Report

Dataset Description

This dataset consists of reviews of fine foods from amazon. The data span a period of more than 10 years, including all ~500,000 reviews up to October 2012. Reviews include product and user information, ratings, and a plain text review. It also includes reviews from all other Amazon categories.

Data Preprocessing

The TL;DR symbol which stands for too long; didn't read can be used as padding between review and summary to let the model know that the review ends here.

We also need to find average sequence length, in our problem average length is 85 so we can conclude that maximum length of 100 will cover the majority of instances.

Processing data samples may be difficult to maintain so pytorch dataset class is made as wrapper class to transform the text reviews to tensors.

Fine tuning GPT2

For training the model adam optimizer is used with learning rate of 3e-4 and batch size of 32.

```
def train(model, optimizer, dl, epochs):
    for epoch in range(epochs):
        print("epoch ", epoch)
        for idx, batch in enumerate(dl):
             with torch.set_grad_enabled(True):
                optimizer.zero_grad()
                batch = batch.to(device)
                output = model(batch, labels=batch)
                loss = output[0]
                loss.backward()
                optimizer.step()
                if idx % 100 == 0:
                    print("loss: %f, %d"%(loss, idx))
```

Then we start the training of the model.

```
epoch  0
loss: 7.385129, 0
loss: 2.681737, 100
loss: 2.593184, 200
loss: 2.391713, 300
loss: 2.372806, 400
loss: 2.466849, 500
loss: 2.257484, 600
loss: 2.303234, 700
loss: 2.547662, 800
loss: 2.183252, 900
loss: 2.396228, 1000
loss: 2.518579, 1100
loss: 2.010458, 1200
loss: 2.324232, 1300
loss: 2.476485, 1400
loss: 2.635899, 1500
loss: 2.239695, 1600
loss: 2.280631, 1700
loss: 2.188337, 1800
```

The loss decreased consistently which indicates that model is learning.

Using the fine tuned model

1) The model takes a input reviews first
2) Then from all the top k reviews, one is chosen
3) The choice is added to summary and current option is fed as input
3) Step 2 and 3 are repeated until the maximum sequence length is reached of EOS token is produced.