Weakly Supervised Grounding for VQA in Vision-Language Transformers

Aisha Urooj Khan[1], Hilde Kuehne[2,3], Chuang Gan[3], Niels Da Vitoria Lobo[1], and Mubarak Shah[1]

[1] University of Central Florida, Orlando, FL, USA
[2] Goethe University Frankfurt, Frankfurt, Hesse, Germany
[3] MIT-IBM Watson AI Lab, Cambridge, MA, USA

Abstract. Transformers for visual-language representation learning have been getting a lot of interest and shown tremendous performance on visual question answering (VQA) and grounding. However, most systems that show good performance of those tasks still rely on pre-trained object detectors during training, which limits their applicability to the object classes available for those detectors. To mitigate this limitation, this paper focuses on the problem of weakly supervised grounding in the context of visual question answering in transformers. Our approach leverages capsules by transforming each visual token into a capsule representation in the visual encoder; it then uses activations from language self-attention layers as a text-guided selection module to mask those capsules before they are forwarded to the next layer. We evaluate our approach on the challenging GQA as well as VQA-HAT dataset for VQA grounding. Our experiments show that: while removing the information of masked objects from standard transformer architectures leads to a significant drop in performance, the integration of capsules significantly improves the grounding ability of such systems and provides new state-of-the-art results compared to other approaches in the field[4].

Keywords: multimodal understanding, visual grounding, visual question answering, vision and language

1 Introduction

Empowering VQA systems to be explainable is important for a variety of applications such as assisting visually-impaired people to navigate [16, 64] or helping radiologists in early diagnosis of fatal diseases [1,65]. A system that only produces a good answering accuracy will not be sufficient in these applications. Instead, VQA systems for such uses should ideally also provide an answer verification mechanism and grounding is a convincing way to obtain this direct verification.

On the heels of success in natural language processing and multi-modal understanding, a variety of transformer-based methods have been introduced for

[4] Code is available at https://github.com/aurooj/WSG-VQA-VLTransformers

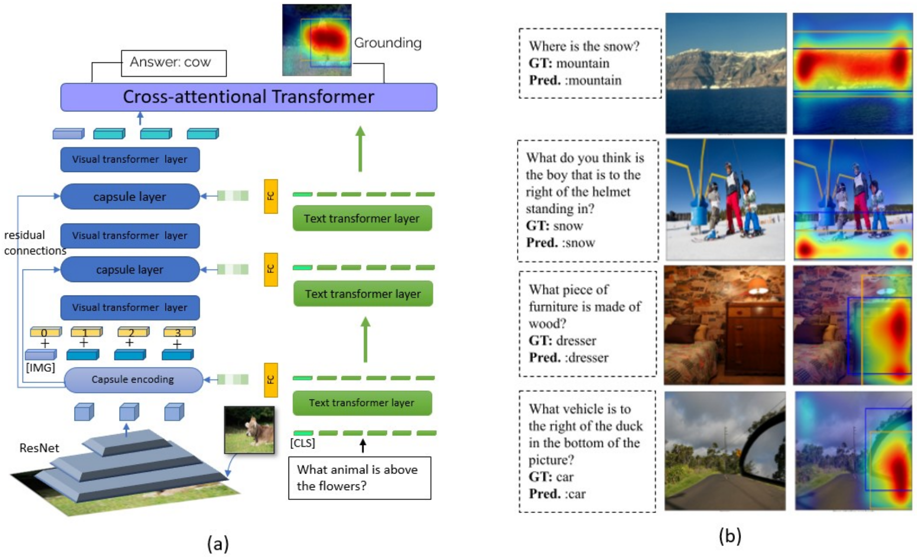


Fig. 1: **(a) Proposed architecture**: Given the question-image pair, grid features are used to obtain visual capsules using a Capsule encoding layer. Output embedding for [CLS] token from text transformer layer is then used to do capsule features selection. Selected capsules encodings with position information is then input to the visual encoder. We use sentence embedding from each textual transformer layer to select capsules for each visual transformer layer. The selected capsules are then input to the next visual transformer layer along with the output of the previous layer in the visual encoder. Finally, a cross-attentional block allows for a fine-grained interaction between both modalities to predict the answer. **(b) Attention from the proposed vision-language transformer with VQA supervision.** We look at the self-attention of the [IMG] token on the heads of the last layer. These maps show that the model automatically learns to ground relevant objects with VQA-only supervision leading to weakly-supervised grounded VQA (GVQA). *blue box:ground truth, orange:predicted box.*

joint representation learning of vision and language and respective downstream tasks, including VQA. Such approaches, e.g., [34, 40, 57] are usually trained based on region masks of detected objects generated by a pre-trained object detector [26]. The assumption that object masks are provided at input time limits detection to pretrained objects only and comes at the risk of image context information being unused.

Detector-free methods avoid this bias toward pre-trained object classes while being simpler and faster because they do not need to extract region-based features from a pre-trained object detector. Other works [10, 22, 28, 49] have therefore focused on removing the dependency on object detectors while achieving comparable if not better performance (e.g., on retrieval and VQA tasks). Among those, [10] and [49] also show good visual representations by qualitative examples, but do not provide an evaluation of their answer (or question) grounding ability.

In this work, we address this issue and focus on the problem of weakly supervised grounding for the VQA task in visual-language transformer based systems. For the input image-question pair, we want to answer the question as well as localize the relevant question and answer objects with only VQA supervision. Compared to detector-free referential expression grounding [7, 38, 61], VQA does not assume that the region description is always part of the input phrase as the answer word may not be present in the input question. It is therefore inadequate to only learn a direct mapping between text and image features. Instead, it requires processing multiple image-text mapping steps with the correct relation between them.

To address the task of VQA grounding in transformer-based architectures with the question-answering supervision alone, we propose a generic extension of the visual encoder part of a visual-language transformer based on visual capsule layers together with a text-guided selection module. Capsule networks learn to group neurons into visual entities. They thus try to capture entities (objectness) and relationships among them with promising results on various visual detection and segmentation tasks [12, 13, 31]. To make use of this ability in the context of transformers, we transfer inputs as well as intermediate layers' feature tokens to capsule encodings, from which the most relevant ones will be selected by the textual tokens of a parallel language encoder. This text-guided selection facilitates choosing features at entity-level, similar to attending object features, instead of an independent feature selection. We interleave transformer layers with such masked residual capsule encodings. This extension provides a combination of visual input routing and text-based masking which significantly improves the visual grounding ability of such systems.

