Visual Question Answering

Andrew Rausch, Chittesh Thavamani, Karl Xiao
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
{arausch, cthavama, kzx}@andrew.cmu.edu

Abstract

Inspired by recent successes in attention-based approaches to tackling Visual Question Answering (VQA), we implement a variety of modifications to a strong baseline by Kazemi and Elqursh [7] to better investigate the challenges inherent to the VizWiz VQA 2.0 dataset. We investigate hyperparameter tuning, the use of image augmentation techniques to expand the training dataset, the addition of specific features to address certain common types of visual questions, and the development of a novel attention model. The results of our experiments demonstrate that image augmentation and our novel attention module provide significant improvement, with our best-performing model improving accuracy by 3.9% relative to the strong baseline.

1 Introduction

In recent years, deep convolutional neural networks (CNNs) have shown great success in object classification and detection, both classic tasks in the computer vision and machine learning fields. However, attaining a more human-level understanding of images, which includes identifying object state and well as relationships between objects, has proven more elusive. The Visual-Question Answering (VQA) task, which seeks to accurately answer a question about an image, addresses the goal of greater understanding by learning the higher-level semantics of objects in a scene.

Besides its theoretical relevance, VQA also presents important practical applications for the visually impaired. These people often cannot individually process the visual information in their surroundings, which could range from expiration dates to credit card numbers. Hiring humans to interpret images taken by the blind or visually impaired is expensive and raises major privacy concerns. However, with a reliable solution to VQA, this interpretation can be automated.

In this work, we move further along the path to reliable VQA for the visually impaired. We improve the performance of a strong baseline [7] on the VizWiz VQA 2.0 dataset [4], which was collected by blind people navigating the real world. Our primary contributions include the use of image augmentation techniques to expand the training dataset, the addition of specific features to address common types of visual questions, and the development of a novel attention model.

The rest of the paper is organised as follows. Section 2 discusses the chosen dataset. Section 3 surveys some related works. Section 4 lists a couple of baseline models we will compare our novel methods to. Those methods are then described in Section 5. Section 6 gives the results of our experiments, which are then analyzed and discussed in Section 7.

2 Data

As noted in the introduction, we use the VizWiz VQA 2.0 (VizWiz) dataset, a collection of human-annotated image-question pairs captured and asked by blind people in the real world [4]. The VizWiz

10-701 Introduction to Machine Learning (PhD). Carnegie Mellon University, Pittsburgh, PA, USA.

dataset contains 20,523 training examples, 4,319 validation examples, and 8,000 test examples. We use the "test-dev" split, which is a subset of 4,000 test examples publicly available for development purposes, to evaluate our experiments. Each example consists of an image-question pair along with ten human-annotated answers to the visual question. We follow the convention set by Antol et al. [2], by which accuracy is defined as $\min(1, \#\text{ agreeing answers}/3)$. Thus, an answer is considered correct (i.e. a *consensus answer*) if it agrees with at least three of the ten human annotations.

This dataset presents significant challenges as a learning problem. While existing image recognition mechanisms work best with high-quality images such as those found in the popular MSCOCO dataset, images in the VizWiz dataset are often poor qualit, because all images are captured by blind people who are unable to verify image quality. For instance, the images may be blurry or fail to capture an object of interest. Furthermore, many of the questions in the dataset are open-ended with many possible responses; in fact, over 10% of questions do not have a single consensus answer. This lack of unity among responses demonstrates the difficulty of the task to humans, let alone computers.

3 Related Work

Proposed in 2015 in [2], VQA is a relatively new task, and one that has proven to be exceedingly complex. One common approach used to simplify the VQA task is to treat it as a classification problem. The training examples are preprocessed to find the top N most frequently occurring answers, and the output space is then restricted to those N possibilities. Although a crude approximation, this removes the complexity of having an answer generation mechanism.

Significant work has also centered around the idea of "visual attention" for the closely related task of image caption generation. As the caption is iteratively generated, the model constantly varies its weights on each constituent image segment to focus on small regions only [12]. Applied to VQA, the question is usually converted into a query vector and combined with the image vector to determine one or more attention maps [5, 9].

