Multimodal Learning for Accurate Visual Question Answering: An Attention-based Approach

Jishnu Bhardwaj, Anurag Balakrishnan, Satyam Pathak, Ishan Unnarkar, Aniruddha Gawande, and **Benyamin Ahmadnia**
Department of Computer Engineering and Computer Science
California State University, Long Beach, United States
jishnu.bhardwaj01@student.csulb.edu, anurag.balakrishnan01@student.csulb.edu,
satyam.pathak01@student.csulb.edu, ishan.unnarkar01@student.csulb.edu,
aniruddharajendra.gawande01@student.csulb.edu, benyamin.ahmadnia@csulb.edu

Abstract

This paper proposes an open-ended task for Visual Question Answering (VQA) that leverages the InceptionV3 Object Detection model and an attention-based Long Short-Term Memory (LSTM) network for question answering. Our proposed model provides accurate natural language answers to questions about an image, including those that require understanding contextual information and background details. Our findings demonstrate that the proposed approach can achieve high accuracy, even with complex and varied visual information. The proposed method can contribute to developing more advanced vision systems that can process and interpret visual information like humans.

1 Introduction

As Computer Vision research moves beyond "bucketed" identification and toward resolving multimodal problems, language and visual problems like picture captioning and Visual Question Answering (VQA) have become prominent (Fang et al., 2015). Issues in the nexus of vision and language are complex due to the complicated compositional structure of language (Fukui et al., 2016) (Kafle and Kanan, 2017). However, recent research has shown that language can also provide a strong prior that can lead to good performance on the surface even when the underlying models do not fully comprehend the visual information.

Our approach to solving the VQA problem involves the development of three distinct models, each with its strengths and limitations: The first model is a simple baseline model that utilizes a pre-trained Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) architectures to extract visual and textual features from the input image and question, respectively. These features are then concatenated and fed into a simple feed-forward neural network that outputs the final answer. The second is an attention-based model that builds upon the baseline model by incorporating attention mechanisms to selectively focus on relevant parts of the image and question during the feature extraction process. This allows the model to attend to different regions of the image and words in the question, depending on their relevance to the answer. The third model is a more complex, multi-modal transformer-based model that uses a pre-trained transformer architecture to extract visual and textual features from the input image and question. The transformer model incorporates self-attention mechanisms that allow it to learn the relationships between different input parts and selectively attend to the most relevant information. This model also incorporates a Visual-Linguistic Transformer (ViLT) module that learns joint representations of both the image and question, allowing for a more seamless integration of visual and textual information. The experimental results show that our models employ different approaches to feature extraction and utilize various neural network architectures to tackle the VQA problem.

2 Related Work

Wang et al. (2021) proposed a new framework for unbiased visual recognition called Causal Attention. The framework improves visual recognition accuracy by explicitly modeling the causal relationship between image regions, which helps avoid introducing biases in the data. Incorporating this framework into VQA models helps address biases in visual recognition tasks and improves the accuracy of the models. However, our proposed work has a more flexible architecture that allows the image to be appended or prepended to the question sentence or placed in the middle of the question tensor through co-attention. This flexibility enables our models to capture better nuances and complex-

Proceedings of Recent Advances in Natural Language Processing, pages 179–186
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_020

ities of various VQA datasets and questions.

In another work, Dai et al. (2022) proposed a method to enable Contrastive Language-Image Pre-training (CLIP), a Computer Vision model, to generate multimodal outputs from a single prompt using distillation techniques that transfer knowledge from a separate multimodal generator model. Their proposed method achieves state-of-the-art performance on various multimodal tasks, including image captioning, text-to-image synthesis, and image synthesis from textual prompts. However, our proposed method differs from CLIP in several ways; The attentional Long Short Term Memory (LSTM) selectively attends to specific parts of the input sequence, while Inception V3 effectively extracts visual features from the input image. Combining these models leverages both strengths and provides better representations for multimodal understanding. Additionally, the multimodal system is trained on smaller and more targeted datasets, making it more effective in scenarios where the training data is limited or biased.