We evaluate existing methods as well as the proposed approach on the challenging GQA and VQA-HAT datasets. To this end, we consider the attention output obtained from these methods and evaluate it on various metrics, namely overlap, intersection over union, and pointing game accuracy. Our results on the original architectures show a significant gap between task accuracy and grounding of existing methods indicating that existing vision-language systems are far from learning an actually grounded representation. The proposed method bridges the gap and outperforms the best contenders in terms of overlap, intersection over union, and pointing game accuracy achieving SOTA performance on the GQA dataset. It also achieves best mean-rank correlation score on VQA-HAT [8] dataset among methods which do not use attention supervision.

We summarize the contributions of our architecture as follows: a) we propose a capsule encoding layer to generate capsule-based visual tokens for transformers; b) we propose a text-guided capsule selection with residual connections to guide the visual features at each encoding step; and c) the proposed generic interleaved processing of capsules and self-attention layers can be integrated in various vision language architectures.

2 Related Work

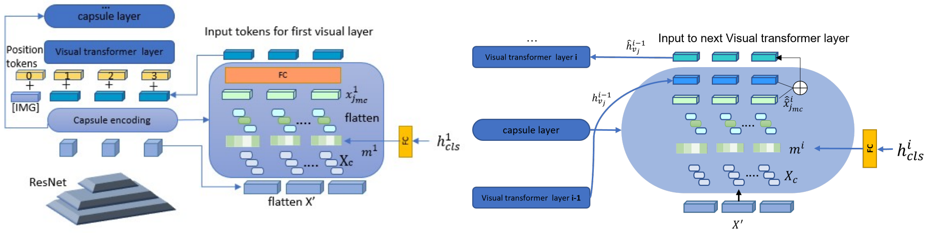
Visual-language Representation Learning. Learning a robust visual-language representation is currently an active area of research [27] with impressive progress on downstream tasks including VQA [6,32,34,35,39,40,43,56,57]. A majority of these methods rely on object detections making the downstream task simpler. Some works have attempted to avoid this dependency on object detections and show comparable performance using spatial features or image patches [21,22,28,33]. Our work falls into the later category and uses grid features as input.

Weakly-supervised Grounding and VQA. Weakly-supervised visual grounding is well studied for phrase-grounding in images [3,5,7,9,38,58,61]. Some works also focused on phrase grounding in videos [20,55,62]. However, less attention has been paid to grounding in VQA despite having significance for many critical applications. There is much research on making questions visually grounded [45,50,54,59,67,68], but only a handful of works focus on evaluating their grounding abilities [8,24,25,48,52,54]. GQA leverages scene graphs from Visual Genome dataset providing visual grounding labels for question and answer making it feasible to evaluate VQA logic grounding. Recently, xGQA [46] has been introduced as a multilingual version of the GQA benchmark. GQA [24] and [25] discuss the evaluation of VQA systems for grounding ability. VQA-HAT [8] on the other hand, provides human attention maps used for answering the question in a game-inspired attention-annotation setup. A handful of methods [41,48,63] evaluate their systems for correlation between machine-generated attention and human attention maps on VQA-HAT. With the emergence of transformers as the current SOTA, the focus moves towards grounding abilities of those systems for the VQA task. Unfortunately, none of these transformer-based methods have yet focused on the evaluation of weakly supervised grounding. Additionally, the fact that only few real-world datasets provide grounding labels makes this task challenging. We therefore finetune three existing detector-free transformer methods on GQA and evaluate them for the weak grounding task.

Transformers with Capsules. Some research has focused on the idea of combining transformers and capsules [11,15,37,42,44,47,60]. For instance, [60] studies text-summarization, image-processing, and 3D vision tasks; [44,47] uses capsule-transformer architecture for image-classification, and [37] uses capsules-transformers for predicting stock movements. To the best of our knowledge, the combination of capsules with transformers for VQA grounding has not been studied.

3 Proposed Approach

Given an input image-question pair with image I and question Q, we want to localize the relevant question and answer objects with only VQA supervision. We start from a two stream visual-language model (fig. 1) where the language encoder L_e guides input and intermediate respresentations of the visual encoder V_e. The



(a) Capsule Encoding layer (b) Capsule layer

Fig. 2: **(a) Capsule encoding layer**: grid features $X' \in \mathbb{R}^{h \times w \times d}$ are transformed into capsules X_c for each spatial position. Output embedding h^1_{cls} for $[CLS]$ token from the first text encoder layer generates a mask m^1 for capsule selection. The selected capsules X^1_{mc} are flattened along capsule dimension to get a set of visual tokens (of length $h * w$) where each token is denoted by $x^1_{jmc}, j = \{1, 2, ..., hw\}$; $x^1_{jmc} \in \mathbb{R}^{d_c}$ where d_c=capsule dimension, is then upsampled to model dimension d using a fully connected layer. A position embedding is added to visual tokens with the special token [IMG] at position 0. The output capsule encodings are then input to the visual transformer for future steps. **(b) Capsule layer** prepares the input for the next visual transformer layer i by combining previous layer's output (layer $i-1$) with selected capsules using output for [CLS] token h^i_{cls} from the i^{th} text encoder layer. Similar to the Capsule encoding, input tokens X' are first transformed into capsules X_c. The aggregated output feature h^i_{cls} from the i^{th} text encoder layer generates a mask m^i to select certain capsules for input to the i^{th} visual encoder layer. The resulting capsules are then flattened and upsampled (denoted by $\hat{x}^i_{jmc}$) and added to the output $h^{i-1}_{v_j}$ of the previous visual transformer layer $i-1$ to obtain input features $\hat{h}^i_{v_j}$ for visual transformer layer i.

input text to language encoder L_e is a sequence of word tokens from a vocabulary V appended with special tokens $[CLS]$ and $[SEP]$ at the start and end of the word tokens. As input to the visual encoder, our model takes convolutional features as image embeddings. The convolutional features $X \in \mathbb{R}^{h \times w \times d1}$ are extracted from a pre-trained ResNet model, h, w are the feature height and width, and $d1$ is the extracted features dimension. A 2D convolutional layer then yields an embedding X' of size $\mathbb{R}^{h \times w \times d}$, where d is the model dimension size. These input embeddings produce capsule encodings X_c as explained in section 3.2.