One of the two areas of innovation concerns how the image and query vectors are generated. For instance, Anderson et al. [1] use a bottom-up Faster R-CNN to generate a set of salient image regions and a Gated Recurrent Unit (GRU) to process the question. The other concerns how to 'combine' the textual and visual features. An elaborate method, used by Fukui et al. [3] and first proposed by [8], involves taking outer products and the Fast Fourier Transform. However, we are not aware of works that exploit the sequential nature of the original attention model.

4 Baselines

In this section, we describe the two baseline models we use for comparison. The first is a naïve linear model used to demonstrate the capability of a generic deep neural network on the VizWiz dataset. The second is a stronger, yet still relatively simple model by Kazemi and Elqursh [7] that incorporates visual attention for better results. In Section 5, we go on to optimize this model for better performance.

4.1 Naïve Linear

Our first, and simplest, baseline model aims to build upon the success of CNNs in object classification and detection with a simple combination of Long Short-Term Memory (LSTM) cells, convolutional layers, and fully-connected layers. As shown in Figure 1, each of the words of the question are converted into embedding tokens, which are in turn sequentially fed into an LSTM. Separately, the input image, which is cropped to be 448×448, is fed through a series of convolutional layers with kernel size 3 and stride size 2. Each subsequent convolutional layer reduces the size of the features by roughly a factor of 2. When the output of the convolutional layer reaches a size of 27×27, it is flattened and appended to the feature result of the LSTM. This concatenation of image and question features is fed through a single hidden layer, and subsequently the output is obtained via the softmax function. The output is interpreted as a probability distribution over the 3000 most frequently appearing answers in the training set. The model is trained using stochastic gradient descent with an ℓ_2 regularization coefficient (weight decay) of 0.001 and the cross-entropy loss function.

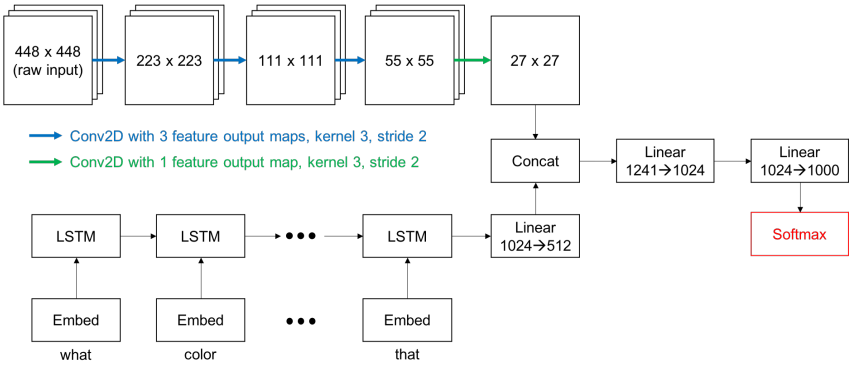

Figure 1: Architecture of our Naïve linear model, consisting of LSTM, convolutional layers and fully-connected linear layers. Concatenation is done by first flattening the 27×27 attention map.

4.2 Attention-based CNN

We use the model proposed by Kazemi and Elqursh as our stronger baseline [7]. The four main components of this model are feature extraction, attention, feature addition, and classification.

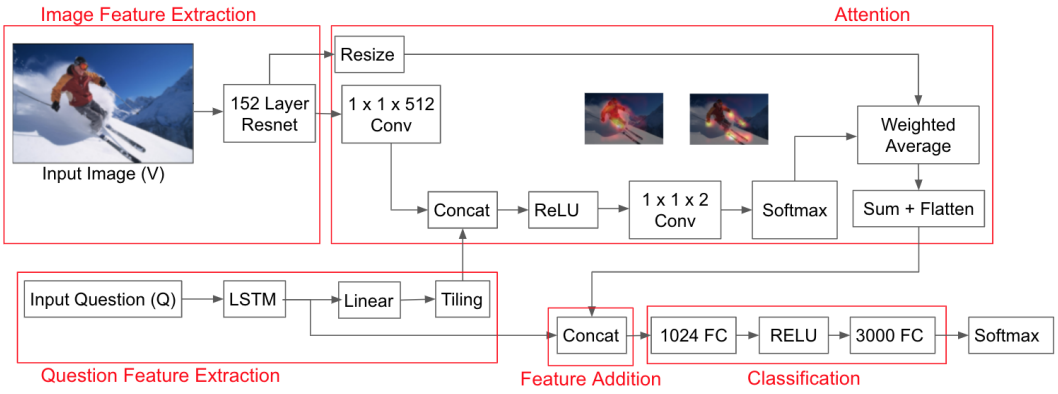

Figure 2: Architecture of the attention baseline model from [7]

