Sentence Attention Blocks for Answer Grounding

Seyedalireza Khoshsirat   Chandra Kambhamettu

Video/Image Modeling and Synthesis (VIMS) Lab, University of Delaware

{alireza, chandrak}@udel.edu

Abstract

Answer grounding is the task of locating relevant visual evidence for the Visual Question Answering task. While a wide variety of attention methods have been introduced for this task, they suffer from the following three problems: designs that do not allow the usage of pre-trained networks and do not benefit from large data pre-training, custom designs that are not based on well-grounded previous designs, therefore limiting the learning power of the network, or complicated designs that make it challenging to re-implement or improve them. In this paper, we propose a novel architectural block, which we term Sentence Attention Block, to solve these problems. The proposed block re-calibrates channel-wise image feature-maps by explicitly modeling inter-dependencies between the image feature-maps and sentence embedding. We visually demonstrate how this block filters out irrelevant feature-maps channels based on sentence embedding. We start our design with a well-known attention method, and by making minor modifications, we improve the results to achieve state-of-the-art accuracy. The flexibility of our method makes it easy to use different pre-trained backbone networks, and its simplicity makes it easy to understand and be re-implemented. We demonstrate the effectiveness of our method on the TextVQA-X, VQS, VQA-X, and VizWiz-VQA-Grounding datasets. We perform multiple ablation studies to show the effectiveness of our design choices.

1. Introduction

1.1. Visual Question Answering

Visual Question Answering (VQA) systems try to accurately answer natural language questions regarding an input image [1]. This topic aims at developing systems that can communicate effectively about an image in natural language and comprehend the contents of images similar to humans.

1.2. Answer Grounding

The answer grounding task is defined as detecting the pixels that can provide evidence for the answer to a given question regarding an image [3]. In other words, the task is to return the image regions used to arrive at the answer for a given visual question (question-image pair) with an answer. Although the VQA community has made significant progress, the best-performing systems are complicated black-box models, raising concerns about whether their answer reasoning is based on correct visual evidence. By understanding the reasoning mechanism of the model, we can evaluate the quality of answers, improve model performance, and provide explanations for end-users. To address this problem, answer grounding has been introduced into VQA systems, which requires the model to locate relevant image regions as well as answer visual questions.

By providing answer groundings in response to visual questions, numerous applications become possible. First, they allow for evaluating whether a VQA model is reasoning correctly based on visual evidence. This is useful as an explanation as well as for assisting developers with model debugging. Second, answer grounding makes it possible to separate important regions from irrelevant background regions. Given that non-professional users can mistakenly have private information in the background of their pictures, answer grounding is a useful tool to obfuscate the background for privacy preserving. Third, suppose a service can magnify the relevant visual evidence. In that case, users will be able to discover the needed information in less time. This is useful in part because VQA answers can be insufficient sometimes.

1.3. Multimodal Deep Learning

By definition, the VQA and answer grounding tasks are multimodal tasks since a method for these tasks should be able to process and correlate two different modalities.

Multimodal Joint-Embedding Models: These models merge and learn representations of multiple modalities in a joint feature space. Joint-embeddings underpin the building of a lot of cross-modal methods as they can bridge the gap between different modalities. In this joint space, the dis-

tance of different points is equivalent to the semantic distance between their corresponding original inputs.

Multimodal Attention-based Models: The main objective of the attention mechanism is to design systems that use local features of image or text, extract features from different regions, and assign priorities to them. The attention portion in an image determines salient regions. Then the language generation component focuses more on those salient regions for additional processing [35, 2, 12, 16, 26].

1.4. Attention Mechanism

In deep learning, attention is a mechanism that mimics cognitive attention. The goal is to enhance the essential features of the input data and vanish out the rest. Attention methods can be classified into two classes based on their inputs: self-attention and cross-attention. Self-attention (also known as intra-attention) is a type of attention that quantifies the interdependence within the elements of a single input. At the same time, cross-attention (also known as inter-attention) finds the interdependence across two or more inputs [40, 18, 17]. Usually, cross-attention methods are used for multimodal inputs [43]. Cross-attention models first process individual modalities using modality-specific encoders, then the encoded features are fed into cross-attention modules.

The Squeeze-and-Excitation method [13] is a channel-wise self-attention mechanism widely used in classification networks [37, 38]. It consists of a global average pooling of the input, followed by two linear layers with an interleaved non-linearity and a sigmoid function. Concretely, the output of this method is:

$$\sigma(FC(RELU(FC(g\_avg\_pool(\mathbf{X}))))) \times \mathbf{X} \qquad (1)$$

where $\mathbf{X}$ is the feature-maps. This module aims to dynamically focus on more important channels, essentially increasing the importance of specific channels over others. This is accomplished by scaling the more important channels by a higher value. While many feature descriptors exist to reduce the spatial dimensions of the feature maps to a singular value, this module uses average pooling to keep the required computation low. The next part of the module maps the scaling weights using a Multi-Layer Perceptron (MLP) with a bottleneck structure. The result values are scaled to a range of 0-1 by passing them through a sigmoid activation layer. Afterward, using a common broadcasted element-wise multiplication, the output is applied directly to the input. This multiplication scales each input feature-maps channel with its corresponding weight learned from the MLP in the Excitation module.