In the following, we first explain the motivation to use capsules in Sec. 3.1 followed by details about the capsule encoding in Sec. 3.2, the text-guided selection of the capsules in Sec. 3.3, as well as the text-based residual connection in Sec. 3.4. We close the section with an overview of the pretraining procedure in Sec. 3.5 and describe the details of the VQA downstream task in Sec. 4.1.

3.1 Capsule Networks

Standard neural networks lack the ability to dynamically represent a distinct part-whole hierarchy tree structure for each image [17]. This inability motivated the introduction of a type of model called Capsule Networks [18] which was later formalized in [53]. A Capsule Network is a neural network that is designed to model part-whole hierarchical relationships more explicitly than Convolutional Neural Networks (CNNs), by using groups of neurons to encode entities and learning the relationships between these entities. The promising performance of capsules can be attributed to their ability to learn part-whole relationships for object entities via routing-by-agreement [53] between different capsule layers. A capsule is represented by a group of neurons; each capsule layer is comprised of multiple capsules and multiple capsule layers can be stacked together. Capsule routing is a non-linear, iterative and clustering-like process that occurs between adjacent capsule layers, dynamically assigning *part* capsules i in layer ℓ to *object* capsules j in layer $\ell + 1$, by iteratively calibrating the routing coefficients $\boldsymbol{\gamma}$ [51]. Unlike most previous works which use a loss over object classes to learn a set of capsule classes, we do not have any object level supervision available for capsules, but instead combine the power of transformers and capsules by interleaving capsules as intermediate layers within the transformer and use VQA supervision to model visual entities as capsules.

3.2 Capsule Encodings

We use matrix capsules [19] as follows: given an image embedding $X' \in \mathbb{R}^{h \times w \times d}$, matrix capsules $X_c \in \mathbb{R}^{h \times w \times d_c}$, as shown in Figure 2(a), are obtained as follows: The image embedding X' is input to a convolutional layer producing primary capsules X_p where each capsule has a pose matrix of size $K \times K$ and an activation weight. The primary capsule layer outputs C_p number of capsules for each spatial location. The output dimensions for poses is $\mathbb{R}^{h \times w \times C_p \times K \times K}$ and for activation is $\mathbb{R}^{h \times w \times C_p \times 1}$. To treat each capsule as a separate entity, the pose matrix and activation are grouped together for each capsule. Hence, the primary capsules X_p have the dimensions $\mathbb{R}^{h \times w \times d_p}$ where $d_p = C_p \times (K \times K + 1)$. The primary capsules are then passed through an EM-Routing layer to vote for capsules in the next layer. Assuming we have C_v number of capsules in the next layer, the routing yields capsule encodings X_c where $X_c \in \mathbb{R}^{h \times w \times d_c}$, $d_c = C_v \times (K \times K + 1)$. We use an equal number of capsules in both layers, i.e., $C = C_p = C_v$. Our system employs the capsule representation X_c as visual embeddings.

Since transformers take a sequence of tokens as input, we flatten the capsule embeddings across spatial dimension to get a sequence of visual tokens of length $h * w$, where each visual token is denoted by $x_j \in \mathbb{R}^{d_c}$ for $j \in 1, 2, ..., hw$. A special trainable token $[IMG]$ is then concatenated to these tokens to form the final set of visual tokens $\{[IMG], x_1, x_2, ..., x_{hw}\}$. A learnable position embedding is added to these visual tokens to keep the notion of spatial position in the sequence. Each of the visual tokens except $[IMG]$ is represented by C capsules.

3.3 Text-guided Capsule Selection

As the language encoder is attending different words at each layer, we select visual capsules based on the text representation at each visual encoder layer. Let h^i_{cls} be the feature output corresponding to $[CLS]$ token from the i^{th} text encoder layer; we take the feature output h^1_{cls} corresponding to $[CLS]$ token from the first text encoder layer and input it to a fully connected layer ϕ. The output is C logits followed by a softmax function to learn presence probability $m^1 \in \mathbb{R}^C$ of attended words at layer 1. This mask is applied to X_c to select the respective capsules and mask out the rest resulting in the masked capsule representation X^1_{mc}.

$$m^1 = softmax(\phi(h^1_{cls})). \tag{1}$$

$$X^1_{mc} = m^1 \odot X_c \tag{2}$$

The masking is only applied to the visual tokens x_j without affecting $[IMG]$ token.

3.4 Text-based Residual Connection

To keep the capsule representation between intermediate layers, we add capsules via a residual connection to the inputs of each intermediate visual encoder layer. The input capsules to the intermediate layer are also selected based on the intermediate features output from the text encoder. Let m^i be the probability mask for attended words in the text feature output h^i_{cls} from the i^{th} layer:

$$m^i = softmax(\phi(h^i_{cls})), \forall i \in \{1, 2, ..., L\}, \tag{3}$$

and $x^i_{j_{mc}}$ denotes the j^{th} token with visual capsules selected using mask m^i.

$$x^i_{j_{mc}} = m^i \odot X_c, \tag{4}$$

The i^{th} visual encoder layer takes features from the $(i-1)^{th}$ layer to produce features $h^i_{v_j}$ for the j^{th} position. Let f^i_v be the i^{th} layer in the visual encoder. The output and input follow the notation below:

$$h^i_{v_j} = f^i_v(h^{i-1}_{v_j}). \tag{5}$$

To keep information flowing from text to image, we propose to add the residual connection from visual capsules for the j^{th} token by adding $x^i_{j_{mc}}$ to the input of the i^{th} encoder layer. However, there is a dimension mismatch between $x^i_{j_{mc}} \in \mathbb{R}^{d_c}$ and $h^{i-1}_{v_j} \in \mathbb{R}^d$. We upsample $x^i_{j_{mc}}$ to dimension size d using a fully connected layer σ and get the upsampled capsule-based features $\hat{x}^i_{j_{mc}} \in \mathbb{R}^d$. The input to the visual encoder layer will be as follows:

$$\hat{h}^{i-1}_{v_j} = f^i_v(h^{i-1}_{v_j} + \hat{x}^i_{j_{mc}}). \tag{6}$$

The output feature sequences from both encoders are then input to our cross attentional module which allows token-level attention between the two modalities. The aggregated feature outputs corresponding to $[CLS]$ and $[IMG]$ tokens after cross attention are input to a feature pooling layer followed by respective classifiers for pretraining and downstream tasks. We discuss the implementation about modality-specific encoders, feature pooling, and cross attention in detail in the supplementary material.

