RNN/CNN-based Natural Language Inference

Shubham Chandel - sc7238@nyu.edu

October 31, 2018

1 Introduction

Natural language inference (NLI) is the task of determining whether a given hypothesis can be inferred from a given premise. This problem can be posed as a classification problem, with the three classes being, neutral, entailment, and contradiction.

In this work, we to evaluate how well a CNN and a RNN based encoder along with a fully-connected head is able to perform this classification.

More specifically, we use the Stanford Natural Language Inference (SNLI) Corpus [1] to evaluate the performance of our model. We also use The Multi-Genre NLI (MNLI) Corpus [2] to evaluate how well the best model on SNLI is able to generalize to a different underlying data-distribution.

A comprehensive study is conducted and results are presented in the report and code is available here.
https://github.com/sksq96/nyu-nlp

2 Model

An encoder, either a CNN or RNN is used to map each string of text (hypothesis and premise) to a fixed-dimension vector representation. Further, both these representations are concatenated together and fed to a fully connected network, with a 3-class softmax.

2.1 Convolutional Neural Network

We use a CNN based encoder to encode both the input sentence into a fixed dimension latent representation. Specifically, we use 2-layer 1D convolutional network with ReLU activation.

Finally, we perform a sum over the time axis to, deal to map variable length hidden vector to fixed length.

2.2 Bi-directional Gated Recurrent Unit

We use a bi-directional GRU [3] to model the text as a sequence of timesteps, with each timestep indicating a token in the sentence.

Since a recurrent network brings along an inductive bias of causality, this proved to be helpful for the problem at hand, as shown in later sections.

3 Experiments

3.1 Dataset

We make use of a subset of The Stanford Natural Language Inference (SNLI) Corpus, which is a set of 100,000 sentence pair for training, and 1,000 for validation. Further, we use a subset of The Multi-Genre NLI (MNLI) Corpus, which is a set of 5,000 sentence pairs for validation.

Each example is a pair of two pieces of text, a premise and a hypothesis, and the correct label; entails, contradicts or neutral.

3.2 Experimental set-up

To evaluate the effect of different hyperparameters on the performance of our model, we fix an initial set of parameters presented in Table 1.

Once we have these set of parameters, we change one of these, keeping all other constant. The motivation behind this is to analyze what effect each of them have on the final model.

All the experiments are run for 10 epochs, with Adam optimizer [4] with learning rate of 1e-3.

encoder	hidden size	pretrained embedding	dropout	kernel size
Bi-GRU	300	fasttext.en.300d	0.2	3

Table 1: Initial set of Parameters

4 Results and interpretation

In this section, we look at detailed ablation study for each hyperparameter evaluated and a brief discussion surrounding each is presented to the readers.

4.1 Encoding Scheme

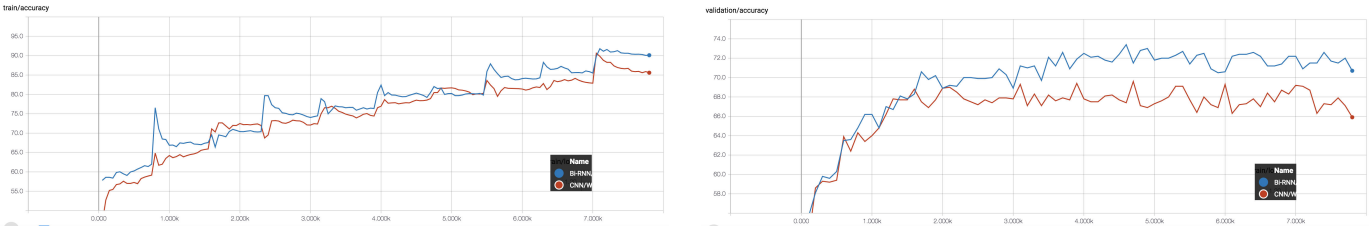
We train both a 2-layer CNN architecture and a bi-directional GRU model in this section. There is notable difference in the validation accuracy, as we can clearly see in Figure 1.

The Bi-GRU out-performs the CNN model, since it is able to capture much richer information, from both direction.

Further, we observe the CNN starts to overfit early on during training, as indicated by the decrease in validation loss. An important observation to note is, the CNN has fewer number of parameter, but still we observe signs of overfitting compared to the GRU. This suggests, the inductive bias of GRU is well suited for this task.

Encoder	Bi-GRU	CNN
$\mid \theta \mid$	7,127,203	6,524,203
Accuracy	72.27%	67.52%

Table 2: Number of trained parameters and validation accuracy for each model.



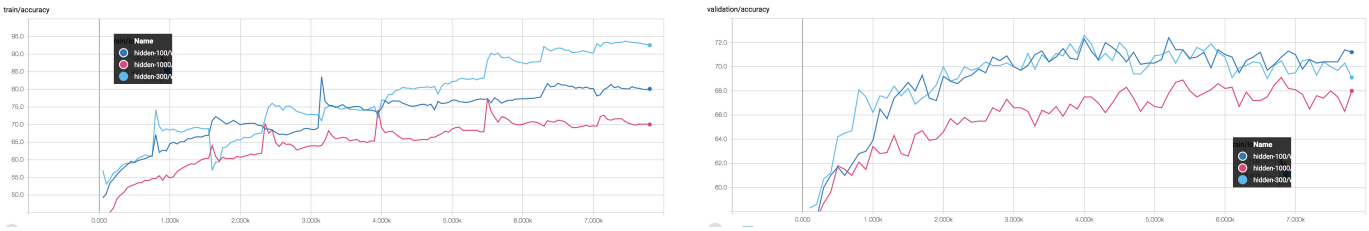
(a) (b)

Figure 1: Train and Validation accuracy as a function of timestep for CNN and GRU.