The following is the summary of our contributions:

- We present a novel attention module based on the Squeeze-and-Excitation method for the answer grounding task.

- We evaluate our method on common datasets and show that it achieves new state-of-the-art results.

- We compare our design with the top-performing networks.

- We perform multiple ablation studies to learn more about this network.

2. Related Work

2.1. Attention Methods for Answer Grounding

Models based on different attention methods have been actively explored for the answer grounding task.

MAC-Caps [39] is based on MAC [15] which has a recurrent reasoning architecture that performs T reasoning steps to answer the question. At each reasoning step, MAC uses an attention block to read from image features and writing memory. MAC-Caps adds capsule layers on top of the convolutional layers to obtain visual capsules from the feature-maps.

Att-MCB [32] is designed to use a Knowledge Base. It is composed of nine modules, out of which two modules use an attention mechanism. One module uses an attention mechanism to generate visual explanations, and the other one to consume and point out relevant information from the Knowledge Base.

A multi-grained attention method is introduced in [14]. It consists of two types of object-level groundings to explore fine-grained information, and a more sophisticated language model for better question representation. A Word-Label Matching attention vector that indicates the weight that should be given to each of the K objects in the image is computed in terms of the semantic similarity between the category labels of the objects and the words in the question. A word-object matching module is exploited to evaluate how likely a question word matches a visual object. A Sentence-Object attention module is used to capture the global semantics of the whole sentence to guide the focus on relevant objects.

Multiple custom-designed attention modules are used in [46] to generate an attention map for the given image and question pair. Questions are tokenized and passed through an embedding layer followed by an LSTM layer to generate the question features. An Image Attention Supervision Module is used as an auxiliary classification task; that is, the ground-truth visual grounding labels are used to guide the model to focus on significant parts of the image to answer each question.

A so-called "accumulated attention" mechanism is introduced in [5]. It consists of three attention modules, one for the input text, one for the objects, and one for the whole image. All three modules have a similar structure of a linear layer followed by a Softmax function.
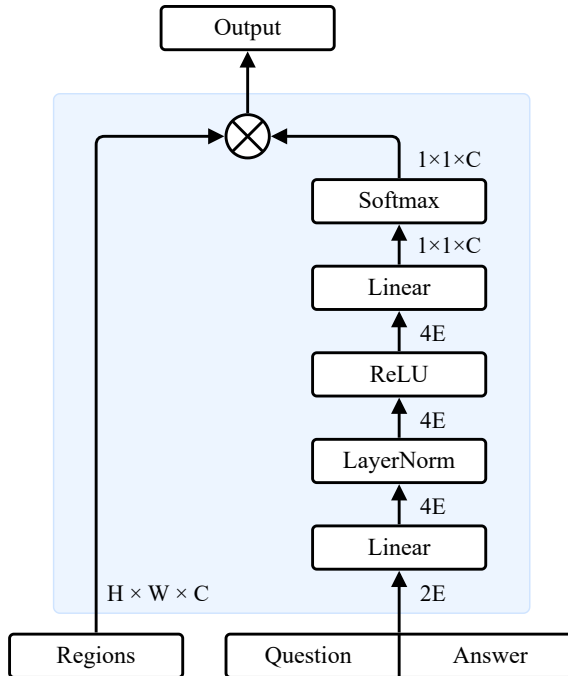
Figure 1. Our proposed sentence attention block. C denotes the number of channels in image feature-maps, and E represents the embedding size of the sentence encoder network. Question and answer embeddings are concatenated into one vector.

2.2. Multimodal Attention Methods

Since its introduction, the attention mechanism has been widely adopted in the computer vision community due to its special capabilities for many multimodal applications.

In [11], a multimodal attention model is proposed based on the encoder-decoder networks and RNNs for video captioning and sentence generation. In particular, the multimodal attention framework combined image, audio, and motion features by selecting each modality's most relevant context vector.

In [44], utilizing stacked attention networks is suggested to look for image regions that correlate with a query answer and pinpoint representative features of a given question more accurately.

More recently, in [9], a normalized variant of the self-attention mechanism, named normalized self-attention (NSA), is introduced. NSA seeks to encode and decode the image and caption features and normalize the distribution of internal activations during training.

For the video question answering task, attention and memory mechanisms are used in [6] to efficiently learn visual features and the semantic correlations that precisely answer questions.

2.3. Answer Grounding Methods

The answer grounding task has also been studied under different names, such as grounded VQA and visual explanation for VQA.

