DS340 Final Report

Authors: Chung-Yeh(Oliver) Yang, Jin Kyu(Jaden) Cho

Introduction

The problem we aimed to solve was determining whether a model could accurately classify a given answer to a question about an image as correct or incorrect. This challenge falls within the domain of Visual Question Answering (VQA), a task that combines computer vision and natural language processing.

The motivation for addressing this problem lies in the potential applications of VQA systems, such as assisting visually impaired individuals, improving automated customer service, and enhancing educational tools. However, accurately evaluating answers in a VQA context is complex, as it requires integrating visual and textual information to make nuanced decisions.

For this project, we developed the fusion model that evaluates the correctness of simple answers provided in response to questions about images. The dataset includes images paired with questions and answers, requiring the model to understand and interpret both modalities effectively.

Methodology

Dataset Processing

We utilized a publicly available Visual Question Answering (VQA) dataset comprising image-question pairs with multiple answers. For each pair, the dataset provides 10 answers labeled as correct or incorrect based on agreement with human annotators. To preprocess this dataset, we implemented a custom Python class, VQADataset, which efficiently pairs images with their corresponding questions and answers using unique IDs from the dataset metadata. This approach ensured structured data handling, streamlining processing during model training and evaluation.

To create balanced training data, incorrect answers were generated by randomly sampling half of the answers from unrelated questions in the dataset. These incorrect answers were labeled as 0, while the original correct answers were labeled as 1. This strategy increased dataset diversity, challenging the model to effectively distinguish between correct and incorrect answers. Additionally, we utilized a DataLoader to batch and shuffle the data efficiently during training

and evaluation. This ensured that each batch contained a mix of correct and incorrect answers, enhancing the model's learning capabilities.

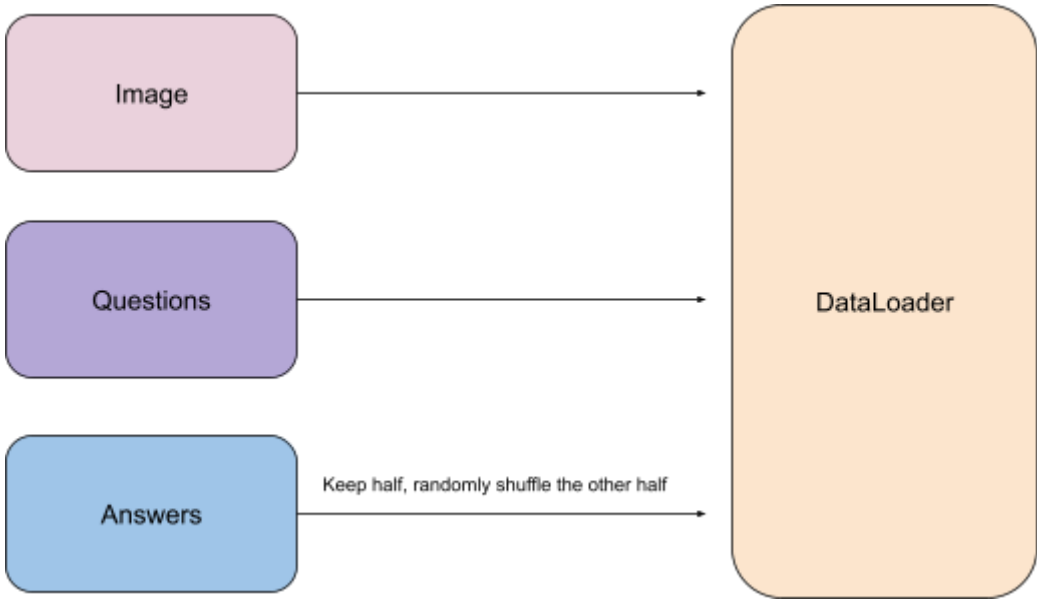

Figure 1: Illustration of preprocessing workflow, where images, questions, and answers are paired and batched using a DataLoader

---

Fusion Model Training

The image processing pipeline employed a pre-trained ResNet50 model from the PyTorch library to extract feature embeddings. Specifically, the output from the penultimate layer (a 2048-dimensional vector) was used as the image embedding, capturing high-level visual features critical for downstream tasks.