3.5 Training

To perform well, transformers require pretraining on large-scale datasets before finetuning for the downstream tasks, i.e., GQA and VQA-HAT in our case. Therefore, we first pretrain our capsules-transformer backbone on three pretraining tasks: image-text matching (ITM), masked language modeling (MLM), and visual question answering (VQA). The system is pre-trained in two stages: first, we do joint training of modality-specific encoders only to learn text-guided capsules representation; the representation learned in encoders is kept fixed during the second stage of pre-training where we add cross-attentional blocks on top of the modality encoders allowing token-level interaction between text features and visual features. While the first stage of pretraining uses pooled features from text and from visual encoders, the second stage pools features after cross attention: therefore, the second stage pre-training tasks uses cross-modal inputs as language and image features. For details about pretraining tasks in context of our method, refer to section 1.2 in supplementary. We finally finetune the pretrained capsules-transformer backbone to solve VQA as our downstream task.

4 Experiments and Results

4.1 Datasets

Pre-training. We use image-caption pairs from MSCOCO [36] and Visual Genome [30] for pretraining our backbone. Specifically, we use the same data as [57] which also include MSCOCO-based VQA datasets: Visual7W, VQAv2.0, and Visual Genome based GQA. However, we exclude the GQA validation set from pretraining and finetuning as we evaluate grounding on this set because scene graphs for GQA test and testdev are not publicly available. We use train sets of MSCOCO and VG with ∼7.5M sentence-image pairs for pretraining. MSCOCO val set is used for validating pretraining tasks.
Downstream. We consider two datasets for the downstream task, GQA [24] and VQA-HAT [8].
GQA poses visual reasoning in the form of compositional question answering. It requires multihop reasoning to answer the question, so GQA is a special case of VQA. GQA is more diverse than VQA2.0 [14] in terms of coverage of relational, spatial, and multihop reasoning questions. It has 22M QA pairs with 113K images. GQA provides ground truth boxes for question and answer objects making it a suitable test bed for our task.
VQA-HAT dataset provides human attention maps for VQA task. This dataset is based on VQA v1.0 [2] dataset and provides 1374 QA pairs with 488 images in the validation set. To evaluate on this dataset, we train our system on VQA v1.0. The answer vocabulary of VQA train set has a long tail distribution. We follow previous works [2,8] and use 1000 most frequent answers. We first combine training (248,349 QA pairs) and validation data (121,512 QA pairs) to get a total of 368,487 QA pairs. We then filter out the questions with out-of-vocabulary answers resulting in 318,827 QA pairs.

4.2 Evaluation Metrics

For GQA, VQA accuracy is reported for task accuracy. For grounding task on transformers, we take attention scores from [IMG] token to visual tokens from the last cross-attentional layer for all heads. Answer (or question) grounding performance is evaluated in terms of the following: **Overlap**– overlap between the ground truth bounding box for answer object and the detected attention region is reported in terms of precision (P), recall (R), and F1-score (F1); **IOU**– intersection over union (IOU) between the ground truth object and detected region is reported in terms of P, R, and F1-score. **Pointing Game**– proposed by [66] is a metric for weakly-supervised visual grounding methods. For pointing game, we consider the point detected from each head as a part of distribution, and perform k-means clustering (k=1) on those points. The cluster center is considered as the detected point from the system and used for evaluating accuracy. For VQA-HAT, we report **mean rank correlation** between system generated attention and human attention maps to compare with previous methods. Mean rank correlation is an order-based metric for finding the degree of association between two variables based on their ranks and thus is invariant to absolute spatial probability values [8].

4.3 Implementation details.

We use $L = 5$ layers in both text and image encoders, and 2 layers in cross-attention module. The transformer encoder layers have the same configuration as BERT with 12 heads and feature dimension $d = 768$. A batch size of 1024 with learning rate $lr = 1e-4$ is used for pretraining. First stage pretraining is done for 20 epochs and further trained for 10-15 epochs during the second stage. We use Imagenet pre-trained ResNet model to extract features of dimensions $7 \times 7 \times 2048$. For finetuning on GQA, we use batch size=32 with $lr = 1e-5$ and 5-10 training epochs. For VQA-HAT, we use batch size=64 with $lr = 4e-5$ trained for 20 epochs. To evaluate the grounding results, we follow [4] and consider the last cross-attentional layer's output for the attention map. To compute overlap and IOU for GQA, we threshold over the attention map with an attention threshold of 0.5 to get high attention regions. Each connected region is considered a detection. For pointing game, we find the single cluster center over maximum attention points from all heads and use it for evaluation. We ignore the test samples with empty ground truth maps for pointing game since there is no ground truth bounding box to check for a hit or a miss. For the VQA-HAT evaluation, we follow [8] and use mean rank correlation between the generated attention maps and the ground truth.

4.4 Comparison to State-of-the-Art

We compare the performance of our method to other best-performing methods in the field of weakly supervised VQA grounding and VQA in general, namely MAC [23] and MAC-Caps [25] as representation of visual reasoning architectures,

Method	Layer	Pointing Game Acc.
Random	-	18.80
Center	-	33.30
MAC [23]	mean	8.90
MAC-Caps [25]	mean	28.46
LXMERT [57]	last	29.00
ALBEF [33]-GC	last	32.13
ALBEF [33]-ATN	last	32.11
ViLT [28]	last	11.99
Ours-no-init (C=16)	last	**34.59**
Ours-no-init (C=32)	last	**34.43**
Ours-nogqa (C=32)	last	**37.04**

Table 1: Pointing game accuracy for GQA. For MAC and MAC-Caps, mean attention maps over reasoning steps are used. For transformer-based methods, maximum attention points from all heads are used for clustering. The cluster center is then used for the pointing game evaluation. For ALBEF, GC=GradCAM output, ATN=attention output. Ours-no-init is the full model trained from scratch (no initialization from BERT or ViT), Ours-nogqa uses no GQA samples at pretraining stage. Numbers are in percentages.

and LXMERT [57], ViLT [28], and ALBEF [33] as state-of-the-art transformer architectures without object features.

