Lab Assignment 4
CSCI 5992 - Neural Networks and Deep Learning

Rajeev R Menon - 110581437

November 3, 2022

Validation Set

Methods

The dataset used in this experiment is the VizWiz-VQA dataset compiled by Gurari, Danna, et al. for VizWiz grand challenge [1]. The dataset was created with the help of blind people recording a question with respect to an image. 10 crowdsourced questions were also added to each image as well. The dataset consists of 20,523 training samples, 4319 validation samples, and 8000 test samples.

The experiment was conducted on a 14-inch MacBook Pro with an 8-core M1 Pro CPU. For this experiment, all 20,523 training samples were used to train various models which are validated using all of the available validation samples. The predictions on the first 1000 test samples are exported to a CSV file. A total of three different models were trained using various architectures and hyperparameters.

The Inception Resnet v2 image classification model, loaded with weights pre-trained on Imagenet was used to extract features from the sample images [2]. A port of the DistilBert TAS-B Model to the sentence-transformers model, trained on MS MARCO corpus was used to extract sentence embeddings from the associated questions. The features from the image samples were concatenated with the sentence embeddings from the questions to form the input. Three different models were trained using this data.

Model 1 consists of 5 fully connected layers using the ReLu activation function with 1024, 512, 128, 64, and 32 nodes respectively, followed by an output layer using the sigmoid activation function. Model 2 consists of 5 fully connected layers using the ReLu activation function with 1024, 512, 128, 64, and 32 nodes respectively, followed by an output layer using the sigmoid activation function. A drop-out with a rate of 30% is applied to each of the fully connected layers of Model 2. Model 3 consists of 5 fully connected layers using the ReLu activation function with 1024, 512, 128, 64, and 32 nodes respectively, followed by an output layer using the sigmoid activation function. Batch normalization is applied to each of the fully connected layers of Model 3. RMSProp algorithm was used for optimization and the learning rate was set to the default value of 0.001. A binary cross-entropy loss function was

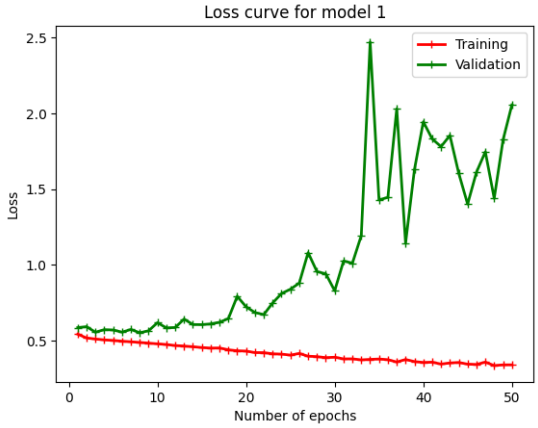

Figure 1: Loss curve for model 1

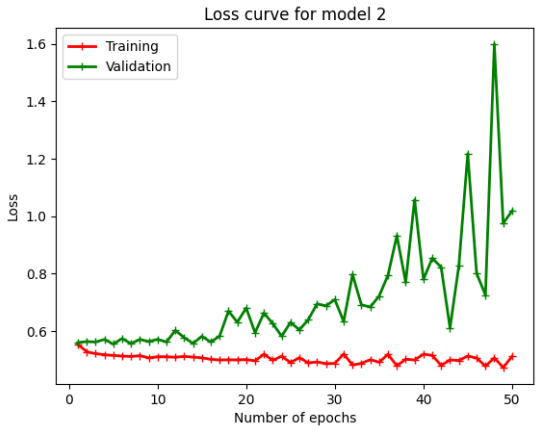

Figure 2: Loss curve for model 2

used to train all models. All the models were trained for 50 epochs using the Keras library for Tensorflow in Python. After training all three models, average precision was calculated for the training and validation data and the loss curves were plotted.

Results

Model 1 reported an average precision of 0.7837 for the validation set. Model 2 reported an average precision of 0.8288 for the validation set. Model 3 reported an average precision of 0.7449 for the validation set. The loss curves for Model 1, 2 and 3 were plotted as shown in figures 1, 2, 3.

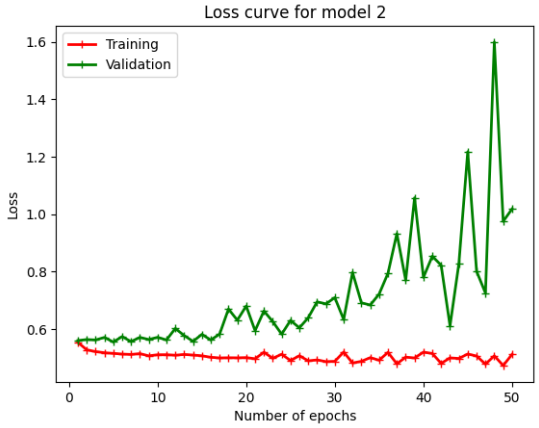

Figure 3: Loss curve for model 2

Analysis

From the loss curve of Model 1, we can see that the model is overfitting to the training data. The absence of any form of regularization could explain the increase in validation loss.

From the loss curve of Model 2, we can see that model 2 is also overfitting to the training data. Drop-out regularization applied in model 2 could have helped in reducing the overfitting compared to model 1.

From the loss curve of Model 3, we can see that model 3 is also overfitting to the training data. Batch normalization applied in model 3 resulted in the worst case of overfitting because the mini-batches might not have been representative of the entire training data.

Model 2 with higher average precision for validation is chosen as the best-performing model. Model 2 was used to predict the answers on 1000 samples from the testing dataset.

Testing Set

Methods

The dataset used in this experiment is the VizWiz-VQA dataset compiled by Gurari, Danna, et al. for VizWiz grand challenge [1]. The dataset was created with the help of blind people recording a question with respect to an image. 10 crowdsourced questions were also added to each image as well. The dataset consists of 20,523 training samples, 4319 validation samples, and 8000 test samples.

The experiment was conducted on a 14-inch MacBook Pro with an 8-core M1 Pro CPU. For this experiment, the best-performing model from the first part was chosen. The predictions on the first 1000 test samples are exported to a CSV file.

The Inception Resnet v2 image classification model, loaded with weights pre-trained on Imagenet was used to extract features from the sample images [2, **?**]. A port of the DistilBert TAS-B Model to the sentence-transformers model, trained on MS MARCO corpus was used

to extract sentence embeddings from the associated questions. The features from the image samples were concatenated with the sentence embeddings from the questions to form the input.

The best-performing model, Model 2 consisted of 5 fully connected layers using the ReLu activation function with 1024, 512, 128, 64, and 32 nodes respectively, followed by an output layer using the sigmoid activation function. A drop-out with a rate of 30% was applied to each of the fully connected layers.

Results

As we are using the same model (Model 2) from the above section, please refer to the above results section for loss curves and average precision.

Analysis

As the model used is same from the above section, please refer the analysis section above.

References

[1] Gurari, Danna, et al. "Vizwiz grand challenge: Answering visual questions from blind people." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[2] Szegedy, Christian, et al. "Inception-v4, inception-resnet and the impact of residual connections on learning." Thirty-first AAAI conference on artificial intelligence. 2017.