The VinVL [45] method is based on OSCAR [21] which proposes a novel vision-language pre-training that incorporates image anchor points as input, taken from an object detection network. VinVL improves the visual representations by using a bigger visual model pre-trained on larger training corpora that combine multiple annotated object detection datasets.

LXMERT [36] presents a cross-modality framework for learning the connections between vision and language. The authors build a large-scale Transformer model that consists of three encoders: an object relationship encoder, a language encoder, and a cross-modality encoder. This model is then pre-trained on a large-scale dataset of image-and-sentence pairs.

MTXNet [30] is an end-to-end trainable architecture that generates multimodal explanations focusing on the text in the image. It contains a Graph Attention Network, and each object location and OCR token is treated as a node in the graph. It also contains a multimodal transformer that operates on three modalities: question words, visual objects, and OCR tokens.

U-CAM [29] is a method that uses gradient-based uncertainty estimates to provide visual attention maps. The certainty measure is improved by adding the certainty gradients to the existing standard Cross-Entropy loss gradients for training the model during back-propagation.

3. Method

Current attention methods have one or more of the following three problems:

- The attention mechanism is not based on well-known and well-tested methods [32, 36, 29]. Well-established methods have been evaluated on many different tasks and datasets, especially, they perform well on large-scale datasets. Their design details have been actively studied, and their advantages and disadvantages are visible to the users.

- The complicated design makes it challenging to re-implement them on other platforms [32, 39, 30]. Recently, many applications have been made based on deep neural networks that are expected to be used by the end-users of different platforms. For example, answer grounding methods are helpful in assistive technologies for people with vision impairments [3]. In order to use such technologies offline, it is crucial to implement the method for each platform.
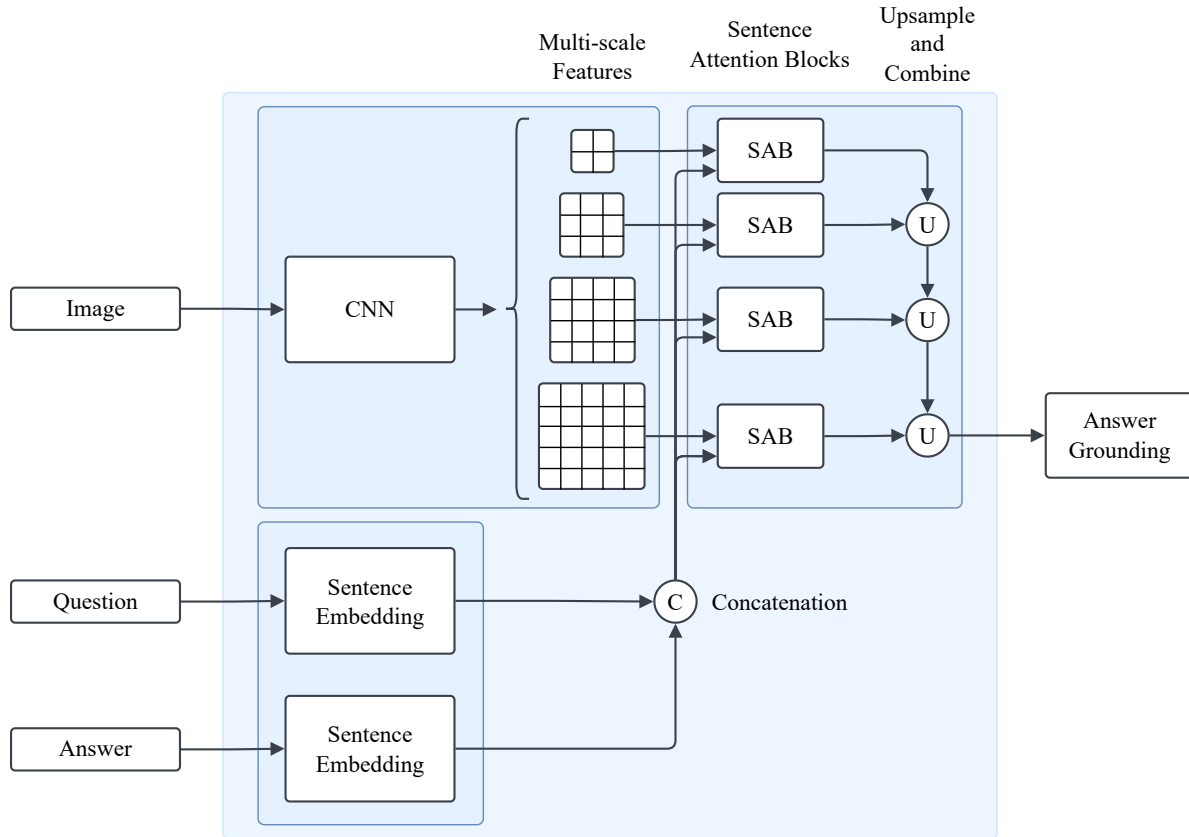
Figure 2. The complete proposed network. The multi-scale channel-wise image features contain different regions that are filtered by our Sentence Attention Block (SAB) based on the question and answer embeddings.