For text processing, questions and answers were combined into a single input string (e.g., *"How many players are on the court? Ten."*). This string was tokenized using the BERT tokenizer, which converts text into subword tokens and numerical representations. The tokenized inputs were processed through a pre-trained BERT model to generate a 768-dimensional embedding from the final hidden state, encapsulating the semantic meaning of the question-answer pair.

In the attention-based model, a 4-head MultiHeadAttention layer was implemented to enhance the integration of image and text features. This layer dynamically computed weighted combinations of the image and text embeddings, emphasizing the most relevant features from

each modality. These outputs were concatenated and passed to the fusion step. Conversely, the baseline model directly concatenated raw image and text embeddings without an attention layer.

The fused embeddings were passed through a classification head, designed to predict whether an answer was correct. This head consisted of two fully connected layers with 512 and 256 neurons, interspersed with ReLU activations and dropout layer (0.3 probabilities) to prevent overfitting. Finally, a single output neuron with a sigmoid activation function provided a probability score indicating the likelihood of the answer being correct.

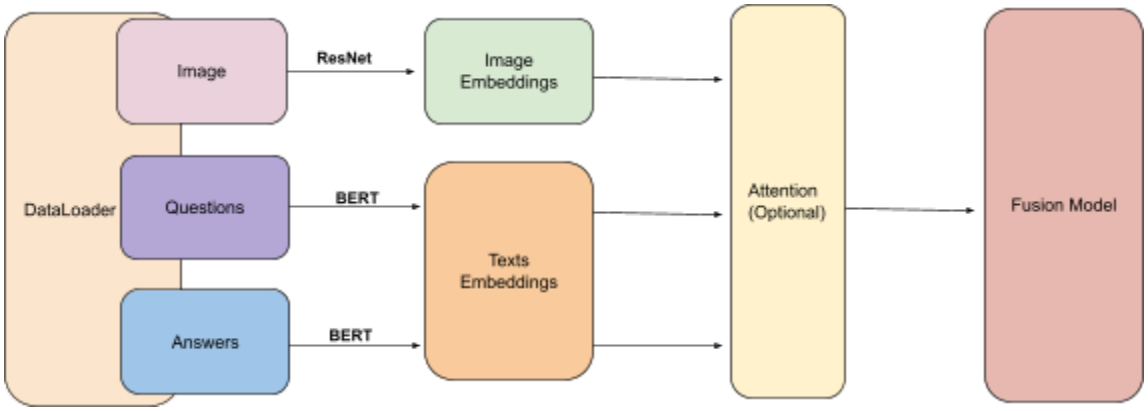

Figure 2: Illustration of model training workflow, transforming datasets into embeddings through pretrained encoders and feeding to the model

---

Training Configuration

The training was framed as a binary classification problem, with the Binary Cross Entropy Loss (BCELoss) function used to compute the loss. BCELoss is well-suited for binary classification tasks where the output is a probability indicating one of two classes (correct or incorrect).

The AdamW optimizer was employed for its robust handling of sparse gradients and built-in weight decay regularization to mitigate overfitting. Its dynamic learning rate adjustment facilitated efficient convergence.

Models were trained for 5 epochs with a batch size of 64. During each epoch, the DataLoader iterated over the training set, feeding batches into the model. After training each epoch, performance was evaluated on a separate validation set. Key metrics—accuracy, precision, recall, and F1 score—were recorded to monitor progress and effectiveness.

Training was conducted on a machine equipped with an NVIDIA GPU, leveraging CUDA support for parallel processing. This setup significantly accelerated computations for image and text embedding extraction and model training.

Validation and Testing

To evaluate model performance, the following metrics were employed:

- **Accuracy:** The percentage of answers correctly classified as correct or incorrect, providing a general performance measure.
- **Precision:** The proportion of true positives (correctly classified correct answers) among all answers predicted as correct, indicating the model's ability to avoid false positives.
- **Recall:** The proportion of true positives among all actual correct answers, reflecting the model's ability to identify all correct answers.
- **F1 Score:** The harmonic mean of precision and recall, balancing false positives and false negatives.

After training, the models were evaluated on an unseen test dataset to assess generalization. Predictions were compared with ground-truth labels, and final metrics (accuracy, precision, recall, and F1 score) were computed to comprehensively evaluate the model's effectiveness in classifying answers as correct or incorrect.