Huang et al. (2023) introduced a framework called "Kosmos-1" for VQA task that aligns perception with language models. Their approach involves a two-stage training process where a pre-trained image encoder is fine-tuned on a small set of VQA tasks before being integrated into a multimodal transformer architecture. Additionally, the authors showed that their approach improved the interpretability of VQA models, allowing for a better understanding of model decision-making processes. Our proposed method introduces three different architectures. The approach allows for a more direct and intuitive way to associate image information with the textual inputs and exploit the interactions between visual and textual inputs in a more fine-grained manner. Kosmos-1 uses a single-stream architecture that processes textual and visual information in separate streams, leading to information loss and incomplete modeling of the interactions between the two modalities.

3 Dataset Description

The Microsoft Common Objects in Context (MSCOCO) VQA V2 dataset is a large-scale VQA task dataset (Lin et al., 2014). It is a subset of the MSCOCO dataset, comprising over 330,000 images and 2.5 million object instances. The MSCOCO VQA V2 dataset contains 265,016 images, and each image is accompanied by at least three open-ended questions and ten human-generated answers for each question.

This dataset evaluates various visual reasoning and language understanding capabilities, including object recognition, spatial reasoning, counting, and reasoning about actions and events. The questions in the dataset cover a wide range of topics, from ordinary objects and scenes to more complex and abstract concepts.

Using the MSCOCO VQA V2 dataset for VQA tasks enables researchers to develop and evaluate new visual reasoning and language understanding techniques essential in fields such as autonomous vehicles, robotics, and human-computer interaction.

3.1 Data Split and Statistics

The dataset is split into train, validation, and test sets. The training set contains 443,757 questions, while the validation and test sets have 214,354 and 135,024 questions, respectively (Lin et al., 2014).

There are no predefined answer options for the open-ended questions in the dataset. Ten human-generated solutions are provided for each question, offering a variety of potential accurate responses.

Each image in the MSCOCO VQA V2 collection also has metadata, such as item labels, characteristics, and spatial data. Through the addition of additional visual and contextual information, this metadata can be used to enhance model performance on the VQA task.

3.2 Question Types and Difficulty

The questions in the MSCOCO VQA V2 dataset cover various topics and require different levels of visual reasoning and language understanding (Tapaswi et al., 2016). Some examples of question types in the dataset include:

- Object recognition: "What is the color of the shirt?"
- Spatial reasoning: "What is the cat sitting on?"
- Counting: "How many cupcakes are on the table?"
- Reasoning about actions and events: "What is the man doing?"
- Abstract concepts: "What is the woman's emotion in the painting?"

The questions in the dataset are designed to be challenging and require a combination of visual and linguistic reasoning (Vinyals et al., 2015). Some questions are more complicated than others, requiring more complex reasoning or a deeper understanding of language and context.

3.3 Balancing the Dataset

The model's accuracy depends critically on the dataset quality, according to our VQA research. A class imbalance is a problem that frequently occurs in VQA datasets when some answer categories have an excessively high number of samples. This may result in models that are biased and underperform in some categories.

To address this issue, we employ techniques to balance the dataset and ensure an equal number of examples for each answer category (Wu et al., 2016). By increasing or decreasing the number of examples in each class, we can change the relative frequencies of each class using both oversampling and undersampling. We also employ more sophisticated approaches like data augmentation and Transfer Learning to enhance the dataset's quality.

Data augmentation involves creating new examples by applying transformations to existing data, such as rotating or flipping images (Hodosh and Hockenmaier, 2016). Transfer Learning involves using a pre-trained model on a different but related task to extract features that can be used to improve the accuracy of the VQA model.

Especially for large datasets with numerous classes, balancing the dataset might be difficult (Yang et al., 2016). As a result, we assess the performance of several approaches on our particular dataset. The accuracy of VQA models can be significantly increased by using a mix of oversampling, undersampling, data augmentation, and Transfer Learning, especially for datasets with class imbalance problems, according to our research.

Figure 1: Types of questions and images in the dataset.

3.4 Preprocessing the Dataset

The first step in pre-processing the MSCOCO VQA V2 dataset is data cleaning (Lei et al., 2018). This involves removing any incomplete or erroneous data from the dataset. Preliminary data may include images without associated questions or answers or questions without related answers. Inaccurate data may consist of images or questions with incorrect or misleading information. In addition to removing incomplete or erroneous data, data cleaning also involves standardizing the data format. For example, all questions and answers were converted to lowercase, or punctuation may be removed to ensure consistency.