- The custom design does not allow the usage of pre-trained models [39, 45, 28, 27]. Answer grounding datasets are less diverse and have fewer samples than the classification datasets. Using pre-trained models can improve the accuracy and make the model more robust to noise and unseen samples.

This paper proposes an attention block that avoids the mentioned problems. We design an attention block based on well-established and well-studied methods. It has a simple design to make it easy to re-implement and uses pre-trained models to achieve maximum accuracy.

Our proposed method has three main parts: region proposal, sentence embedding, and attention fusion.

3.1. Region Proposal

This module's primary goal is to process the image and generate all the possible region candidates. Unlike Region Proposal Networks [31] which generate anchor-based bounding boxes, this module generates dense regions (segmentations).

A pre-trained classification network is used as the back-

bone. However, the final classifier is removed to get the multi-scale feature-maps. During the end-to-end training of the proposed method, the backbone network learns all the candidate regions. Later, these candidate regions are filtered and combined by the proposed Sentence Attention Blocks. In Section 5.3, we show that the backbone network works as expected by visualizing the feature-maps.

3.2. Sentence Embedding

Since the input questions and answers have variable lengths, they are processed by a sentence embedding network. A pre-trained sentence encoder is employed to encode the given question and answer to a vector space separately. The result is two vectors, one for the question and one for the answer. The two vectors are concatenated and passed to the Sentence Attention Blocks as one vector.

3.3. Attention Fusion

Given multiple potential regions for an image, the answer grounding task reduces to choosing and/or combining these regions based on the given question and/or answer. The multi-scale channel-wise image features from the back-

Method	Mean IoU
MTXNet [30]	18.9
LXMERT [36]	24.9
VinVL [45]	25.6
MAC-Caps [39]	26.1
Ours	**29.0 (+2.9)**

Table 1. Comparison results on the test set of the TextVQA-X dataset [30].

Method	Mean IoU
DeconvNet [7]	29.8
Mask Aggregation [7]	32.6
LXMERT [36]	33.3
VinVL [45]	33.9
MAC-Caps [39]	34.3
Ours	**36.8 (+2.5)**

Table 2. Comparison results on the test set of the VQS dataset [7].

bone network contain candidate regions. The goal is to filter and combine these regions based on the question and answer vector from the sentence embedding network. Towards this goal, we design an attention block based on the Squeeze-and-Excitation (SE) block [13].

We start with an SE block and make four modifications as follows:

1. The first modification is changing it from self-attention to cross-attention. In other words, unlike a self-attention block, the proposed block pays attention to an external source: the encoded question and answer.

2. The second modification is adding a normalization layer. Empirically we found that adding a normalization layer improves the accuracy. Similar to [40], we use LayerNorm since it performs better on sequences.

3. The third modification is replacing the sigmoid function with softmax. Since the goal is to force the model to choose a few of the channel-wise image features, the sigmoid function is replaced with softmax.

4. The last modification is inverting the bottleneck. Instead of shrinking the embedding space, we expand the question and answer embeddings by a factor of two.

In Section 5.1, the effect of each modification on the network accuracy is shown. Figure 1 depicts the final structure of the proposed block. Precisely, the proposed attention block is as follows:

$$Softmax(FC(ReLU(LN(FC(\mathbf{QA}))))) \times \mathbf{R}_i \quad (2)$$

where FC is a single-layer perceptron, LN is LayerNorm, $\mathbf{QA}$ is the joint question and answer embeddings, and $\mathbf{R}_i$ is the image regions at scale i. For each scale, a separate instance of this attention block is created so that each block learns special weights for its corresponding feature-map scale.

After computing the attention block for all the scales, all the outputs at different scales are combined into one output. Starting from the lowest resolution output, a 1×1 convolution is used to convert the number of channels to the number of channels of the next lowest resolution output. Then, Bilinear interpolation is used for upsampling the

output feature-map by a factor of 2. Finally, the output is added to the next lowest resolution output. This process is repeated until no other output is left. In the end, a 1×1 convolution is used to reduce the number of channels of the final output to two. Therefore, the final output has two channels representing answer grounding and its background. We use Argmax to output a binary image.

The complete design of the proposed method is depicted in Figure 2.

3.4. Comparison to Existing Methods

We compare the design of our proposed method to the top-performing existing methods for the answer grounding task.

MAC-Caps [39] is one of the best-performing networks that currently exist. This network is based on MAC [15] which employs attention blocks in a recurrent reasoning architecture. MAC-Caps adds capsule layers to process the feature-maps further. Although similar to our method, MAC-Caps employs an attention mechanism. However, the recurrent design brings the usual problems of RNNs, such as gradient vanishing/exploding, longer training time, and not performing well on long sequences. Since our proposed method does not have a recurrent design, it avoids such problems.