Results

The sample result of the model is presented in **Figure 3**. The model predicts a binary label (0 or 1) to determine whether a given answer correctly matches the associated image and question. As shown in the figure, the question "How many hot dogs are shown?" is paired with the answer "1". The model accurately predicts the label as "1" (correct), demonstrating its ability to learn

Figure 3: A sample result which successfully predicts the correct match for the question and image

relationships between images, questions, and answers.

To evaluate the model's performance under different settings, we experimented with datasets of varying sizes (1,000, 10,000, and 100,000 data points) and compared models with and without attention mechanisms in the fusion layer. The table below summarizes the results, providing the training and validation accuracy for each configuration.

1,000 Data Points, Model Without Attention

```
Starting Epoch 1/5
Batch [0/16], Loss: 0.6990
Batch [10/16], Loss: 0.7029
Epoch [1/5] - Train Loss: 0.6969, Train Accuracy: 0.4979, Val Loss: 0.6926, Val Accuracy: 0.4989
Starting Epoch 2/5
Batch [0/16], Loss: 0.6962
Batch [10/16], Loss: 0.6942
Epoch [2/5] - Train Loss: 0.6943, Train Accuracy: 0.5051, Val Loss: 0.6900, Val Accuracy: 0.5158
Starting Epoch 3/5
Batch [0/16], Loss: 0.6890
Batch [10/16], Loss: 0.6931
Epoch [3/5] - Train Loss: 0.6922, Train Accuracy: 0.5255, Val Loss: 0.6871, Val Accuracy: 0.5334
Starting Epoch 4/5
Batch [0/16], Loss: 0.6875
Batch [10/16], Loss: 0.6873
Epoch [4/5] - Train Loss: 0.6897, Train Accuracy: 0.5344, Val Loss: 0.6853, Val Accuracy: 0.5405
Starting Epoch 5/5
Batch [0/16], Loss: 0.6982
Batch [10/16], Loss: 0.6874
Epoch [5/5] - Train Loss: 0.6889, Train Accuracy: 0.5402, Val Loss: 0.6824, Val Accuracy: 0.5550
Training completed successfully.
```

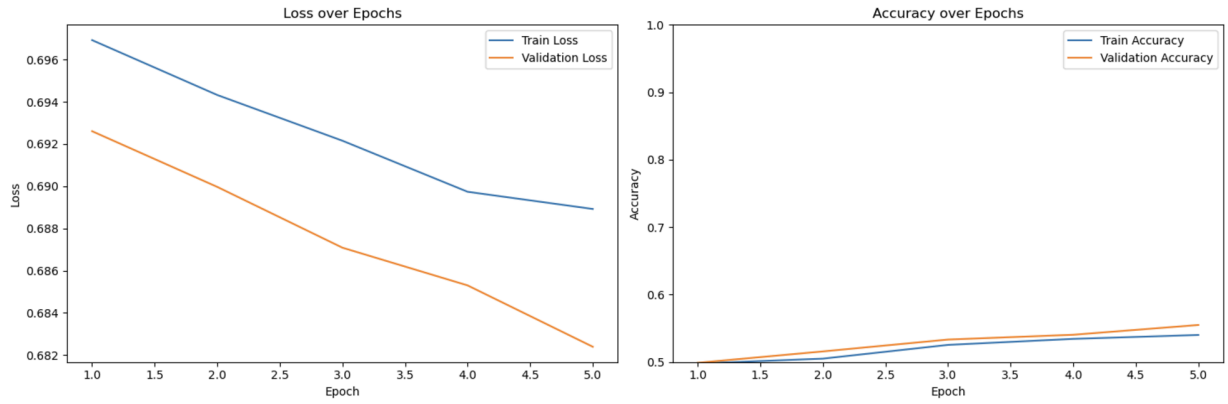

Figure 4: Training and Validation Performance Over Epochs With 1,000 Data Points Without Attention(Left-to-Right) Loss, Accuracy

The metrics and graphs for the model trained on 1,000 data points without attention indicate limited learning due to the small dataset size. The final training loss is 0.6899, and the validation loss is 0.6824, showing minimal improvement across epochs and indicating that the model struggles to generalize effectively. The training accuracy ends at 54.02%, while the validation

accuracy is slightly higher at 55.50%, suggesting mild generalization but no significant optimization. In the "Loss over Epochs" graph, both training and validation losses show a slight downward trend, but the minimal separation between the two indicates limited capacity for improvement. The "Accuracy over Epochs" graph reflects a slow and steady rise in both training and validation accuracy, with little divergence, signifying underfitting due to insufficient data. Overall, the results emphasize the need for a larger dataset or more complex features to enhance learning and generalization.