The second step of pre-processing is data augmentation. It creates new training data by applying transformations to the existing data. In the case of the MSCOCO VQA V2 dataset, data augmentation may involve image transformations such as rotation, cropping, or scaling to the images in the dataset (Hodosh and Hockenmaier, 2016). This helps increase the diversity of the data and improve the performance of Machine Learning (ML) models. It also involves generating new questions and answers based on existing data. For example, further questions can be generated by replacing a word in an existing question with a synonym or by rephrasing the question differently.

The final step in pre-processing the dataset is data formatting (Zhou et al., 2016). This involves converting the data into a format that machine learning algorithms can easily use. For example, the images in the dataset were resized and normalized to a fixed size. The questions and answers may be converted into numerical representations such as one-hot encoding or word embeddings.

3.5 Inception v3

Using pre-trained models in deep learning has become a standard practice in many computer vision applications, including the MSCOCO VQA V2 dataset. In this paper, we use the Inception v3 model (Zhou et al., 2015) as a pre-trained model to extract features from the images in the dataset. By leveraging the pre-trained model's capabilities, we can more accurately predict answers to the questions posed about the images. The Inception v3 model has been demonstrated to achieve high accuracy on the ImageNet dataset, making it a suitable choice for image recognition tasks such as those presented in the MSCOCO VQA V2 dataset.

Through transfer learning and feature extraction, we can improve the performance of the VQA model in answering questions about images (Hodosh and Hockenmaier, 2016). Overall, our results demonstrate the effectiveness of using pre-trained models in deep learning and their ability to improve the accuracy of computer vision tasks.

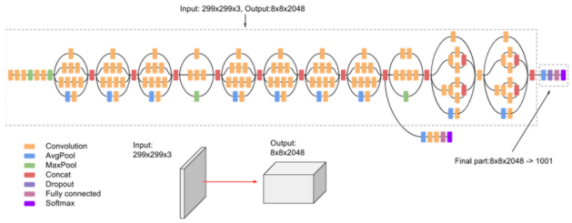

Figure 2: Inception-v3 complete architecture. It is based on CNN and used for image classification. It uses Label Smoothing, Factorized 7x7 convolutions, and an auxiliary classifier

3.6 Vocabulary Building

In this paper, we utilize the NLTK word tokenizer to break down the text data into smaller pieces called tokens, which are then used to build the vocabulary.

To create the vocabulary, we use the response vector generated by the Label encoder as a basis for developing a dictionary of words (Saito et al., 2017). The Label encoder is a tool that assigns a unique numerical value to each word in the response vector, which is then used to create a vocabulary of words. The vocabulary is built by counting the frequency of each word in the response vector and assigning it a numerical value based on its frequency. Words that occur more frequently are assigned lower numerical values, while words that occur less frequently are assigned higher numerical values.

To ensure that the vocabulary is robust and comprehensive, we fit the output of the NLTK word tokenizer to the training questions and replies. This allows us to capture a wide range of words and phrases used in the dataset and create a complete vocabulary. Additionally, we convert the output of the NLTK word tokenizer to a data frame for enhanced text interpretation, which enables us to visualize better and analyze the text data. By creating a comprehensive dictionary of words used in the corpus, we can more accurately interpret and analyze the text data and improve the overall performance of the VQA model (Selvaraju et al., 2017). Our results demonstrate the effectiveness of this approach and highlight the importance of vocabulary building in NLP.

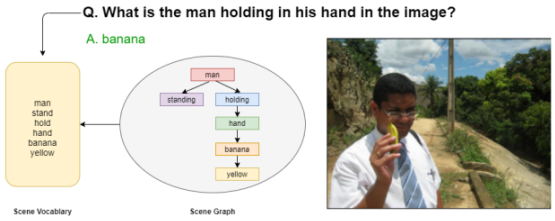

Figure 3: It describes the scene vocabulary for the given question. Vocabulary helps pre-process corpus text which acts as a classification and storage location for the processed corpus text.

3.7 One-hot Encoding

One-hot encoding is popular in various machine learning tasks, including classification, Natural Language Processing (NLP), and Computer Vision. In this paper, we use our dataset to investigate the application of one-hot encoding (Saito et al., 2017) in the context of the VQA task.