For LXMERT, we take the provided backbone pre-trained on object features and finetune it using image patches of size $32 \times 32 \times 3$ on GQA. In case of ViLT, we use the provided pre-trained backbone and finetune it on GQA. Following ViLT, we generate a heat map using the cosine similarity between image and word tokens evaluating the similarity scores as well as raw attention for grounding performance for all three metrics. For ALBEF we report results on the last layer as well as on layer 8 which is specialized for grounding [33] using the visualizations from gradcam as well as raw attention maps.

GQA We first look at the results of our evaluation on GQA, considering pointing game accuracy in Table 1 and for overlap and IOU in Table 2. Our method outperforms both MAC and MAC-Caps for answer grounding on the last attention map. We achieve an absolute gain of 16.47% (overlap F1-score) and 2.67% increase in IOU F1-score, and an improvement of 25.69% in pointing game accuracy for MAC. When compared with MAC-Caps, our best method (C=16, no-init) improves overlap F1-score by 15.71% ↑, IOU F1-score by 2.32% ↑, and pointing game accuracy by 6.13% ↑. Similar performance gain is observed for question grounding with an improvement of 38.2% ↑ for overlap F1-score and 3.67% ↑ gain for IOU F1-score.

To evaluate LXMERT finetuned on image patches (LXMERT-patches), we take the attention score maps from the last cross modality layer. We improve over LXMERT by 12.01% ↑ absolute points w.r.t overlap F1-score, 2.23% ↑ w.r.t. IOU F1-score and 5.43% ↑ gain in pointing game accuracy. For question grounding, LXMERT achieves an overlap F1-score: 43.08% (vs. ours 59.69%) and IOU F1-score of 4.62% (vs. ours 5.96%).

						Overlap			IOU		
Method	Obj.	Backbone	Pre-training	Layer	Acc.	P	R	F1	P	R	F1
MAC [23]	A	ResNet	-	last	57.09	5.05	24.44	8.37	0.76	3.70	1.27
MAC-Caps [25]	A	ResNet	-	last	55.13	5.46	27.9	9.13	0.97	4.94	1.62
LXMERT-patches	A	Faster RCNN	MSCO,VG	last	48.65	7.13	64.21	12.83	0.95	8.66	1.71
ALBEF [33]-GC	A	ViT+BERT		last	64.16	6.94	99.92	12.98	0.89	13.43	1.67
ALBEF [33]-ATN	A	ViT+BERT	MSCO,VG,	last	64.20	5.13	99.92	9.75	0.64	12.98	1.21
ALBEF [33]-GC	A	ViT+BERT	SBU,GCC	8	64.20	4.41	99.92	8.44	0.54	12.85	1.04
ALBEF [33]-ATN	A	ViT+BERT		8	64.20	4.82	99.92	9.19	0.60	12.88	1.14
ViLT [29]	A	ViT	MSCO,VG,	last-cos	66.33	0.34	6.13	0.65	0.04	0.63	0.07
ViLT [29]	A	ViT	SBU,GCC	last-ATN	66.33	0.28	4.10	0.53	0.08	1.20	0.15
Ours (C=16)	A	ResNet	MSCO,VG	last	57.21	**14.53**	85.47	**24.84**	**2.30**	**13.61**	**3.94**
MAC [23]	Q	ResNet	-	last	57.09	10.79	16.38	13.01	1.39	2.09	1.67
MAC-Caps [25]	Q	ResNet	-	last	55.13	17.39	28.10	21.49	1.87	2.96	2.29
LXMERT-patches	Q	Faster RCNN	MSCO,VG	last	48.65	32.46	64.02	43.08	3.48	6.87	4.62
ALBEF [33]-GC	Q	ViT+BERT		last	64.20	22.15	99.90	36.26	1.96	9.22	3.24
ALBEF [33]-ATN	Q	ViT+BERT	MSCO,VG,	last	64.20	16.50	99.90	28.33	1.40	8.90	2.43
ALBEF [33]-GC	Q	ViT+BERT	SBU,GCC	8	64.20	14.21	99.90	24.88	1.19	8.71	2.09
ALBEF [33]-ATN	Q	ViT+BERT		8	64.20	15.51	99.90	26.85	1.31	8.77	2.27
ViLT [29]	Q	ViT	MSCO,VG,	last-cos	66.33	1.02	5.64	1.73	0.10	0.54	0.17
ViLT [29]	Q	ViT	SBU,GCC	last-ATN	66.33	0.34	1.56	0.56	0.08	0.38	0.14
Ours (C=16)	Q	ResNet	MSCO,VG	last	57.21	**47.03**	81.67	**59.69**	**4.72**	8.08	**5.96**

Table 2: Results on GQA validation set (for last layer). All methods are evaluated for weak VQA grounding task. For transformer-based models, attention was averaged over all heads. Results are based on grounding of objects referenced in the answer (A) and the question (Q). C=no.of capsules, we report results from our best model with C=16. Refer to table 4 for more variants. For ViLT, we obtain results using cosine similarity (cos.) between text and image features as proposed by the authors as well as from raw attention scores (ATN). For ALBEF, GC is the gradcam output used for evaluation, ATN is the attention output. ALBEF uses layer 8 as grounding layer, we also report grounding performance on this layer. Our method outperforms all baselines for overlap F1-score and IOU F1-score. See section 4.4 for more details. Numbers are in percentages.

ViLT outperforms all methods in terms of VQA accuracy of 66.33%. However, on the grounding task, it demonstrates the lowest performance for all metrics (table 1: row 7, table 2: rows 8-9). Similar behavior is observed for the question grounding task.

ALBEF produces visualization using GradCAM. We compare with ALBEF using both GradCAM output and attention maps. ALBEF has a very high recall (R) both in terms of overlap and IOU. However, it lacks in precision (P) leading to lower F1-scores for both metrics. Our best model outperforms ALBEF-VQA by a significant margin on both answer grounding and question grounding.

VQA-HAT We further evaluate our system on the VQA-HAT dataset. To this end, we follow the protocol of VQA-HAT and resize the human attention maps and the output attention maps from our system to the common resolution of 14x14. We then rank both of them. VQA-HAT val set provides three human attention maps for each question. We compute the rank correlation of generated attention map with each human attention map and take the average score. Mean rank correlation score over all QA pairs is reported.