10,000 Data Points, Model Without Attention

```
Starting Epoch 1/5
Batch [0/157], Loss: 0.7003
Batch [50/157], Loss: 0.6955
Batch [100/157], Loss: 0.6873
Batch [150/157], Loss: 0.6797
Epoch [1/5] - Train Loss: 0.6905, Train Accuracy: 0.5256, Val Loss: 0.6529, Val Accuracy: 0.6233
Starting Epoch 2/5
Batch [0/157], Loss: 0.6667
Batch [50/157], Loss: 0.6085
Batch [100/157], Loss: 0.5339
Batch [150/157], Loss: 0.4946
Epoch [2/5] - Train Loss: 0.5845, Train Accuracy: 0.7093, Val Loss: 0.4769, Val Accuracy: 0.7702
Starting Epoch 3/5
Batch [0/157], Loss: 0.4791
Batch [50/157], Loss: 0.4638
Batch [100/157], Loss: 0.4471
Batch [150/157], Loss: 0.4760
Epoch [3/5] - Train Loss: 0.4715, Train Accuracy: 0.7689, Val Loss: 0.4450, Val Accuracy: 0.7738
Starting Epoch 4/5
Batch [0/157], Loss: 0.4641
Batch [50/157], Loss: 0.4637
Batch [100/157], Loss: 0.4914
Batch [150/157], Loss: 0.4665
Epoch [4/5] - Train Loss: 0.4449, Train Accuracy: 0.7731, Val Loss: 0.4394, Val Accuracy: 0.7732
Starting Epoch 5/5
Batch [0/157], Loss: 0.4360
Batch [50/157], Loss: 0.4528
Batch [100/157], Loss: 0.3925
Batch [150/157], Loss: 0.4008
Epoch [5/5] - Train Loss: 0.4217, Train Accuracy: 0.8002, Val Loss: 0.3867, Val Accuracy: 0.8349
Training completed successfully.
```

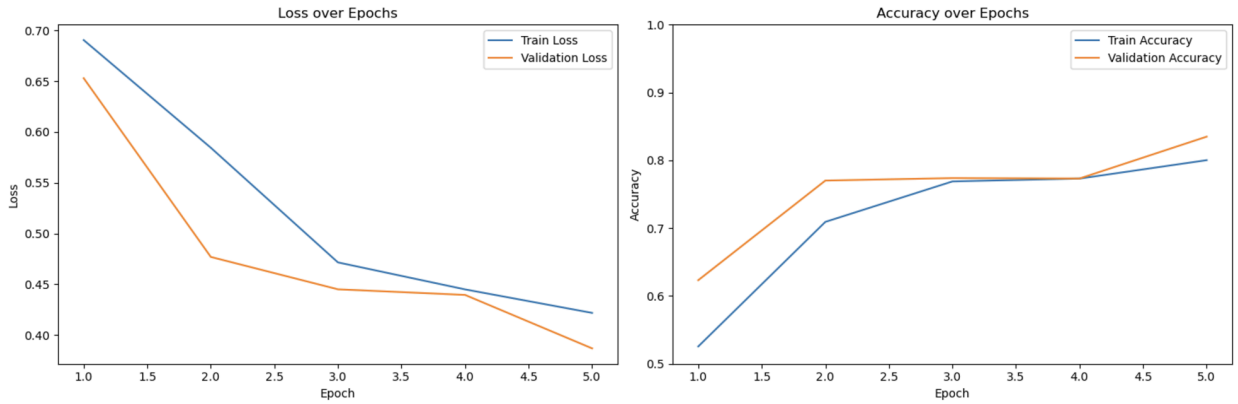

Figure 5: Training and Validation Performance Over Epochs With 10,000 Data Points Without Attention(Left-to-Right) Loss, Accuracy

The metrics and graphs indicate that the model has successfully learned from the training data and generalizes well to the validation set. Both the training and validation losses steadily decrease across epochs, with the final training loss at 0.4217 and validation loss at 0.3867, showing effective optimization without signs of overfitting. The validation accuracy slightly

surpasses training accuracy (83.49%, 80.02%, respectively), suggesting the model generalizes robustly to unseen data. In the "Loss over Epochs" graph, both training and validation losses follow a downward trend, reflecting consistent learning. The "Accuracy over Epochs" graph shows steady improvements in both metrics, with validation accuracy plateauing slightly higher than training accuracy.

