Week 2

Oscar Morris

10 Jun 2022

1 Experiment 1

1.1 Aims

The first set of training was run to establish answers to the following questions:

1. Can outliers be detected by comparing multiple models?
    - are the outlier answers all the same, or do they change depending on the model?
2. How does the distribution of machine-given marks compare with that of the human markers?
3. How do models' predictions compare with each other (similar to question 1)?
4. Do the results found in the EE hold for more repeats and training samples?

1.2 Planned Method

Three models were going to be trained on the Tin Iodide synthesis dataset (approx. 100 samples) to determine differences, these models were the highest performing in the EE:

1. BERT (at the bert-base-cased checkpoint)
2. ALBERT (at the albert-base-v2 checkpoint)
3. Longformer (at the longformer-base-4096 checkpoint)

Each model was trained for 30 epochs with a batch size of 8. Tests were done evaluating the model on unseen data, and on training data.

1.3 Results

During training, the BERT model achieved similar values as in the EE, with r^2 around 0 and MAE around 0.1.

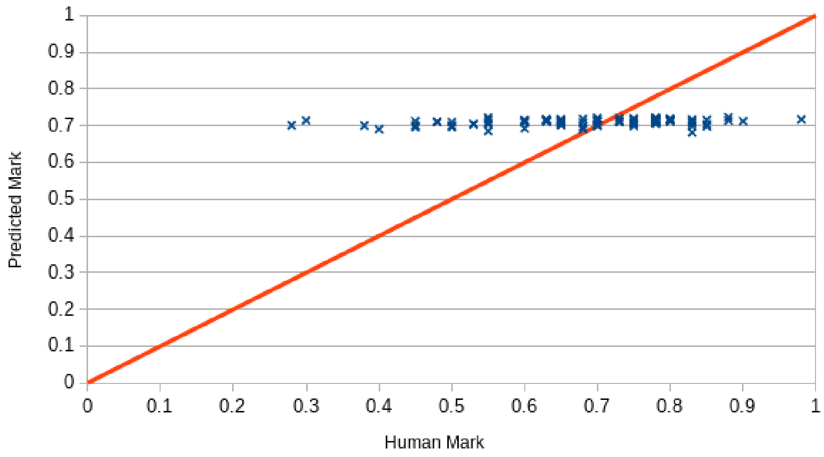

Figure 1: BERT predictions against Human markers on training data, Orange is the ideal solution

As can be seen from Fig. 1, the BERT model is always predicting approximately the same value, this value is approximately the mean of the human given marks (0.69). This is a common issue when training regression models, since most of the time it is better to predict the mean of the training data rather than randomly (and consequently learn the task). On re-examination of the data collected in the EE, it is very likely this this was also occurring in those experiments. This problem is known as "underfitting," either to the training data or to the evaluation data, in this case it appears that the model is underfitting to both splits as the distribution is very similar between seen and unseen samples.

1.3.1 Cause hypotheses

On the viewing of the data collected in this experiment, a few hypotheses were created for the causes (and potential solutions) to this problem:

1. The transformer-based models were too large to effectively learn from the data

2. The distribution of human-given marks in the training data was too unbalanced, leading to underfitting.

3. The AES task cannot be effectively learned on this dataset.

1.3.2 Cause Solutions

Possible solutions to each hypothesis are:

1. Use a smaller pre-trained transformer based on BERT [1], or create and train a small LSTM or self-attention based model

2. Testing on a more balanced dataset, or adjusting the loss metric to account for unbalanced data

3. Testing on a different dataset that has been proven to be effective in previous work.

Each solution has been tested on the Kaggle AES dataset [2] in further experiments.

2 Experiment 2

2.1 Aims

The second experiment aimed to determine the effectiveness of creating, and training a self-attention model from scratch.

2.2 Methods

The basic model architecture is shown in Fig. 2, with the architecture of the encoder shown in Fig. 3.

Each model was trained for 5 epochs, the model has a hidden LSTM size of 256 and an embedding length of 300.

The model was then pre-trained and evaluated on the Kaggle AES (Automated Essay Scoring) dataset [2] and the Kaggle SAS (Short Answer Scoring) dataset [3].

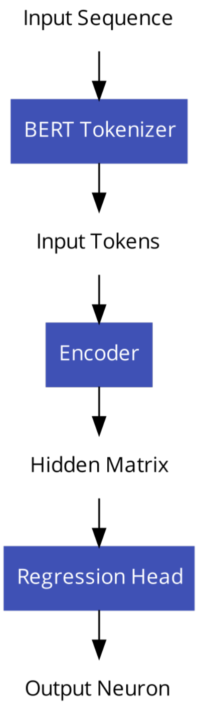

Figure 2: Basic regression transformer architecture

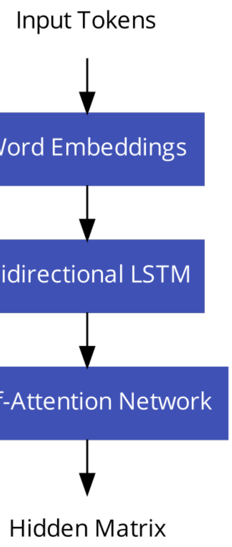

Figure 3: Encoder architecture

2.3 Results

The results for this model are very similar to the results for Experiment 1. This shows that the problem likely does not like with the model. However, due to the similar resuts to BERT, this model was used for all further Experiments due to the increase in control and data that a custom model gives.

3 Experiments 3-4

No further progress was made in Experiment 3.

In Experiment 4 the model was adjusted slightly, replacing the regression head with a classification head, this was to test the feasibility of a classification system rather than a regression system. Previous work has found it to be ineffective when compared to regression [4]. However, because of the difficulty in creating a system that allows a regression model to learn effectively it was still attempted. The model achieved approximately 0.6, almost double what would be expected if the model was guessing randomly, although the model is still not performing well. However, with further tuning and model improvements it is possible that this score could be significantly improved.

References

[1] I. Turc, M.-W. Chang, K. Lee, and K. Toutanova, “Well-Read Students Learn Better: On the Importance of Pre-training Compact Models,” Sept. 2019.

[2] “The Hewlett Foundation: Automated Essay Scoring.” https://kaggle.com/competitions/asap-aes.

[3] “The Hewlett Foundation: Short Answer Scoring.” https://kaggle.com/competitions/asap-sas.

[4] S. Johan Berggren, T. Rama, and L. Øvrelid, “Regression or Classification? Automated Essay Scoring for Norwegian,” in *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, (Florence, Italy), pp. 92–102, Association for Computational Linguistics, 2019.