Method	Mean Rank-Correlation
Random	0.000 ± 0.001
Human	0.623 ± 0.003
Unsupervised	
SAN [63]	0.249 ± 0.004
HieCoAtt [41]	0.264 ± 0.004
Ours (C=16)	**0.479 ± 0.0001**
Supervised	
HAN [48]	0.668 ± 0.001

Table 3: Results on VQA-HAT val dataset. *Unsupervised*: no attention supervision, *Supervised*: use attention refinement.

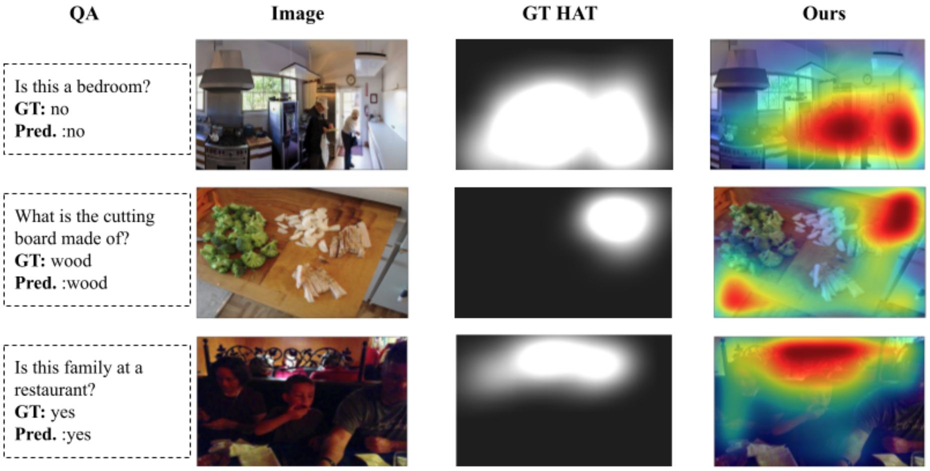


Fig. 3: **Success cases for VQA-HAT dataset.** VQA-HAT provides 3 human attention maps for each image. Here, we show the best matched ground truth map (GT HAT).

We compare our approach on VQA-HAT with three different baselines: SAN [63] and HieCoAtt [41] as unsupervised bounding box free systems, and HAN [48] which uses attention supervision during training. The evaluation is shown in table 3. It shows that the proposed system is able to significantly outperform both methods using VQA-only supervision.

Without any attention supervision during training, we are able to narrow the gap between unsupervised methods and methods such as HAN, which use human ground truth attention maps during training. Figure 3 shows success cases on VQA-HAT, comparing our generated attention result to the closest human attention map.

4.5 Ablations and Analyses

Impact of Residual Connections. We compare our full system with an ablated variant without residual connections. We observe a drop in performance

		Overlap			IOU			Pointing
Method	Acc.	P	R	F1	P	R	F1	Game
(1) no skip(C=32)	56.83	11.06	77.60	19.37	1.39	9.85	2.43	29.81
(2) w/skip (C=32)	55.41	**10.09**	**71.95**	**17.70**	**1.41**	**10.09**	**2.47**	**34.43**
(3) w/skip (C=16)	57.21	**14.53**	**85.47**	**24.84**	**2.30**	**13.61**	**3.94**	**34.59**
(4) w/skip (C=24)	56.26	10.90	74.03	19.00	1.54	10.56	2.69	31.08
(5) w/skip (C=32)	55.41	10.09	71.95	17.70	1.41	10.09	2.47	34.43
(6) w/skip (C=48)	53.65	10.28	68.94	17.89	1.59	10.73	2.78	29.70
(7) no-init (C=32)	55.41	**10.09**	**71.95**	**17.70**	**1.41**	**10.09**	**2.47**	**34.43**
(8) vit-bert init (C=32)	58.86	11.11	74.67	19.34	1.55	10.44	2.69	27.06

Table 4: Ablations over the design choices for the proposed architecture on GQA val set. Average attention over all heads in the last transformer layer is used to evaluate the grounding performance. We perform ablation study with C=32 caspules except rows 3-6 where we train the proposed architecture with varying number of capsules. Ablation (1) no skip is our system without residual connections, (2) w/skip is the full model. Results for the final design choices are shown in bold.

in terms of overlap, but a slight increase in terms of IOU. Without residual connections, pointing game accuracy is lower than with residual connections (4.62% ↓in table 4). We conclude that using residual connections is beneficial for pointing game.

Number of capsules. We ablate our system with varying number of capsules. We train the proposed system with C=16, 24, 32, and 48 capsules. We observe that increase in number of capsules not only decreases VQA accuracy, but also hurts the overlap and IOU in terms of precision, recall and F1-score. Our best method uses 16 capsules with residual connections and pre-trained from scratch.

ViT + BERT + Ours. ViLT and ALBEF initialize their image and text encoders from ViT and/or BERT weights. Although our model is shallower than both models (5 layers in modality specific encoders compared to 12 layers in ViLT and ALBEF), we experimented to initialize our text encoder with BERT weights and image encoder with ViT weights from last 5 layers. We find a gain in VQA accuracy (58.86% vs. 57.21%) but it get less grounding performance.

4.6 Qualitative Analysis

In figure 4, for all examples including the ones where our system mispredicted the answer, the grounding attention was correct (row 1,4 and 5). Also, the answers are plausible. For instance, in row 3, the correct answer is 'aircraft', and our method predicted it as 'airplane' with the correct localization. Overall, we notice that compared to our method, the baselines were either attending most of the image (ALBEF in rows 1,3, and 5 which explains the high recall in table 2), or generate small attention maps (MAC-Caps, ViLT) or look at the wrong part of the image (LXMERT). More examples and analysis are in the supplementary material.