VinVL [45] is another top-performing network for the answer grounding task. This method is based on OS-CAR [21] which incorporates image anchor points as inputs. VinVL improves the visual representations by using a bigger visual model pre-trained on larger training corpora that combine multiple annotated object detection datasets. This method relies on object detection networks to provide the anchor points, which adds an extra task for the network to learn. Also, using large generic corpora may not improve the accuracy for special datasets such as VizWiz-VQA-Grounding [3]. In contrast, our proposed method relies only on the feature-maps and does not define another task.

Att-MCB [32] is the state-of-the-art network on the VQA-X dataset [27]. This method is designed to use a Knowledge Base. It comprises nine modules and four loss functions, making it a tedious design. This complicated design is resource-demanding and makes the training sensitive

Method	Rank Correlation
PJ-X [27]	0.342
CCM [28]	0.368
U-CAM [29]	0.372
VinVL [45]	0.373
Att-MFH [32]	0.376
MAC-Caps [39]	0.389
Att-MCB [32]	0.396
Ours	**0.421 (+0.025)**

Table 3. Comparison results on the test set of the VQA-X dataset [27].

Team	Mean IoU
MindX	66.9
MGTV	69.7
hsslab_inspur	70.1
GroundTruth	70.3
Aurora	70.6
SAB (Ours)	**72.4 (+1.8)**

Table 4. The top-performing teams on the VizWiz-VQA-Grounding Challenge 2022 leaderboard [3]. At the time of this writing, none of the methods have been published.

Method	# of params	inference time
LXMERT [36]	153M	56ms
VinVL [45]	406M	69ms
MAC-Caps [39]	80M	54ms
Ours	88M	**49ms**

Table 5. Comparison of our proposed method to existing methods regarding the number of parameters and inference time.

to hyper-parameters. Our proposed method comprises three main modules and one loss function. This minimalistic design helps the gradient flow during back-propagation, and the training is less sensitive to hyper-parameters.

4. Experiments

We train and test our proposed method on four answer grounding datasets. We use the same setup for all of the experiments.

4.1. Setup

We use a pre-trained EfficientNet [37] to extract the multi-scale image features and we use a MiniLMv2 [42] sentence encoder network to encode the questions. In Section 5.2, we perform an ablation study on different backbones. we use BEiT [41] to generate answers for the datasets where the answers are not publicly available. We use RMI [47] with its default settings along with cross-entropy. Therefore, the overall loss function to minimize is:

$$\mathcal{L}_{all}(y,p) = \frac{1}{B}\sum_{b=1}^{B}\lambda\mathcal{L}_{ce}(y^{(b)},p^{(b)}) \\ + (1-\lambda)\mathcal{L}_{rmi}(y^{(b)},p^{(b)}) \quad (3)$$

where $\lambda \in [0,1]$ is a weight factor, B denotes the number of samples in a mini-batch, $\mathcal{L}_{ce}(y^{(b)},p^{(b)})$ is the standard cross entropy loss between the b-th sample and its corresponding prediction, respectively, and $\mathcal{L}_{rmi}(y^{(b)},p^{(b)})$ is the RMI loss as follows:

$$\mathcal{L}_{rmi}(\mathbf{Y},\mathbf{P}) = \sum_{c=1}^{C}\frac{1}{2d}Trace(log(\mathbf{M})) \quad (4)$$

and,

$$\mathbf{M} = \mathbf{\Sigma_Y} - Cov(\mathbf{Y},\mathbf{P})(\mathbf{\Sigma_P^{-1}})^T Cov(\mathbf{Y},\mathbf{P})^T \quad (5)$$

where $\mathbf{Y} = y^{(b)}$, $\mathbf{P} = p^{(b)}$, $\mathbf{M} \in \mathbb{R}^{d\times d}$, d denotes the number of pixels, C is the number of object classes, here $C = 2$ for the answer grounding and its background, $\mathbf{\Sigma_Y}$ is the variance matrix of Y, and $Cov(\mathbf{Y},\mathbf{P})$ is the covariance matrix of $\mathbf{Y}$ and $\mathbf{P}$. Similar to [47], we set $\lambda = 0.5$.

We use AdamW optimizer [25] with a weight decay of 0.05 and batch size of 16. We apply the "polynomial" learning rate policy with a poly exponent of 0.9 and an initial learning rate of 0.0001. Synchronized batch normalization is used across multiple GPUs. We use RandAugment [4] for data augmentation.

4.2. TextVQA-X

The TextVQA-X dataset [30] is a subset of the TextVQA dataset [33] such that the answer groundings are generated through a manual segmentation annotation process. TextVQA-X consists of 10,379 training images and 3,354 test images. Since only a few models have been evaluated on this dataset, we train three more models for a fair comparison. Similar to [3], we train top-performing methods that their code is publicly available; specifically, LXMERT [36], VinVL [45], and MAC-Caps [39]. We use the same setup as in [3] to train the three models with the difference that we train the models on the TextVQA-X training set. Similar to the existing methods, we use the test set to evaluate our method and report the mean IoU measure.