100,000 Data Points, Model Without Attention

```
Starting Epoch 5/5
Batch [0/1563], Loss: 0.2787
Batch [100/1563], Loss: 0.2845
Batch [200/1563], Loss: 0.2885
Batch [300/1563], Loss: 0.2577
Batch [400/1563], Loss: 0.3145
Batch [500/1563], Loss: 0.3270
Batch [600/1563], Loss: 0.2990
Batch [700/1563], Loss: 0.3170
Batch [800/1563], Loss: 0.2841
Batch [900/1563], Loss: 0.2713
Batch [1000/1563], Loss: 0.3194
Batch [1100/1563], Loss: 0.2971
Batch [1200/1563], Loss: 0.3090
Batch [1300/1563], Loss: 0.2563
Batch [1400/1563], Loss: 0.2905
Batch [1500/1563], Loss: 0.2615
Epoch [5/5] - Train Loss: 0.2938, Train Accuracy: 0.8907, Val Loss: 0.3641, Val Accuracy: 0.8471
Training completed successfully.
```

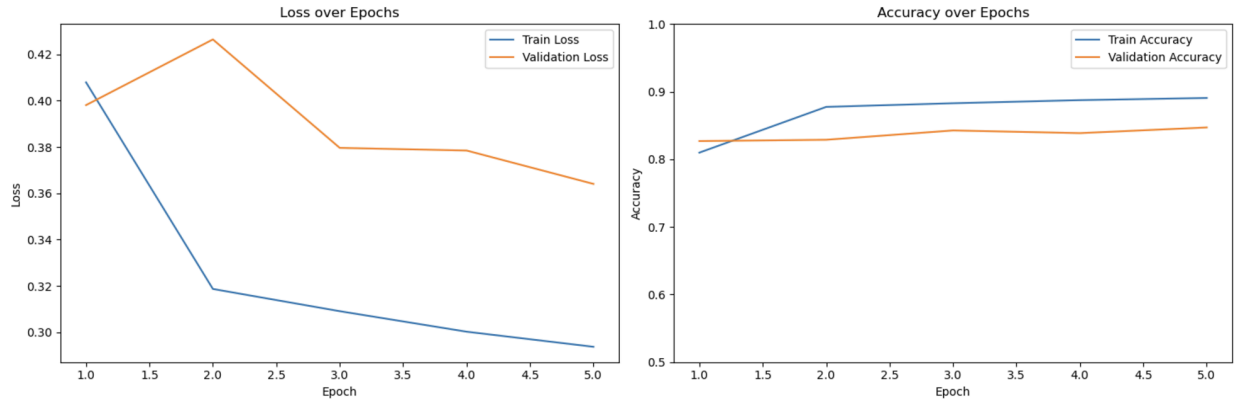

Figure 6: Training and Validation Performance Over Epochs With 100,000 Data Points Without Attention: (Left-to-Right) Loss, Accuracy

The metrics and graphs for the 100,000 data points model without attention demonstrate strong performance and effective optimization. The final training loss is 0.2938, and the validation loss is 0.3641, with a training accuracy of 89.07% and validation accuracy of 84.71%, reflecting good generalization. The "Loss over Epochs" graph shows a consistent decline in training loss, while the validation loss fluctuates slightly but remains lower than the initial epochs, indicating stable learning. The "Accuracy over Epochs" graph illustrates steady improvement in both training and validation accuracy, with the training accuracy slightly surpassing validation accuracy in the

final epochs, suggesting that the model fits the training data well while maintaining strong generalization.