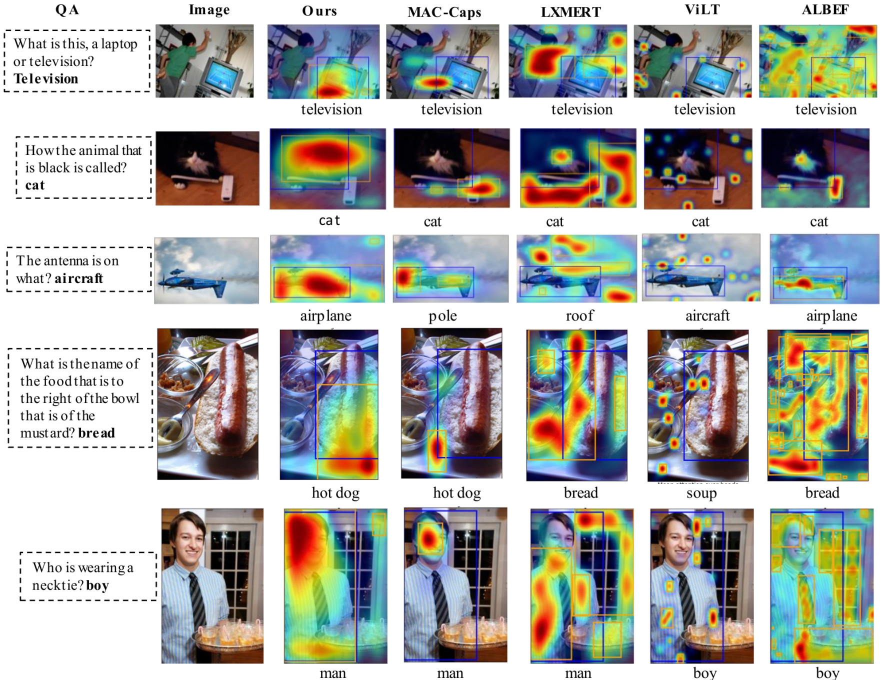

QA
Image
Ours
MAC-Caps
LXMERT
ViLT
ALBEF
What is this, a laptop or television? **Television**
television
television
television
television
television
How the animal that is black is called? **cat**
cat
cat
cat
cat
cat
The antenna is on what? **aircraft**
airplane
pole
roof
aircraft
airplane
What is the name of the food that is to the right of the bowl that is of the mustard? **bread**
hot dog
hot dog
bread
soup
bread
Who is wearing a necktie? **boy**
man
man
man
boy
boy

Fig. 4: **Qualitative comparison:** each row shows the input example, and the last layer's attention visualizations (averaged over all heads) with the predicted answer from all methods. Column 1 shows the question and ground truth answer, column 2 is the input image, column 3 shows the attention (grounding) output from our method, column 4-7 are results from the baselines. Blue box is the ground truth bounding box for the answer object, orange boxes are the detected regions from each system. We can see that ours is attending relevant answer object with the plausible predicted answer even when the prediction mismatches with the ground truth answer (row 3-5). In row 4, the question is vague; therefore we can say, except LXMERT, all methods choose the correct answer. ALBEF has attention spread over all image which explains the high recall it achieves for overlap and IOU. Refer to section 4.6 for more details and discussion. Best viewed in color.

5 Conclusion

In this work, we show the trade-off between VQA accuracy and the grounding abilities of the existing SOTA transformer-based methods. We use text-guided capsule representation in combination with transformer encoder layers. Our results demonstrate significant improvement over all baselines for all grounding metrics. Extensive experiments demonstrate the effectiveness of the proposed system over the baselines.

Acknowledgements. Aisha Urooj is supported by the ARO grant W911NF-19-1-0356. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. **Disclaimer:** The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ARO, IARPA, DOI/IBC, or the U.S. Government.

References

1. Abacha, A.B., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H.: Vqa-med: Overview of the medical visual question answering task at imageclef 2019. (2019)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
3. Arbelle, A., Doveh, S., Alfassy, A., Shtok, J., Lev, G., Schwartz, E., Kuehne, H., Levi, H.B., Sattigeri, P., Panda, R., et al.: Detector-free weakly supervised grounding by separation. arXiv preprint arXiv:2104.09829 (2021)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers (2021)
5. Chen, K., Gao, J., Nevatia, R.: Knowledge aided consistency for weakly supervised phrase grounding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4042–4050 (2018)
6. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Learning universal image-text representations (2019)
7. Chen, Z., Ma, L., Luo, W., Wong, K.Y.K.: Weakly-supervised spatio-temporally grounding natural sentence in video. arXiv preprint arXiv:1906.02549 (2019)
8. Das, A., Agrawal, H., Zitnick, C.L., Parikh, D., Batra, D.: Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2016)
9. Datta, S., Sikka, K., Roy, A., Ahuja, K., Parikh, D., Divakaran, A.: Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2601–2610 (2019)
10. Desai, K., Johnson, J.: VirTex: Learning Visual Representations from Textual Annotations. In: CVPR (2021)
11. Duan, S., Cao, J., Zhao, H.: Capsule-transformer for neural machine translation. arXiv preprint arXiv:2004.14649 (2020)
12. Duarte, K., Rawat, Y., Shah, M.: Videocapsulenet: A simplified network for action detection. In: Advances in Neural Information Processing Systems. pp. 7610–7619 (2018)
13. Duarte, K., Rawat, Y.S., Shah, M.: Capsulevos: Semi-supervised video object segmentation using capsule routing. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8480–8489 (2019)
14. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6904–6913 (2017)
15. Gu, S., Feng, Y.: Improving multi-head attention with capsule networks. In: CCF International Conference on Natural Language Processing and Chinese Computing. pp. 314–326. Springer (2019)
16. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3608–3617 (2018)