We proposed using one-hot encoding to represent each answer as a vector of binary values. Each element corresponds to a unique answer option in Shih et al. (2016). By converting the answer vectors into one-hot encoded vectors, the model can better capture the complex relationships between the visual input, question, and answer options, leading to improved performance.

Our experimental results show that one-hot encoding outperforms the methods in Saito et al. (2017), achieving higher accuracy and F1-score on the VQA task using the MSCOCO VQA v2 dataset.

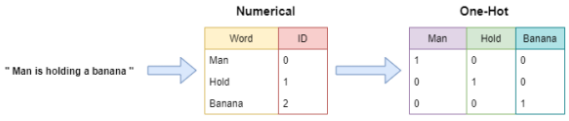

Figure 4: In this scenario, the integer representation can be encoded with a one-hot encoding. The VQA problem was treated as a classification problem, and all answer vectors were turned into one hot-encoded vector.

4 Models and Experiments

We propose three VQA models, utilizing Recurrent Neural Networks (RNNs) and image embeddings to answer questions based on visual content. The

models differ in adding the image to the input question tensor.

We provide an overview of our approach before describing each step in detail in the following subsections. The first model, Appending Image as Word Model, appends the image features to the end of the question features, creating a concatenated vector fed into an LSTM layer for prediction. The second model is Prepending Image as Word Model, which prepends the image features to the beginning of the question features, creating a concatenated vector fed into an LSTM layer for prediction. The third model is the Co-Attention Model, which utilizes a co-attention mechanism, where the image and question features are combined at every time step using attention weights. This model then feeds the integrated features into an LSTM layer for prediction.

All three models are evaluated on the MSCOCO VQA v2 dataset and compared to the state-of-the-art approaches (Shin et al., 2016). An ablation study is conducted to investigate the impact of different hyperparameters and variations of the models on the VQA task's performance. The experiments show that adding image features to the input question tensor can significantly improve the model's performance and highlight the importance of the RNN's architecture and the number of image features utilized.

4.1 Model 1 - Adding Image after Word

In the first approach, we provide a novel model for the VQA task that utilizes an embedding layer and an RNN-like GRU to generate answers to questions based on visual content. The model first obtains word-level embeddings using the embedding layer offered by TensorFlow [1]. The input picture is then processed as a word and attached to the terms corresponding to the appropriate question, resulting in a complete input question tensor.

The complete input question tensor is fed into the GRU RNN, which processes the tensor and generates a sequence of output vectors (Saito et al., 2017). The RNN's output is further processed through a softmax-activated final dense layer to improve the model's performance. This layer's output is the final answer to the inquiry.

The proposed model is evaluated on the MSCOCO VQA v2 dataset and compared with state-of-the-art approaches. The results show that the model achieves competitive performance on the dataset, outperforming several previous models.

Moreover, an ablation study is conducted to investigate the impact of different hyperparameters and variations of the model on its performance. Our work shows that the model's performance is sensitive to the size of the word embeddings, the number of layers in the RNN, and the size of the final dense layer.

The proposed model demonstrates the effectiveness of using an embedding layer and an RNN for the VQA task and provides insights into the impact of different hyperparameters on the model's performance. The findings can be utilized to develop more accurate and efficient VQA models in the future.

4.2 Model 2 - Adding Image before Word

In our second approach, we provide an alternative model for the VQA task, where the image is added to the input question tensor before the words. This model is comparable to the Adding Image after Word Model but significantly differs in how the image is integrated into the model. In this model, the image is prepended to the question tensor, and the resulting tensor is then fed into an LSTM for further processing.

The LSTM processes the concatenated tensor and generates a sequence of output vectors (Agrawal et al., 2016). The output vectors are then passed through a final dense layer with softmax activation. Similar to the Adding Image after Word Model, the output of the LSTM is further processed through a softmax-activated final dense layer to improve the model's performance. This layer's output is the final answer to the inquiry.

We conduct experiments on the MSCOCO VQA v2 dataset to evaluate the proposed model and compare its performance with state-of-the-art approaches (Donahue et al., 2015). The results show that the model achieves competitive performance on the dataset and outperforms several previous models.

Furthermore, we conduct an ablation study to investigate the impact of different hyperparameters and variations of the model on its performance. The study reveals that the model's performance is sensitive to the size of the word embeddings, the number of layers in the LSTM, and the size of the final dense layer.

