick, J., Battaglia, P., Tenenbaum, J.B.: Internal physics models guide probabilistic judgments about object dynamics. In: Proceedings of the 33rd Annual Meeting of the Cognitive Science Society, Boston, MA. (2011)
35. Zheng, B., Zhao, Y., Yu, J., Ikeuchi, K., Zhu, S.C.: Beyond point clouds: Scene understanding by reasoning geometry and physics. In: CVPR. (2013)
36. Fouhey, D.F., Zitnick, C.L.: Predicting object dynamics in scenes. In: CVPR. (2014)
37. Walker, J., Gupta, A., Hebert, M.: Patch to the future: Unsupervised visual prediction. In: CVPR. (2014)
38. Pirsiavash, H., Vondrick, C., Torralba, A.: Inferring the why in images. CoRR abs/1406.5472 (2014)
39. Zhu, Y., Fathi, A., Fei-Fei, L.: Reasoning about object affordances in a knowledge base representation. In: ECCV. (2014)
40. Lin, X., Parikh, D.: Don’t just listen, use your imagination: Leveraging visual common sense for non-visual tasks. In: CVPR. (2015)
41. Vedantum, R., Lin, X., Batra, T., Zitnick, C.L., Parikh, D.: Learning common sense through visual abstraction. In: ICCV. (2015)
42. Sadeghi, F., Divvala, S.K., Farhadi, A.: Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In: CVPR. (2015)
43. Zhu, Y., Zhang, C., Re, C., Fei-Fei, L.: Building a large-scale multimodal knowledge base for visual question answering. arXiv preprint arXiv:1310.1531 (2013)
44. Vondrick, C., Pirsiavash, H., Torralba, A.: Anticipating the future by watching unlabeled video. arXiv preprint arXiv:1504.08023 (2015)
45. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR. (2015)
46. Chen, X., Lawrence Zitnick, C.: Mind’s eye: A recurrent visual representation for image caption generation. In: CVPR. (2015)
47. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML. (2015)
48. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn). In: ICLR. (2015)
49. Ma, L., Xu, Z., Shang, L., Li, H.: Multimodal convolutional neural networks for matching image and sentence. In: ICCV. (2015)
50. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012)
51. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2014)
52. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR. (2015)
53. Cho, K., van Merriënoor, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014)
54. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: NIPS. (2014)
55. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. (2009)
56. Elliott, D., Keller, F.: Comparing automatic evaluation measures for image description. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. (2014) 452–457
57. Vedantum, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. In: CVPR. (2015)
58. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8) (1997) 1735–1780
59. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. JMLR (2014)
60. Klein, B., Lev, G., Sadeh, G., Wolf, L.: Associating neural word embeddings with deep image representations using fisher vectors. In: CVPR. (2015)
61. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. arXiv preprint arXiv:1310.1531 (2013)
62. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. arXiv preprint arXiv:1506.02142 (2015)