4.3. VQS

The VQS dataset [7] builds upon the images, instance segmentation masks, and bounding boxes in COCO [22] and the questions and answers in the VQA dataset [1]. VQS contains 26,995 training, 5,000 validation, and 5,873 test images. Similar to Section 4.2, we train top-performing methods that their code is publicly available; specifically,

Backbone	Mean IoU
ResNet-50x3 [19]	71.9
Swin Transformer V2-B [24]	72.8
EfficientNetV2-L [38]	73.1
EfficientNet-B7 [37]	**73.5**
RoBERTa [23]	72.8
MPNet [34]	73.1
MiniLMv2 [42]	**73.5**

Table 6. The ablation study of the pre-trained backbone networks. **Top Section:** Using a MiniLMv2 as a fixed sentence embedding network and replacing the image feature extraction networks. **Bottom Section:** Using a fixed image feature extraction network (EfficientNet-B7) and replacing the sentence embedding networks.

LXMERT [36], VinVL [45], and MAC-Caps [39]. We use the same setup as in [3] to train the three models with the difference that we train the models on the VQS training set. Similar to the existing methods, we use the test set to evaluate our method and report the mean IoU measure.

4.4. VQA-X

The VQA-X dataset [27] does not have visual annotations for its training set. It has visual annotations only for its validation and test sets. Therefore, we train our method on the training set of the VQS dataset [7], which has related images to the VQA-X validation and test sets. The images of both datasets are from the COCO dataset [22]. The other methods are trained on different datasets, including Visual Genome [20] and VQA v2.0 [8]. The VQA-X dataset consists of 24,876 training, 1,431 validation, and 1,921 test images. Similar to the other methods, we use the test set of the VQA-X dataset to evaluate our method and report the Rank Correlation as in [27].

4.5. VizWiz-VQA-Grounding

The VizWiz-VQA-Grounding dataset [3] is based on the VizWiz-VQA dataset [10], and the images and questions come from visually impaired people who shared them to request visual assistance in their day-to-day lives. VizWiz-VQA-Grounding contains a total of 9,998 VQAs divided into 6,494/1,131/2,373 VQAs for training, validation, and testing. The 2022 VizWiz-VQA-Grounding Challenge is designed around the aforementioned VizWiz-VQA-Grounding dataset. This challenge was held recently, and at the time of this writing, current methods in this challenge have not been published yet. Therefore, we evaluate our method using the VizWiz-VQA-Grounding Challenge online leaderboard [3]. The leaderboard evaluates all the methods against the VizWiz-VQA-Grounding test set. The ground truth labels for the test set are not published yet.

Modification	Mean IoU
Switching to cross-attention	70.1
Adding normalization layer	71.2 (+1.1)
Switching to Softmax	72.7 (+1.5)
Expansion of the embeddings	**73.5 (+0.8)**

Table 7. The results of our ablation study of the design modifications. The modifications are done cumulatively from top to bottom.

4.6. Results

We evaluate our proposed method on four answer grounding datasets. Tables 1, 2, and 3 compare the results of our proposed method with the existing methods on the TextVQA-X, VQS, and VQA-X datasets respectively. Our method achieves new state-of-the-art accuracy on three datasets. More specifically, improvements of 2.9, 2.5, and 0.025 on the TextVQA-X, VQS, and VQA-X datasets, respectively.

Table 4 lists the top-performing teams on the VizWiz-VQA-Grounding Challenge 2022 leaderboard [3]. At the time of this writing, our method holds first place with considerably higher accuracy than the other methods.

Furthermore, Table 5 compares our proposed method to existing methods regarding the number of parameters and inference time. The efficient design of our method makes it the fastest method.

5. Ablation Studies

In this section, we perform four ablation studies. All of the experiments are done on the VizWiz-VQA-Grounding [3] validation set and using the setup from Section 4.1 unless otherwise stated.

5.1. Design Modifications

We started the design of our sentence attention block from the standard SE block and made four modifications as explained in Section 3.3. In this section, we perform an ablation study on these modifications. Table 7 shows the impact of each modification on the final accuracy. These modifications are done cumulatively. The first modification is changing from self-attention to cross-attention, which makes our baseline for this ablation study. The second modification is adding a normalization layer. This modification improves our baseline by 1.1 percent. The third modification is replacing the Sigmoid function with Softmax. This modification has the highest impact on our method, 1.5 percent. And the fourth modification is the expansion of the embeddings. This last modification brings an addition of 0.8 percent.