The first component of the model extracts features from the images and the questions. The image feature extraction is done using a pre-trained 152-layer Resnet with output size $2048 \times 14 \times 14$ followed by a convolution which reduces the depth dimension to 512. As is typical, the original image is transformed with a center-crop and normalized before feeding it into the Resnet. Question features are extracted using an LSTM with state size 1024 followed by a fully-connected linear layer, which converts each question into a query vector Q of dimension 512. This vector is then broadcasted to a $512 \times 14 \times 14$ size, allowing it to be added with the image features.

In the attention component, the combined features of size $512 \times 14 \times 14$ are passed through a ReLU nonlinearity and a convolution, which reduces the depth dimension to 2. These two 14×14 maps M intuitively represent the two layers of attention. After applying a softmax and resizing, we take the weighted average of image features to get image glimpses U, a vector of dimension $2 \cdot 2048$.

Then, in the feature addition layer, U is concatenated again with Q for a vector T, where $\dim(T) = 2048 \cdot 2 + 1024 = 5120$.

Finally, in the classification component, this 5120 vector is reduced to a 3000-vector through a series of fully-connected layers and a softmax. As in the previous model, the output is interpreted as a probability distribution $P(a_k \mid U, Q)$. The loss function is defined to be $\mathcal{L} = \sum_{k=1}^{3000} -\frac{n_k}{10} \log P(a_k \mid U, Q)$, where n_k is the number of times a_k appears in the 10 actual answer annotations.

When training on the VizWiz dataset, we use default parameters, which are given by Adam Optimization, a Learning Rate of 0.001 (with Half-life 50,000), and a dropout of 0.5 on input features of all layers. However, we notice that the validation loss begins to increase even as the training loss is still decreasing, which is evidence of overfitting. To solve this, we add a weight decay of 1e-5 to the optimizer, which has effects similar to ℓ_2-regularization. Figure 3 shows plots of validation and training loss before and after using weight decay. This is a technique we reuse later on in Section 5 in our own experiments.

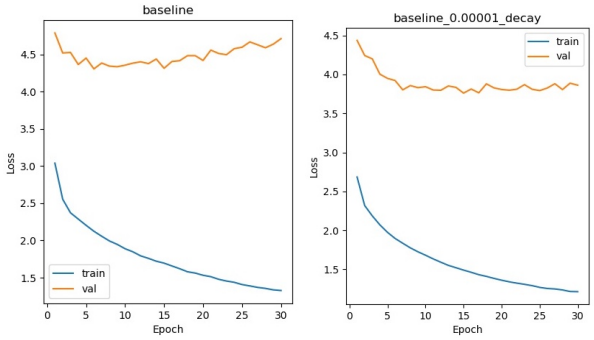

Figure 3: Effect of weight decay on training and validation losses. Left: no weight decay. Right: 1e-5 weight decay

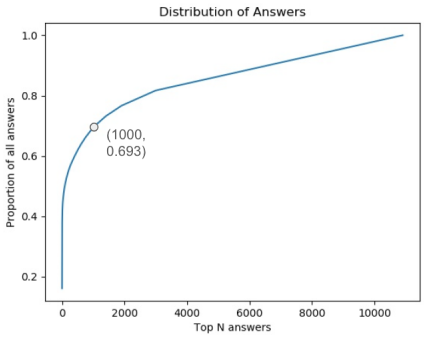

Figure 4: Cumulative frequencies of answers, ordered by most to least frequent.