4.2 Hidden Dimension

As we expected, the network with larger hidden dimension of 300 is able to perform better both at training and validation, since it can store richer representation of the input sentences.

On the other hand, if we increase the hidden size to 1000, the network starts to perform worse. This is indicated in Figure 2, we see the network with larger hidden size had lower validation accuracy compared to those by 100, 300.



(a) (b)

Figure 2: Train and Validation accuracy as a function of timestep for different hidden size.

Size	h=100	h=300	h=1000
$\mid \theta \mid$	6,204,803	7,127,203	14,135,603
Accuracy	70.30%	71.42%	66.31%

Table 3: Number of trained parameters and validation accuracy for each model.

This experiment with different hidden sizes, indicates one of the core problem of encoding all the input sentences into a fixed length representation.

4.3 Regularization

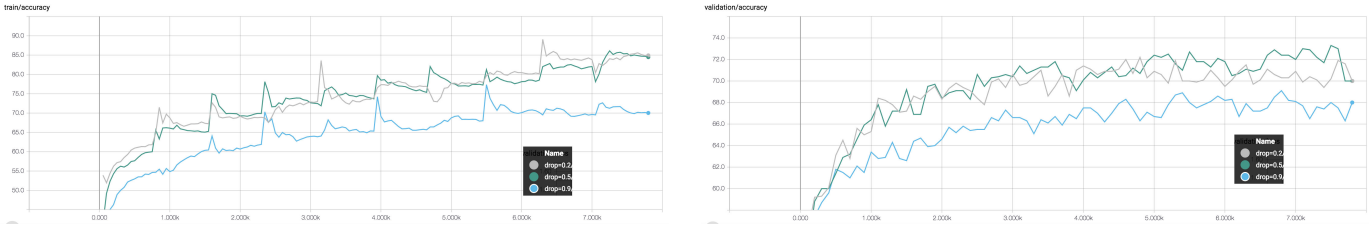
In this section, we explore the effect of regularization schemes. Specifically, we study the effect of dropout on the model.

As expected, a network with very high drop rate, ie. 0.9, is performing very poorly, both on validation and training, shown in Figure 3.

However, there exists a sweet spot around the drop rate of 0.5, where we achieve both high training and validation accuracy. This confirms the notion, in our case a decent amount of dropout is helping prevent the network to overfit on the training dataset.

Dropout	0.2	0.5	0.9
Accuracy	70.21%	73.30%	68.03%

Table 4: Validation accuracy for each model.



(a) (b)

Figure 3: Train and Validation accuracy as a function of timestep for different drop rates.

5 The Multi-Genre NLI

We explore transfer learning in this section, from the best validated model in the previous experiments. The aim is to see, how well our model is able to handle change of underlying data distribution.

Specifically, we use the trained model on SNLI corpus and evaluate it's performance on the MNLI courpus.

The validation scores across each genre is reported in the Table 5 for the best RNN model and Table 6 for the best CNN model.

fiction	government	slate	telephone	travel
49.55%	47.05%	42.51%	45.57%	46.95%

Table 5: Validation accuracy for different genere in the MNLI corpus for Bi-GRU.

fiction	government	slate	telephone	travel
47.24%	45.44%	41.54%	44.14%	44.09%

Table 6: Validation accuracy for different genere in the MNLI corpus for CNN.

The results evaluated on MNLI are about 20% less than those validated on SNLI. This change of underlying data-distribution reflects the real world challenges any Machine Learning model faces in production.

When we evaluated on the MNLI dataset, the data is no longer i.i.d., that is does not belongs to the same distributions as the training set, implying how brittle neural network, and machine learning models in general can be.

Further,the fiction genere has the highest accuracy and slate genere has the lowest accuracy. Fiction, by virtue of it, is found in day to day statements, and therefore is somewhat near to the original dataset.

6 Incorrect and Correct predictions

6.1 Incorrect predictions

1. Predicted: Neutral, True: Contradiction

   A small dog wearing a denim miniskirt. A dog is having all its hair shaved off.

Analysis: It is difficult to say if the premise entails or contradicts from the given statement. The model predicts neutral, which is likely correct in this case.

2. Predicted: Neutral, True: Entailment

   Two players are on a wet field and one is on the ground. There are only two people in the field.

   Analysis: The premise says no information about the total number of people in the field. The model again predicts neutral, which seems likely the correct answer in this scenario.

3. Predicted: Entailment, True: Neutral

   A teenage is on a surfboard. A teen is catching a big wave.

   Analysis: The model outputs incorrect answer, probably because it aligns surfboard with wave and thinks premise entails the hypothesis.

6.2 Correct predictions

1. Predicted: Entailment, True: Entailment

   A town worker working on electrical equipment. The worker is working.

2. Predicted: Contradiction, True: Contradiction

   People are walking into a store. The store is closed.

3. Predicted: Entailment, True: Entailment

   A boy is posing next to his scooter. A boy is next to a scooter.

References

[1] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning large annotated corpus for learning natural language inference https://nlp.stanford.edu/projects/snli/

[2] Williams, Adina and Nangia, Nikita and Bowman, Samuel A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference http://aclweb.org/anthology/N18-1101

[3] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation https://arxiv.org/abs/1406.1078

[4] Kingma, Diederik P and Ba, Jimmy. Adam: A method for stochastic optimization