5.2. Backbone Networks

The flexibility of our method enables us to use a variety of pre-trained networks. In this ablation study, we compare and analyze different pre-trained networks. To compare the image backbone networks, we keep the sentence embedding network fixed and change the image backbone network. Similarly, to compare the sentence embedding networks, we keep the image backbone network with the highest accuracy and change the sentence embedding network. We use well-known and well-performing networks whose pre-trained weights are publicly available. Specifically, we use EfficientNet [37], EfficientNetV2 [38], Swin Transformer V2 [24], and ResNet-50x3 [19] for image feature extraction and MiniLMv2 [42], MPNet [34], and RoBERTa [23] for sentence embedding. Table 6 shows the results of this ablation study. The highest accuracy is achieved by using EfficientNet-B7 and MiniLMv2. Since the EfficientNetV2-L reduces the input resolution more aggressively than EfficientNet-B7, it performs worse. Transformer networks usually perform well on large datasets. However, since VizWiz-VQA-Grounding is not a large dataset, Swin Transformer V2 does not achieve the highest accuracy. The ResNet-50x3 is pre-trained on ImageNet-21k, but it achieves the least accuracy in this experiment.

5.3. Region Proposals

Feature-map visualization will provide insight into the internal representations of each layer in a network for a specific input image.

This ablation study aims to show that the image backbone network has learned the potential regions for an image, which are then processed by the proposed sentence attention block. To this aim, we visualize the final output feature-maps of the image backbone network, before feeding them to an attention block. We use the image backbone of our fully trained network on the VizWiz dataset. Figure 3 (a) shows a few manually-chosen feature-maps for a sample image. These feature-maps show that after training, the backbone network has been able to learn the potential important regions of images. Having all the candidate regions, the task of the proposed attention block is to filter the irrelevant regions to the question and answer. Figure 3 (b) shows the final output of the complete network for the same image in (a), using different questions and answers (for brevity, the answers are not shown). This figure shows that our proposed block is able to filter the channel-wise feature-maps based on the input questions and answers.

5.4. Multi-scale Fusions

There is a trade-off in the answer grounding task such that some images are best handled at lower inference resolution and others better handled at higher inference resolution. Fine details, such as the edges of objects or thin
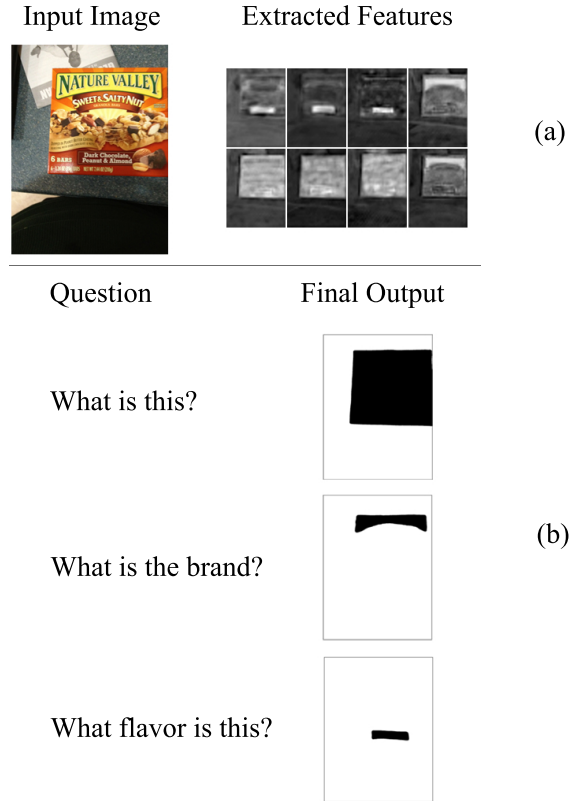

Figure 3. **(a)** Eight manually-selected sample feature-maps for an image. These feature-maps are the output of the feature extraction backbone network, before feeding them to the sentence attention block. The network has learned to output potential regions in different channels. **(b)** The final output of our complete network for the same image in (a), given different questions and answers (for brevity, the answers are not shown). This figure shows that our proposed attention block can filter the channel-wise feature-maps based on the input questions and answers.

Scale	Mean IoU
1/32	71.4
1/32 + 1/16	72.5 (+1.1)
1/32 + 1/16 + 1/8	73.1 (+0.6)
1/32 + 1/16 + 1/8 + 1/4	**73.5 (+0.4)**

Table 8. The gradual effect of multi-scale fusion. From top to bottom, the first row uses only the lowest resolution feature-map, and each row adds higher resolution feature-maps. Each scale is the ratio of the feature-map size to the input image.

structures, are often better predicted with scaled-up images. At the same time, the prediction of large structures, which requires more global context, is often done better at scaled-down images because the network's receptive field can observe more of the necessary context.

This ablation study shows the impact of adding the extra scales on the accuracy. Table 8 shows the results of this

study. From top row to bottom, we start with only one scale and one attention block and gradually add higher resolution feature-maps. The lowest resolution is the final output feature-map of the backbone network.