5 Methods

The first few modifications we tried were changes to hyperparameters and training procedure. This includes changing the maximum answer cutoff in Section 5.1 and image augmentation in Section 5.2. We then proceed to modify the attention baseline model (Section 4.2) itself by modifying the attention network in Section 5.3 and augmenting the inputted features in Section 5.4. We put together the augmented features and modified attention network in Section 5.5.

5.1 Max Answer Cutoff

The first modification we make to the attention baseline is motivated by the somewhat arbitrary choice of classifying using the top $N = 3000$ most frequently occurring answers. We analyze the distribution of answers to determine if this choice of N is justified. Figure 4 shows a plot of cumulative frequencies of answers, ordered by most frequent to least frequent answers. This distribution is very front-heavy – only 2986 answers even appear at least twice. This justifies the choice of $N = 3000$, as it excludes almost exactly those answers which appear exactly once in the training set.

However, we hypothesize that we can further exploit the front-heavy distribution by decreasing the value of N to 1000. This would still include 69.3% of all answers, which is potentially a small price to pay in return for greatly simplifying the classification problem. Furthermore, the 2000 excluded answers are less likely to be relevant for the validation and test sets.

5.2 Image Augmentation

The increased complexity of machine learning models in recent years has increased dependence on an abundance of training data to avoid overfitting. The technique of data augmentation, which, as described by Shorten and Khoshgoftaar [10], encompasses a suite of techniques that enhance the size and quality of training datasets, has proven invaluable in reducing overfitting without requiring the collection of additional data.

In the present work, we leverage image augmentation in an attempt to mitigate many of the challenges presented by the VizWiz dataset. We use Gaussian blur and salt-and-pepper augmentations in hope of improving model performance on blurry and poor-quality images, which occur frequently in the dataset. In addition, we use cropping and rotation augmentations to improve performance on images captured in an unusual frame of reference, which are also common in the dataset. For each of the

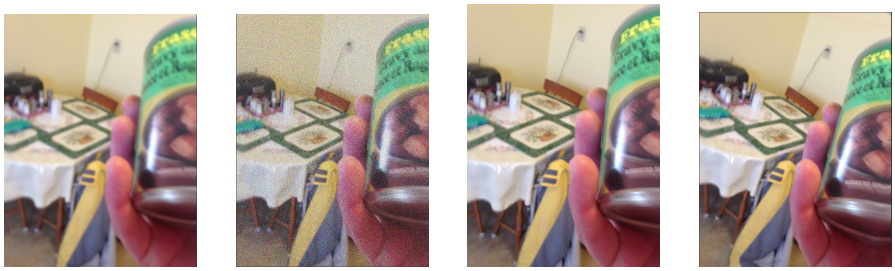

Figure 5: Types of data augmentation. From left to right: Gaussian blur, salt and pepper, cropping, and rotation.

training images, random permutations of these four types of augmentation, each of which has its own random parameters, are applied in order to generate new training data.

5.3 Modified Attention

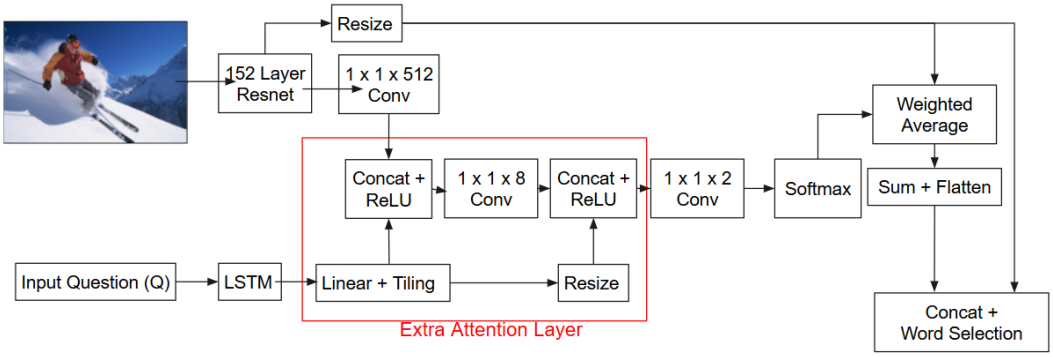

Figure 6: Architecture of the modified attention model.