17. Hinton, G.: How to represent part-whole hierarchies in a neural network. arXiv preprint arXiv:2102.12627 (2021)
18. Hinton, G.E., Krizhevsky, A., Wang, S.D.: Transforming auto-encoders. In: International conference on artificial neural networks. pp. 44–51. Springer (2011)
19. Hinton, G.E., Sabour, S., Frosst, N.: Matrix capsules with em routing. In: International conference on learning representations (2018)
20. Huang, D.A., Buch, S., Dery, L., Garg, A., Fei-Fei, L., Niebles, J.C.: Finding" it": Weakly-supervised reference-aware visual grounding in instructional videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5948–5957 (2018)
21. Huang, Z., Zeng, Z., Huang, Y., Liu, B., Fu, D., Fu, J.: Seeing out of the box: End-to-end pre-training for vision-language representation learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
22. Huang, Z., Zeng, Z., Liu, B., Fu, D., Fu, J.: Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. CoRR **abs/2004.00849** (2020), `https://arxiv.org/abs/2004.00849`
23. Hudson, D.A., Manning, C.D.: Compositional attention networks for machine reasoning. International Conference on Learning Representations (ICLR) (2018)
24. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
25. Khan, A.U., Kuehne, H., Duarte, K., Gan, C., Lobo, N., Shah, M.: Found a reason for me? weakly-supervised grounded visual question answering using capsules (2021)
26. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. arXiv preprint arXiv:2101.01169 (2021)
27. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. arXiv preprint arXiv:2101.01169 (2021)
28. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 5583–5594. PMLR (18–24 Jul 2021), `http://proceedings.mlr.press/v139/kim21k.html`
29. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 5583–5594. PMLR (18–24 Jul 2021), `http://proceedings.mlr.press/v139/kim21k.html`
30. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1), 32–73 (2017)
31. LaLonde, R., Bagci, U.: Capsules for object segmentation. arXiv preprint arXiv:1804.04241 (2018)
32. Li, G., Duan, N., Fang, Y., Gong, M., Jiang, D.: Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11336–11344 (2020)
33. Li, J., Selvaraju, R.R., Gotmare, A.D., Joty, S., Xiong, C., Hoi, S.: Align before fuse: Vision and language representation learning with momentum distillation. In: NeurIPS (2021)

34. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
35. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: European Conference on Computer Vision. pp. 121–137. Springer (2020)
36. 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: European conference on computer vision. pp. 740–755. Springer (2014)
37. Liu, J., Lin, H., Liu, X., Xu, B., Ren, Y., Diao, Y., Yang, L.: Transformer-based capsule network for stock movement prediction. In: Proceedings of the First Workshop on Financial Technology and Natural Language Processing. pp. 66–73 (2019)
38. Liu, Y., Wan, B., Ma, L., He, X.: Relation-aware instance refinement for weakly supervised visual grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5612–5621 (2021)
39. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. arXiv preprint arXiv:1908.02265 (2019)
40. Lu, J., Goswami, V., Rohrbach, M., Parikh, D., Lee, S.: 12-in-1: Multi-task vision and language representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10437–10446 (2020)
41. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. Advances in neural information processing systems **29** (2016)
42. Mazzia, V., Salvetti, F., Chiaberge, M.: Efficient-capsnet: Capsule network with self-attention routing. arXiv preprint arXiv:2101.12491 (2021)
43. Miech, A., Alayrac, J.B., Laptev, I., Sivic, J., Zisserman, A.: Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9826–9836 (2021)
44. Mobiny, A., Cicalese, P.A., Nguyen, H.V.: Trans-caps: Transformer capsule networks with self-attention routing (2021), `https://openreview.net/forum?id=BUPIRa1D2J`
45. Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.S., Wen, J.R.: Counterfactual vqa: A cause-effect look at language bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12700–12710 (2021)
46. Pfeiffer, J., Geigle, G., Kamath, A., Steitz, J.M.O., Roth, S., Vulić, I., Gurevych, I.: xgqa: Cross-lingual visual question answering. arXiv preprint arXiv:2109.06082 (2021)
47. Pucci, R., Micheloni, C., Martinel, N.: Self-attention agreement among capsules. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 272–280 (2021)
48. Qiao, T., Dong, J., Xu, D.: Exploring human-like attention supervision in visual question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
49. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)
50. Ramakrishnan, S., Agrawal, A., Lee, S.: Overcoming language priors in visual question answering with adversarial regularization. arXiv preprint arXiv:1810.03649 (2018)

51. Ribeiro, F.D.S., Duarte, K., Everett, M., Leontidis, G., Shah, M.: Learning with capsules: A survey. arXiv preprint arXiv:2206.02664 (2022)
52. Riquelme, F., De Goyeneche, A., Zhang, Y., Niebles, J.C., Soto, A.: Explaining vqa predictions using visual grounding and a knowledge base. Image and Vision Computing **101**, 103968 (2020)
53. Sabour, S., Frosst, N., Hinton, G.E.: Dynamic routing between capsules. In: NIPS (2017)
54. Selvaraju, R.R., Lee, S., Shen, Y., Jin, H., Ghosh, S., Heck, L., Batra, D., Parikh, D.: Taking a hint: Leveraging explanations to make vision and language models more grounded. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2591–2600 (2019)
55. Shi, J., Xu, J., Gong, B., Xu, C.: Not all frames are equal: Weakly-supervised video grounding with contextual similarity and visual clustering losses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10444–10452 (2019)
56. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vl-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530 (2019)
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 (2019)
58. Wang, L., Huang, J., Li, Y., Xu, K., Yang, Z., Yu, D.: Improving weakly supervised visual grounding by contrastive knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14090–14100 (2021)
59. Whitehead, S., Wu, H., Ji, H., Feris, R., Saenko, K.: Separating skills and concepts for novel visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5632–5641 (June 2021)
60. Wu, L., Liu, X., Liu, Q.: Centroid transformers: Learning to abstract with attention. arXiv preprint arXiv:2102.08606 (2021)
61. Xiao, F., Sigal, L., Jae Lee, Y.: Weakly-supervised visual grounding of phrases with linguistic structures. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5945–5954 (2017)
62. Yang, X., Liu, X., Jian, M., Gao, X., Wang, M.: Weakly-supervised video object grounding by exploring spatio-temporal contexts. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 1939–1947 (2020)
63. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 21–29 (2016)
64. Zeng, X., Wang, Y., Chiu, T.Y., Bhattacharya, N., Gurari, D.: Vision skills needed to answer visual questions. Proceedings of the ACM on Human-Computer Interaction **4**(CSCW2), 1–31 (2020)
65. Zhan, L.M., Liu, B., Fan, L., Chen, J., Wu, X.M.: Medical visual question answering via conditional reasoning. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2345–2354 (2020)
66. Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. International Journal of Computer Vision **126**(10), 1084–1102 (2018)
67. Zhang, S., Qu, L., You, S., Yang, Z., Zhang, J.: Automatic generation of grounded visual questions. arXiv preprint arXiv:1612.06530 (2016)

68. Zhang, Y., Niebles, J.C., Soto, A.: Interpretable visual question answering by visual grounding from attention supervision mining. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 349–357 (2019). https://doi.org/10.1109/WACV.2019.00043