1,000 Data Points, Model With Attention

```
Starting Epoch 1/5
Batch [0/16], Loss: 0.6959
Epoch [1/5] - Train Loss: 0.6937, Train Accuracy: 0.5042, Val Loss: 0.6926, Val Accuracy: 0.5274
Starting Epoch 2/5
Batch [0/16], Loss: 0.6937
Epoch [2/5] - Train Loss: 0.6932, Train Accuracy: 0.5067, Val Loss: 0.6915, Val Accuracy: 0.5499
Starting Epoch 3/5
Batch [0/16], Loss: 0.6909
Epoch [3/5] - Train Loss: 0.6918, Train Accuracy: 0.5262, Val Loss: 0.6887, Val Accuracy: 0.5586
Starting Epoch 4/5
Batch [0/16], Loss: 0.6890
Epoch [4/5] - Train Loss: 0.6900, Train Accuracy: 0.5411, Val Loss: 0.6847, Val Accuracy: 0.5683
Starting Epoch 5/5
Batch [0/16], Loss: 0.6850
Epoch [5/5] - Train Loss: 0.6883, Train Accuracy: 0.5486, Val Loss: 0.6822, Val Accuracy: 0.5708
Training completed successfully.
```

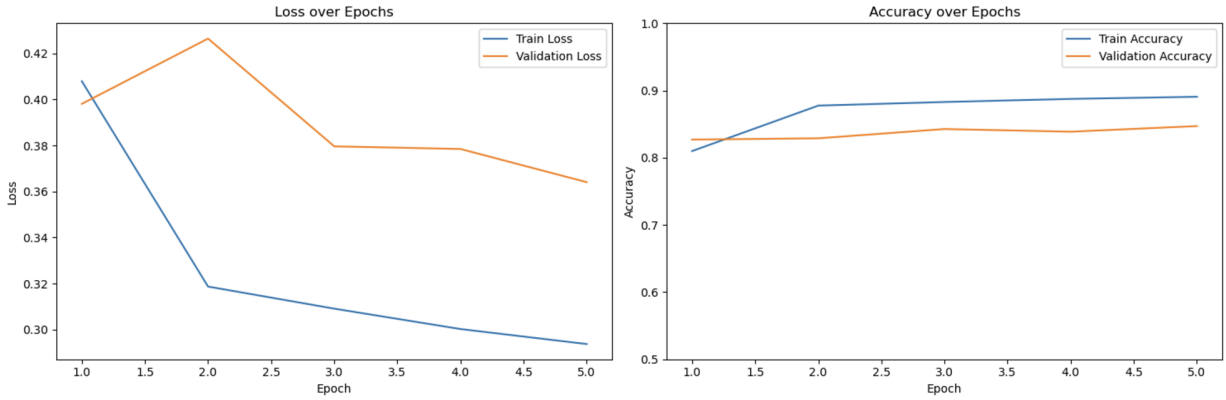

Figure 7: Training and Validation Performance Over Epochs With 1,000 Data Points With Attention(Left-to-Right) Loss, Accuracy

The metrics and graphs for the 1,000 data points model with attention indicate slight improvement compared to the non-attention counterpart, but the overall learning remains constrained by the small dataset size. The final training loss is 0.6883, and the validation loss is 0.6822, showing minimal difference, which suggests limited generalization. The training accuracy reaches 54.86%, while the validation accuracy improves slightly to 57.08%, indicating that the attention mechanism helps the model generalize marginally better. In the "Loss over Epochs" graph, the training loss steadily decreases, but the validation loss fluctuates slightly, showing inconsistent optimization due to the limited data. The "Accuracy over Epochs" graph shows a gradual improvement in both training and validation accuracy, with validation accuracy consistently exceeding training accuracy, suggesting that the attention mechanism provides some

benefit in this context. However, the model's performance remains underwhelming, emphasizing the need for more data to effectively leverage attention mechanisms.

10,000 Data Points, Model With Attention

```
Starting Epoch 5/5
Batch [0/157], Loss: 0.4024
Batch [50/157], Loss: 0.3223
Batch [100/157], Loss: 0.3038
Batch [150/157], Loss: 0.3346
Epoch [5/5] - Train Loss: 0.3423, Train Accuracy: 0.8612, Val Loss: 0.3288, Val Accuracy: 0.8707
Training completed successfully.
```