Kazemi and Elqursh [7] rely on different weight initializations to generate two distinct attention maps. However, we hypothesize that processing the question once only does not sufficiently capture the contextual details of a question. For example, in the question "What is the number on the building", a model has to first locate the building (the context), before searching for any inscribed number. Stacked Attention Models, proposed by Yang et al. [13], combine image feature vectors and the query vectors to form a new query at each layer of the stack. However, we decided to keep the question unchanged when reading it a second time, instead summarizing the image based on the first attention layer into eight layers. The second pass returns two attention maps M as in the original model.

Since the question has already been processed twice with the image during the attention layer, we no longer pass it into the Concat layer. Instead, the original image features are averaged in the breadth and width dimensions, obtaining an average feature pixel of dimension 2048. This is then concatenated with the image glimpses and fed into the classifier. Passing the average feature pixel into the classifier retains information about the background, leaving the attention component free to focus on objects in the foreground. The revised architecture is presented in Figure 6.

By examining the contents of M for individual images, we see from Figure 10 that feeding the input query a second time creates more relevant attention maps.

5.4 Feature Augmentation

We engineer three additional features. Two binary classifiers determine whether an input question can be answered by a color and whether the image should be labelled 'unsuitable', while a further multinomial classifier seeks to identify the color associated with an image based on the most common color-related training answers. Using pretrained weights for these models, we test their effectiveness by combining their outputs with the baseline model at the Concat layer (Figure 11).

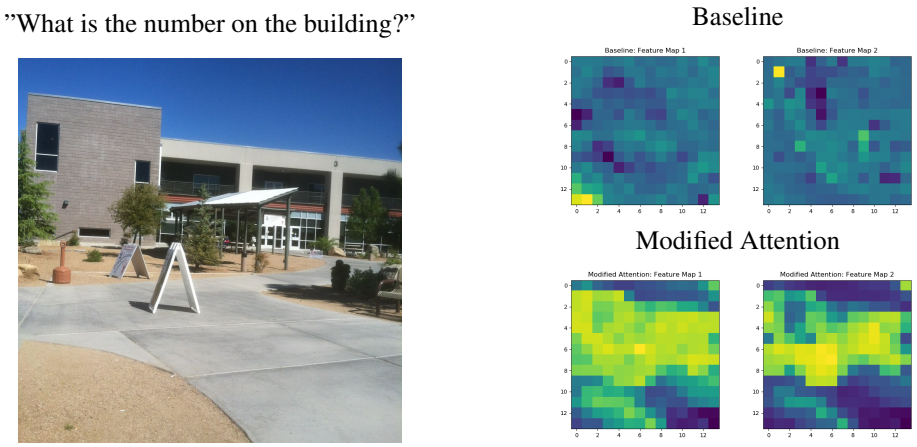

Figure 10: The modified attention model generates a better attention map by locating the building.

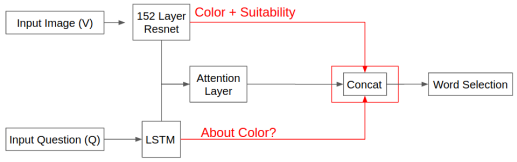

Figure 11: Combining augmented features into the baseline model.

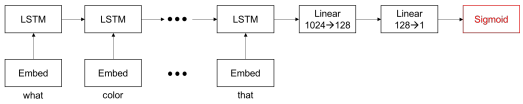

Figure 12: Architecture for determining whether a question is about color.

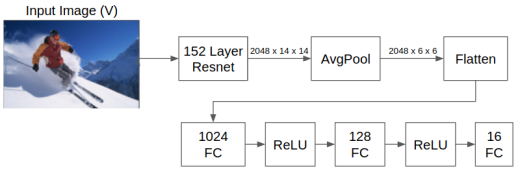

Figure 13: Architecture for predicting the color of an image.

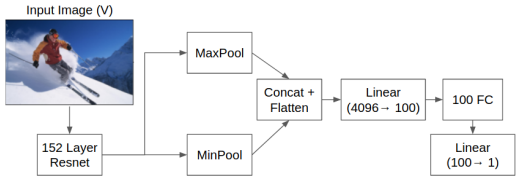

Figure 14: Architecture for determining whether an image is suitably clear.