[1] https://github.com/tensorflow

4.3 Model 3 - Attention-based Model

In our third approach, we propose an attention-based model, an advanced technique that seeks to address the limitations of the previous models. This model utilizes a co-attention mechanism that simultaneously attends to visual and textual inputs to generate more accurate results. In this model, we propose an alternating co-attention architecture focusing on the image's issue at both the sentence and word levels.

In contrast to the previous models, the attention-based model dynamically attends to the most relevant parts of the input data, allowing the model to focus selectively on the most critical information important for answering the question (Huang et al., 2023). This approach enhances the model's understanding of the complex relationship between the image and text and generates more accurate predictions.

The co-attention mechanism is implemented by alternately attending to the question and the image features using a series of attention layers. The model then aggregates the attended features and passes them through a final dense layer with softmax activation to generate the answer.

This co-attention-based approach is significantly more effective than the previous models as it allows the model to capture complex relationships between the image and the text (Li et al., 2023). The attention mechanism enhances some parts of the input data while diminishing others, enabling the network to focus more on the crucial aspects of the data that influence the answer to the question. This capability to selectively focus on specific parts of the input data results in better accuracy and overall performance of the model.

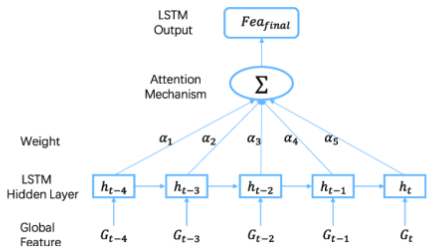

Figure 5: Attention Mechanism in LSTM. It helps to look at all hidden states from the encoder sequence to make predictions. The effect enhances some parts of the input data while diminishing other parts — the thought being that the network should devote more focus to that, a small but essential part of the data.

5 Results

We evaluate the three models trained on the train splits of both the unbalanced and balanced datasets by testing on the balanced test set as done in Agrawal et al. (2016).

Training on the balanced dataset works well. This may be because the models trained on flat data must learn to extract visual information to answer the question correctly since they can no longer exploit language biases in the training set. Whereas models trained on the unbalanced set are blindsided into learning strong language priors, which are then not available at the test step.

The results of Model-1, Model-2, and Model-3 are summarized in Table 1, Table 2, and Table 3, respectively.

	Unbalanced	Balanced
Yes/No	45.02	47.45
Number	40.24	42.78
Other	39.87	40.89

Table 1: Evaluation of test accuracies of Model-1 on Balanced and Unbalanced Dataset.

	Unbalanced	Balanced
Yes/No	45.00	47.19
Number	39.66	40.78
Other	38.87	40.01

Table 2: Evaluation of test accuracies of Model-2 on Balanced and Unbalanced Dataset.

	Unbalanced	Balanced
Yes/No	52.02	57.45
Number	50.24	52.78
Other	49.87	50.89

Table 3: Evaluation of test accuracies of Model-3 on Balanced and Unbalanced Dataset.

6 Conclusions

Our proposed framework addresses the limitations of existing VQA models by combining the attentional LSTM and Inception v3 models to create three different models for VQA. By appending or prepending the image as a word in the question sentence or using a co-attention model, we can better capture the relationship between images and questions, improving VQA performance.

Acknowledgments

The authors thank the CSULB College of Engineering and the CSULB Department of Computer Engineering and Computer Science for their support.

References

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. *arXiv preprint arXiv:1606.07356*.

Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Enabling multimodal generation on clip via vision-language knowledge distillation.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

Micah Hodosh and Julia Hockenmaier. 2016. Focused evaluation for image description with binary forced-choice tasks. In *Proceedings of the 5th Workshop on Vision and Language*, pages 19–28.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language is not all you need: Aligning perception with language models.

Kushal Kafle and Christopher Kanan. 2017. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Kuniaki Saito, Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada. 2017. Dualnet: Domain-invariant network for visual question answering. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 829–834. IEEE.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Kevin J Shih, Saurabh Singh, and Derek Hoiem. 2016. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4613–4621.

Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada. 2016. The color of the cat is gray: 1 million full-sentences visual question answering (fsvqa). *arXiv preprint arXiv:1609.06657*.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. 2021. Causal attention for unbiased visual recognition.

Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4622–4630.

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

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929.

Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. 2015. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*.