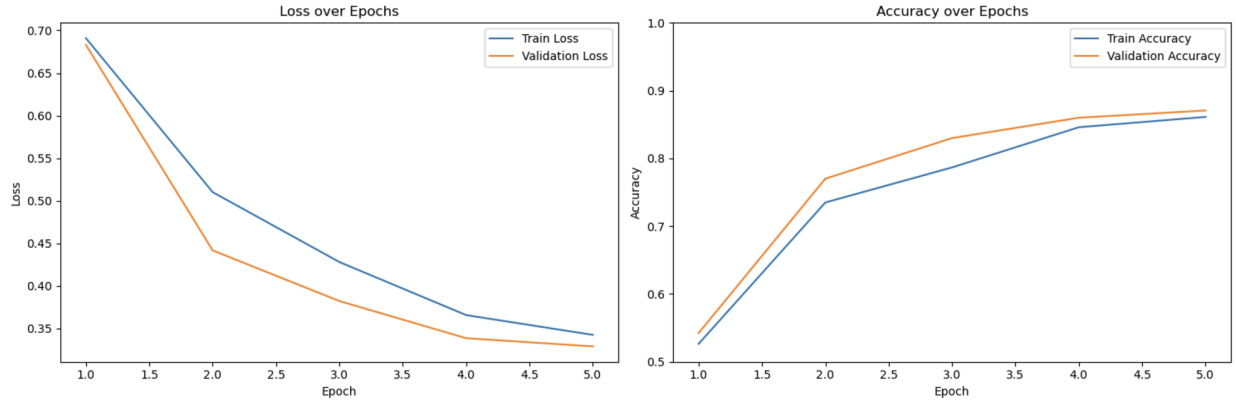

Figure 8: Training and Validation Performance Over Epochs With 10,000 Data Points With Attention: (Left-to-Right) Loss, Accuracy

The metrics and graphs for this model with attention indicate steady learning and effective generalization, though the performance is moderate compared to prior models. The final training loss is 0.3423, and the validation loss is 0.3288, showing minimal difference and suggesting balanced training without overfitting. The training accuracy reaches 86.12%, while the validation accuracy is slightly higher at 87.07%, indicating good generalization to unseen data. In the "Loss over Epochs" graph, both training and validation losses decrease sharply in the first epoch and stabilize thereafter, highlighting efficient early learning and convergence. The "Accuracy over Epochs" graph shows a steep initial rise in both accuracies, with the validation accuracy consistently exceeding training accuracy, confirming effective generalization. While the model demonstrates stability, the lower accuracy suggests room for improvement through further optimization.

100,000 Data Points, Model With Attention

```
Starting Epoch 5/5
Batch [0/1563], Loss: 0.3487
Batch [100/1563], Loss: 0.3048
Batch [200/1563], Loss: 0.3553
Batch [300/1563], Loss: 0.3925
Batch [400/1563], Loss: 0.4395
Batch [500/1563], Loss: 0.3967
Batch [600/1563], Loss: 0.3252
Batch [700/1563], Loss: 0.3760
Batch [800/1563], Loss: 0.3567
Batch [900/1563], Loss: 0.3416
Batch [1000/1563], Loss: 0.3451
Batch [1100/1563], Loss: 0.3316
Batch [1200/1563], Loss: 0.3442
Batch [1300/1563], Loss: 0.3093
Batch [1400/1563], Loss: 0.3227
Batch [1500/1563], Loss: 0.3516
Epoch [5/5] - Train Loss: 0.3462, Train Accuracy: 0.8599, Val Loss: 0.3385, Val Accuracy: 0.8622
Training completed successfully.
```

