CSA 250 Deep Learning
Project-3

System Specifications

python Version=3.7.3
pytorch Version=1.3.1
torchnlp (pytorch-nlp) Version=0.5.0
torchtext Version=0.3.1

Observations

1.Logistic Regression

Logistic regression classifier is trained on SNLI dataset with following specifications.
Penality=L2
Solver= SAGA
Feature Extractor= TFID
The following results are obtained.

Dataset	Accuracy	Macro F1 Score
Training	0.5953	0.4462
Validation	0.5690	0.4297
Testing	**0.5646**	0.4268

Table 1- Accuracy and Macro F1 score for training, testing and validation dataset.

2.Logistic Regression hypothesis only

Logistic regression classifier is trained with hypothesis only model with same hyperparameters as logistic regression with both premise and hypothesis and the following results are obtained.

Dataset	Accuracy	Macro F1 Score
Training	0.6625	0.4971
Validation	0.6366	0.4810
Testing	**0.6352**	0.4802

Table 2- Accuracy and Macro F1 score for training, testing and validation dataset for hypothesis only model.

The accuracy of logistic regression hypothesis only model is significantly better than logistic regression model with both premise and hypothesis. It means that simple logistic regression model is underperforming. Therefore deep learining models are learned and their performance is examined below.

3. Deep learning models

1. Recurrent neural network-Recurrent neural network is learned on the dataset. The baseline model used for all the comparisons is given below.

Architecture of model
Layer 1- Embedding layer which converts a sparse high dimentional vector to a low dimentional vector(300).
Layer 2- Tranlating embedding vector(300d) to a hidden vector(300d) and add ReLU non-linearity.
Layer 3- Bidirectional RNN layer with 1 hidden layer
Layer 4-6- Fully connected layers with Relu non-inearity. Layers take 600d input and ouputs a 600d vector for each sentence. (300d premise and 300d hypothesis vector are concatenated to make 600d vector in layer 4)
Layer 7- Fully connected layer with 600d input and output a 3 dimention vector (1 for each class).

The baseline model uses Adam optimizer with learning rate of 0.001 and cross entropy loss funtion. These two are used in similar way in the susequent model also. The hyperparameters are tuned with random search, their values along with the results are in table given below.

The hyperparameters are tuned with random search. Model with minimum validation error is chosen for each run and

RNN S.NO.	Number of Trainable parameters	Fully connected layers	Hidden layers in RNN	Bidirectional	Glove vectors	Trainable embedding	Training Accuracy	Validation Accuracy	Testing Accuracy
1	4.7787 M	4	1	yes	no	no	0.7188	0.6939	**0.6915**
2	11.8485M	4	1	yes	no	yes	0.7465	0.7142	**0.7136**
3	8.9661M	**2**	1	yes	no	yes	0.7504	0.7200	**0.7152**
4	5.319M	4	**2**	yes	no	no	0.6936	0.6751	**0.6815**
5	1.354M	4	1	**No**	no	no	0.3329	0.3323	**0.3342**
6	4.7787 M	4	1	yes	**yes**	no	0.7301	0.7174	**0.7208**

Table 3- Tuning of hyperparmeters of RNN and accuracy obtained.

Obsevations- The following obsevations for the hypertuning of RNN model is observed.
1- Number of fully connected layer= RNN 2 and RNN 3 are models with 4 and 2 ully connected layers respectifully. RNN 3 is giving slightly better results with less number of trainable parameters.
2- Hidden Layers in RNN= RNN 1 is having 1 hidden layer and RNN 4 is having 2 hidden layers. While comparing both the RNN it is observed that RNN with 1 hidden layer is performing better. Less complex model is able to capture more imformation than more complex model.
3- Bidirectional- RNN 5 is a unidirectional model and it performs poorly. As there is one premise and there hypothesis statments for same premise, the unidirectional RNN model treats all three as similar and thus gives only 33%(approx) accuracy. Therefore bidirection model is preffered choice as it caputures context of entence better.
4- Pretrained Embedding- Pretrainned embeddings are providing better results as they are pretrained with similar words together in the embedding.
5-Trainable Embedding- Here if we are using pre-trained embedding then training it again will only decrease the perfromance and if we are not using any pretrained embeddings the training embedding wil capure contextual similarity between words.

Finallu using all the hyperparameters RNN is trained with Glove vectors, 1 hidden layer, 4 fully connected layers and 20 epochs. It gives testing accuracy of 0.7315.

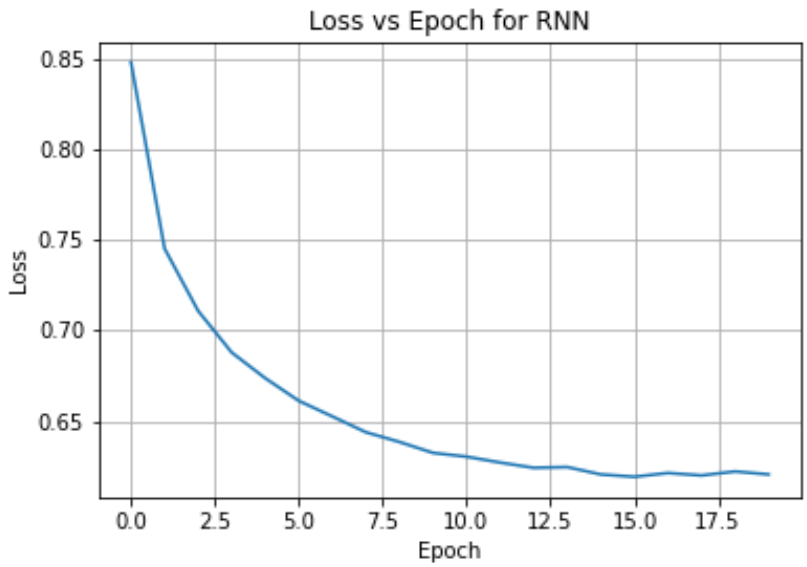

Figure 1= Training Loss vs number of Epochs for RNN.