5.4.1 “Is the question about color?”

This feature, in tandem with a prediction of the color of an input image itself, aims to improve performance relative to the baseline on those images which are answered by colors.

More formally, given a question $Q \in \Sigma^*$, we seek to learn a prediction function $C : \Sigma^* \to \{0, 1\}$ such that $C(Q) = 1$ if Q may be answered by a color (at least one of the ten human answers to the question is a color), and $C(Q) = 0$ otherwise.

To learn this function, we train a simple neural network consisting of an LSTM cell followed by a single hidden layer, with sigmoid output, as shown in Figure 12. The sigmoid activation means that our neural network actually learns a relaxation $f : \Sigma^* \to [0, 1]$, where the output is real-valued rather than discrete, to allow for backpropagation. We train it using stochastic gradient descent with an ℓ_2 regularization (weight decay) coefficient of 0.01 and binary cross-entropy loss function

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} y_i \log f(q_i) + (1 - y_i) \log(1 - f(q_i)) \tag{1}$$

where q_i represents the i^{th} question, y_i is a binary label indicating whether the question pertains to color, and $f(q_i) \in [0, 1]$ is the prediction generated by the neural network, with θ representing the parameters of this model.

5.4.2 "What color best describes the image?"

This feature aims to predict the color of the image. The architecture for this model, as shown in Figure 13 takes the $2048 \times 14 \times 14$ Resnet features for an image I, downsamples them through an average pool, and then passes that information through a series of fully connected layers and a softmax. The final output represents a probability distribution $P(c_k \mid I)$ over the 16 most common colors in the training dataset. The loss is defined as $\mathcal{L} = \sum_{k=1}^{16} -\frac{n_k}{10} \log P(c_k \mid I)$, where n_k is the number of times the color c_k appears among the 10 answers for image I. The model is trained with an Adam optimizer with 1e-2 weight decay using the subset of the training data having at least one color answer.

5.4.3 "Is the image suitable?"

Approximately 20% of the training dataset was classified as either "unsuitable" or "unsuitable image", which presents an opportunity to explicitly engineer a feature to indicate that an image is overly blurry. This avoids the common image detection pitfall of identifying an object when none is present. Again, such a signal could be delivered by an Image Quality Assessment model that takes as input only the Resnet convolutional output of each image. Following Kang et al. [6], we pass the $2048 \times 14 \times 14$ Resent output into MaxPool and MinPool layers to obtain concise representations of each layer as its max and min (Figure 14). The linear layer further reduces the dimension to 100, before a fully-connected layer with sigmoid activation generates a normalized value to indicate the quality of an image. To prevent the model from simply classifying everything in this imbalanced dataset as "suitable", we doubled the penalty on "unsuitable" images before applying binary cross-entropy loss as in 1. In this case, y_i equals 1 if "unsuitable" or "unsuitable image" was one of the consensus answers. This model is also trained with an Adam optimizer with 1e-2 weight decay.

5.5 Combined Model

We concatenate the output of the three smaller models discussed in section 5.4 with the image glimpses and compressed features from the modified attention model of section 5.3 at the Concat layer. The weights in our feature augmentation models are pretrained and therefore not updated during the training of the combined model. We use the same loss function, optimizer, and weight decay as in the attention baseline.

6 Results

We tested our models on a `p2.8xlarge` AWS instance with 8 GPUs. We trained models using the original dataset for 30 epochs, and for 12 epochs when using the augmented datasets. Unless otherwise stated, we used 3000 max answers, a decay of 1e-5, and no image augmentation. Each configuration was trained thrice, and we submitted our predicted answers on the test set to the official website. Table 1 summarizes the average test accuracies. Each set of three trials saw a standard deviation of below 1%, indicating high consistency.