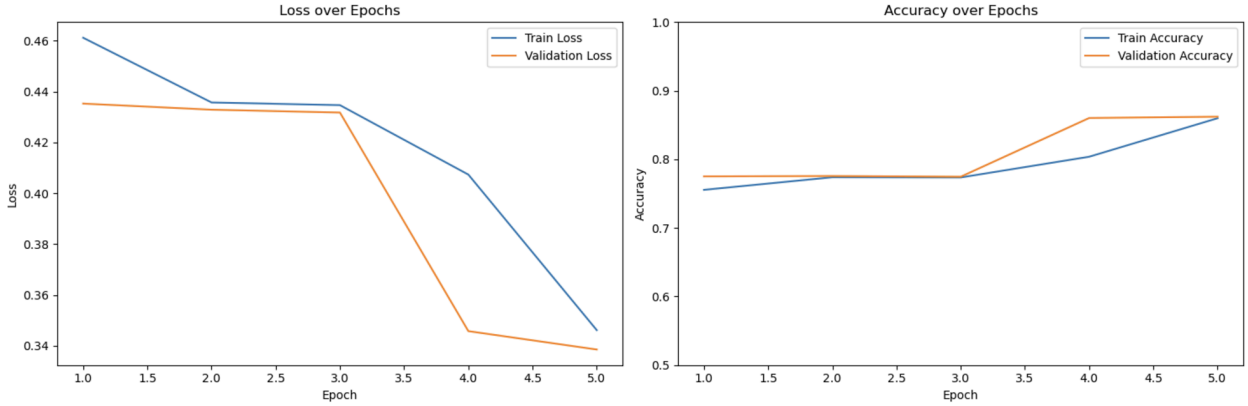

Figure 9: Training and Validation Performance Over Epochs With 100,000 Data Points With Attention: (Left-to-Right) Loss, Accuracy

The metrics and graphs for the model trained on 100,000 data points with attention reveal strong performance and effective generalization. The final training loss is 0.3462, and the validation loss is significantly lower at 0.3385, indicating the model performs better on unseen data, likely due to robust regularization or attention mechanisms. The training accuracy reaches 85.99%, while the validation accuracy is higher at 86.22%, reflecting strong generalization and minimal overfitting. In the "Loss over Epochs" graph, both training and validation losses decrease steadily, with a sharp decline early on and continued improvement toward the final epochs. The "Accuracy over Epochs" graph shows a steady increase, with validation accuracy surpassing training accuracy throughout, indicating the model captures meaningful patterns in the data. These results highlight the model's ability to scale effectively with larger datasets, leveraging attention mechanisms to improve performance.

The model trained on 10,000 data points with attention performs the best, achieving the highest validation accuracy (87.07%) and lowest validation loss (0.3288). This shows that attention mechanisms are effective at handling moderate-sized datasets by capturing detailed patterns and

improving generalization. The attention layers help the model focus on the most relevant features, making it ideal for tasks requiring the integration of multimodal information. While simpler models without attention can work well for straightforward datasets, attention mechanisms provide a better balance of complexity and learning capacity, avoiding overfitting and enhancing performance. With the right dataset size, attention mechanisms significantly improve how the model processes and integrates input data.

Interestingly, in most scenarios, the validation accuracy surpasses the training accuracy. This phenomenon can occur for several reasons: First, techniques such as dropout or weight decay may cause the model to perform better on unseen data by preventing overfitting during training. Moreover, during training, the model sees each batch of data with its own distribution due to batch normalization, which can slightly disrupt performance. However, during evaluation, batch normalization uses the entire dataset's statistics, leading to smoother and potentially better performance. Lastly, the validation set may contain simpler or more representative samples than the training set, making it easier for the model to achieve higher accuracy on validation data. These factors underline the importance of analyzing model behavior holistically, as higher validation accuracy does not always indicate an inherent flaw but could reflect effective generalization and robust model design.

Training Accuracy

	1,000 data points	10,000 data points	100,000 data points
Model w/o attention	54.02 %	80.02 %	89.07 %
Model with attention	54.86 %	86.12 %	85.99 %

Validation Accuracy

	1,000 data points	10,000 data points	100,000 data points
Model w/o attention	55.50 %	83.49 %	84.71 %
Model with attention	57.08 %	87.07 %	86.52 %

Conclusions

This project introduced a fusion-based deep learning model for Visual Question Answering (VQA) to classify the correctness of simple answers to questions about images. We developed two variants: one with attention mechanisms and one without, to assess their effectiveness in integrating multimodal data. Our results showed that the model trained on 10,000 data points with attention achieved the highest validation accuracy (87.07%) and lowest validation loss (0.3288). This demonstrates the ability of attention mechanisms to capture nuanced patterns and improve generalization for moderately sized datasets. Our findings highlight the importance of balancing model complexity with task requirements, offering valuable insights for enhancing future VQA systems.

References

OpenAI. *ChatGPT*. Version 4, 2024, https://openai.com/chatgpt. Used for code organization.

Sahu, Tezan. "Visual Question Answering with Multimodal Transformers." *Medium*, Data Science at Microsoft, 8 Mar. 2022, medium.com/data-science-at-microsoft/visual-question-answering-with-multimodal-transformers-d4f57950c867.