2. Gated Recurrent unit-Keeping other hyperparameters same gated recurrent unit(GRU) is trained in place of RNN.
Number of Epochs=12 or Till optimal convergence
Loss function= Cross Entropy Loss function
Optimizer= Adam
Learning Rate=0.001

GRU S.NO.	Number of Trainable parameters	Fully connected layers	Hidden layers in GRU	Testing Accuracy
1	5.5M	4	1	**0.8076**
2	2.6187M	2	1	**0.8061**
3	4.2423M	2	2	**0.7914**

Table 5- GRU with different hyper-parameters and accuracy obtained.

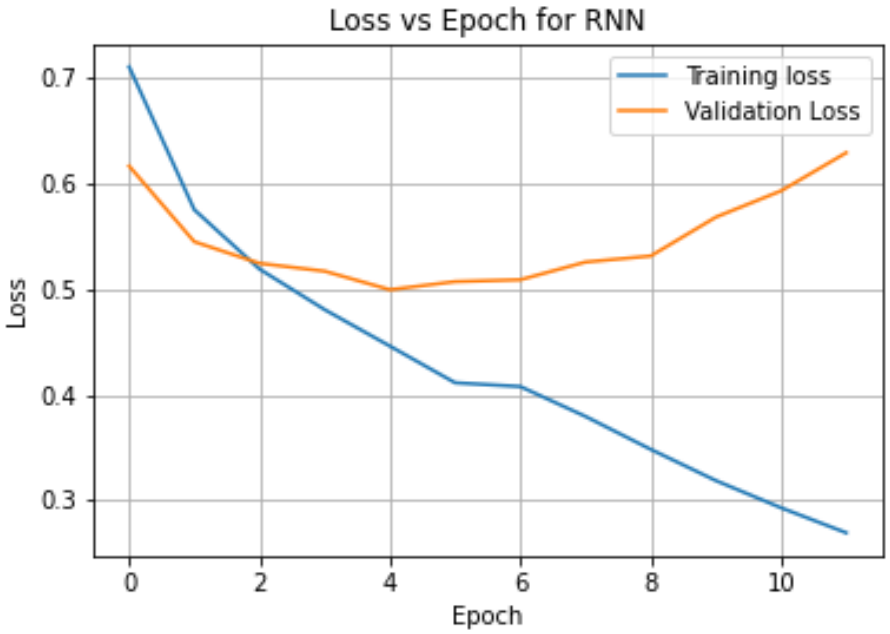

Figure 2= Training and validation loss of optimal GRU model with number of epochs.

The best GRU model with 4 fully connected layers with ReLU impurity and 1 hidden layer in GRU gives 80.76% accuracy.

3. Long and short term memory(LSTM)- Bidirectional LSTM is trained on dataset with with crossentropy loss function and Adam optimizer with learning rate 0.001. Glove vectors with 300 dimentional embeddings are used and by tuning various hyperparmeters the following results are obtained.

LSTM S.NO.	Number of Trainable parameters	Fully connected layers	Hidden layers in LSTM	Testing Accuracy
1	5.8623M	4	1	0.7963
2	2.9799M	2	1	0.8070
3	5.1447M	2	2	0.8010

Table 6- LSTM with different hyper-parameters and accuracy obtained.

From the above table, best results come with LSTM 2 which have 2.9799M hyperparameters, 2 fully connected layers and 1 hidden layer. In short, best accuracy come comes from least complex model.

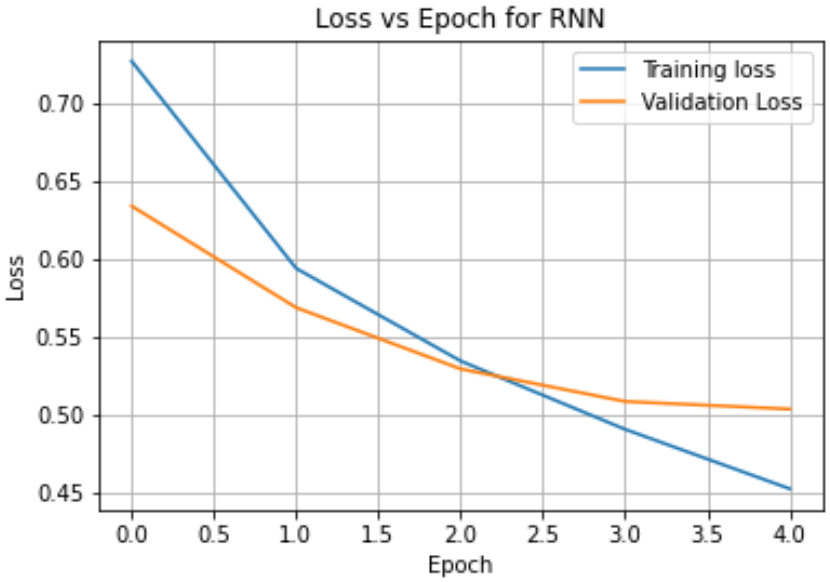

Figure 3= Training and validation loss of optimal LSTM model with number of epochs.

4. Final Models Selection

Logistic Regression with TFID=For Logistic regression hypothesis only model is seleted as it is giving better accuracy.

Deep learning model=For deep learning model, model with bi directional GRU with golve vectors is selected as it has least parameters and was giving best accurcay.

GRU S.NO.	Number of Trainable parameters	Fully connected layers	Hidden layers in GRU	Testing Accuracy
1	5.5M	4	1	**0.8076**

Table 7- Details of finally selected deep learning model.