Table 1: Final Results

Method	Test Accuracy (%)
Naïve Linear + 1e-3 Weight Decay (4.1)	36.7 ± 0.07
Attention (4.2, [7])	46.4 ± 0.51
Attention + 1000 Max Answers (5.1)	46.4 ± 0.44
Attention + Image Augmentation (5.2)	49.2 ± 0.25
Modified Attention + Image Augmentation (5.3)	50.3 ± 0.07
Feature Augmentation + Image Augmentation (5.4)	48.3 ± 0.39
Modified Attention + Feature Augmentation + Image Augmentation (5.5)	49.8 ± 0.19

7 Discussion and Analysis

The results of our experiments, shown in Table 1, demonstrate that image augmentation and modified attention, two of our novel contributions in this paper, significantly improve performance on the VizWiz dataset. Image augmentation alone achieves 3.2% higher test accuracy than the attention baseline in [7], and our strongest-performing network, modified attention with image augmentation, achieves 3.9% higher test accuracy than the baseline.

The performance improvement suggests that our image augmentations at least partially satisfied their intended purpose of providing the neural networks with more exposure to poor-quality images, images containing uncommon objects, and images taken from an unusual vantage point. This underscores the importance of plentiful and variable training data, especially in complex tasks such as VQA, in building a more flexible model.

Modified Attention, by gaining a more accurate understanding of an image's composition, has also proven effective. By sequentially reading each question twice, the revised model exhibits markedly hierarchical attention maps (Figure 10) that capture the contextual relationships between image segments. At the bottom of the hierarchy are the smallest details of an image that actually answer the question, and our model is now able to generate answers based on that alone.

The results also demonstrate that adjusting the maximum answer cutoff and engineering custom features, our two other primary experiments, did not improve performance relative to this baseline. This was initially surprising given the individual performance of our engineered features: the `is_color` feature attained 96.7% accuracy; the `suitable` feature, 79.6%; and the `color` feature, 42.8%. Despite encouraging numbers, each of the features struggled on difficult examples. The `is_color` feature essentially devolved into checking whether the word "color" was present in the question, misclassifying examples such as "what is this marker?". The `color` feature fails on images with multiple objects, where the desired color depends on information present in the question. The `suitable` feature tends to output values between 0.4 and 0.6, which fails to sufficiently discriminate between suitable and unsuitable images.

We also hypothesize that our engineered color and suitability features yielded no benefit because these features provided no further insight than what the network had already implicitly learned from the many questions pertaining to color and those with unsuitable images. Furthermore, since these features were extracted and used by separate models, the entire system is no longer optimized end-to-end. As for the 1000 Max Answers experiment, the rarity of answers outside the most common 1000 probably incentivized the network to avoid such answers, and thus our adjustment from a possible 3000 answers to a possible 1000 had little effect on its output.

Taking a broader perspective, the results highlight the efficacy of attention for visual question-answering: The test accuracy of the Naïve linear model was at least 10% lower than that of every attention-based model. Without Attention, a model would treat each evenly-sized patch of an image as equally important, preventing it from emphasizing the regions suggested by the question.

Despite achieving strong results, several limitations remain with our approach. Each of our networks treats VQA as a classification task, in which the most probable of 1000 or 3000 possible answers is chosen, rather than a generative one in which every question generates a unique response. While this is unlikely to hinder performance significantly on the test accuracy metric, it would affect the experience of a visually-impaired person who needs an individualized answer that is potentially outside the set of preselected answers.

One possible solution is to use transformers, first used by Vaswani et al. [11] for machine translation. Analogous to Recurrent Neural Networks, a possibly more robust model could use our N-vector output to generate answers word-by-word by taking argmax over probabilities on a much larger English corpus. After each word, the "residual" vector can be further analyzed and additional words generated, until a special end-of-output symbol is reached. A model that has learned 'grey cat' and 'black' would now also be able to generate 'black cat'. This could be a further improvement to our already strong model.

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[3] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[4] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. 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.

[5] A. Jabri, A. Joulin, and L. Van Der Maaten. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer, 2016.

[6] L. Kang, P. Ye, Y. Li, and D. Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1733–1740, 2014.

[7] V. Kazemi and A. Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*, 2017.

[8] T. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1449–1457, 2015. doi: 10.1109/ICCV.2015.170.

[9] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4613–4621, 2016.

[10] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 2019.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008, 2017.

[12] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[13] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016.