Real or Not? NLP with Disaster Tweets

Team members:
Arsen Kuzmin
Ildar Yalalov
Nickolay Gaivoronskiy

Project overview

We are going to try to evaluate different NLP models on this Kaggle competition . In this competition, our task is to predict which Tweets are about real disasters and which ones are not.

Competition Description

Twitter has become an important communication channel in times of emergency.
The ubiquitousness of smartphones enables people to announce an emergency they're observing in real-time. Because of this, more agencies are interested in programatically monitoring Twitter (i.e. disaster relief organizations and news agencies).
But, it's not always clear whether a person's words are actually announcing a disaster.

What was done:

In this week we tried to apply BERT on this task. We wanted to try some ideas that were used in the original paper. We applied following ideas:

- *No pooling, directly use the CLS embedding*. The original paper uses the output embedding for the `[CLS]` token when it is finetuning for classification tasks, such as sentiment analysis. Since the `[CLS]` token is the first token in our sequence, we simply take the first slice of the 2nd dimension from our tensor of shape `(batch_size, max_len, hidden_dim)`, which result in a tensor of shape `(batch_size, hidden_dim)`.
- *No Dense layer*. Simply add a sigmoid output directly to the last layer of BERT, rather than experimenting with different intermediate layers.
- *Fixed learning rate, batch size, epochs, optimizer*. As specified by the paper, the optimizer used is Adam, with a learning rate between 2e-5 and 5e-5. Furthermore, they train the model for 3 epochs with a batch size of 32. We wanted to see how well it would perform with those default values.

Results:

We obtained the model that shows 87% accuracy on validation set and 82.5 F score on leaderboard.