6. Conclusions

In the study, a new architectural block termed the Sentence Attention Block was introduced to address certain challenges. This block recalibrates channel-wise image feature-maps by modeling inter-dependencies between image feature-maps and sentence embedding. The design began with a recognized attention method, and with minor adjustments, enhanced results were achieved, reaching top-tier accuracy. We showed the block's ability to filter out irrelevant feature-map channels based on sentence embedding. The method's adaptability allows for various pre-trained backbone networks, and its straightforward nature facilitates comprehension and re-implementation. The method's efficacy was demonstrated on the TextVQA-X, VQS, VQA-X, and VizWiz-VQA-Grounding datasets, and several ablation studies highlighted the validity of the design decisions.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 38(8):2939–2970, 2022.

[3] Chongyan Chen, Samreen Anjum, and Danna Gurari. Grounding answers for visual questions asked by visually impaired people. *arXiv preprint arXiv:2202.01993*, 2022.

[4] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[5] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual grounding via accumulated attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7746–7755, 2018.

[6] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1999–2007, 2019.

[7] Chuang Gan, Yandong Li, Haoxiang Li, Chen Sun, and Boqing Gong. Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused

[8] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[9] Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. Normalized and geometry-aware self-attention network for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10327–10336, 2020.

[10] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018.

[11] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE international conference on computer vision*, pages 4193–4202, 2017.

[12] Parisa Hosseini, Seyedalireza Khoshsirat, Mohammad Jalayer, Subasish Das, and Huaguo Zhou. Application of text mining techniques to identify actual wrong-way driving (wwd) crashes in police reports. *International Journal of Transportation Science and Technology*, 2022.

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[14] Pingping Huang, Jianhui Huang, Yuqing Guo, Min Qiao, and Yong Zhu. Multi-grained attention with object-level grounding for visual question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3595–3600, 2019.

[15] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*, 2018.

[16] Seyedalireza Khoshsirat and Chandra Kambhamettu. Semantic segmentation using neural ordinary differential equations. *Advances in Visual Computing: 17th International Symposium, ISVC 2022, San Diego, CA, USA, October 3–5, 2022, Proceedings, Part I*, pages 284–295, 2022.

[17] Seyedalireza Khoshsirat and Chandra Kambhamettu. Embedding attention blocks for the vizwiz answer grounding challenge. *VizWiz Grand Challenge Workshop*, 2023.

[18] Seyedalireza Khoshsirat and Chandra Kambhamettu. Empowering visually impaired individuals: A novel use of apple live photos and android motion photos. *25th Irish Machine Vision and Image Processing Conference*, 2023.

[19] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *European conference on computer vision*, pages 491–507. Springer, 2020.

[20] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[21] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[24] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. *arXiv preprint arXiv:2111.09883*, 2021.

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[26] Elham Maserat, Reza Safdari, Hamid Asadzadeh Aghdaei, Alireza Khoshsirat, and Mohammad Reza Zali. 43: Designing evidence based risk assessment system for cancer screening as an applicable approach for the estimating of treatment roadmap. *BMJ Open*, 7(Suppl 1):bmjopen–2016, 2017.

[27] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8779–8788, 2018.

[28] Badri Patro, Shivansh Patel, and Vinay Namboodiri. Robust explanations for visual question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1577–1586, 2020.

[29] Badri N Patro, Mayank Lunayach, Shivansh Patel, and Vinay P Namboodiri. U-cam: Visual explanation using uncertainty based class activation maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7444–7453, 2019.

[30] Varun Nagaraj Rao, Xingjian Zhen, Karen Hovsepian, and Mingwei Shen. A first look: Towards explainable textvqa models via visual and textual explanations. *arXiv preprint arXiv:2105.02626*, 2021.

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[32] Felipe Riquelme, Alfredo De Goyeneche, Yundong Zhang, Juan Carlos Niebles, and Alvaro Soto. Explaining vqa predictions using visual grounding and a knowledge base. *Image and Vision Computing*, 101:103968, 2020.

[33] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.

[34] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020.

[35] Jabeen Summaira, Xi Li, Amin Muhammad Shoib, Songyuan Li, and Jabbar Abdul. Recent advances and trends in multimodal deep learning: A review. *arXiv preprint arXiv:2105.11087*, 2021.

[36] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[37] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[38] Mingxing Tan and Quoc V Le. Efficientnetv2: Smaller models and faster training. *arXiv preprint arXiv:2104.00298*, 2021.

[39] Aisha Urooj, Hilde Kuehne, Kevin Duarte, Chuang Gan, Niels Lobo, and Mubarak Shah. Found a reason for me? weakly-supervised grounded visual question answering using capsules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8465–8474, 2021.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[41] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

[42] Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. *arXiv preprint arXiv:2012.15828*, 2020.

[43] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10941–10950, 2020.

[44] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016.

[45] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, pages 5579–5588, 2021.

[46] Yundong Zhang, Juan Carlos Niebles, and Alvaro Soto. Interpretable visual question answering by visual grounding from attention supervision mining. In *2019 ieee winter conference on applications of computer vision (wacv)*, pages 349–357. IEEE, 2019.

[47] Shuai Zhao, Yang Wang, Zheng Yang, and Deng Cai. Region mutual information loss for semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019.