Incorporating BERT into Abstractive Text Summarization

Running Instructions:

Below are the instructions to setup a working environment and running the project.

Environment Setup

Create an anaconda environment using the following command

```
conda create -n <env_name> python=3.8
```

Activate the environment using

```
conda activate <env_name>
```

To install project dependencies, use

```
sh install_deps.sh
```

Changing Config file

The config file contains instructions to construct the datasets and models. configs/default_config.yml is the default configuration file and it is highly recommended not to meddle with the transformer portion of the file so it is easier to utilize training checkpoints.

Dataset:

CNN Dailymail dataset has been used to train the model. Please download the dataset directly from tensorflow and save it in .tfrecord format for easier consumption. To optimize input pipeline, training portion of the dataset has been sharded into 1000 files and placed in a directory. Once you have the dataset, please save them locally and specify their paths in the config file along with batch size and maximum length to consider for each training example.

Training

To initiate training, execute the following command

```
python train.py --num_epochs <num_epochs> --config_path configs/default_config.yml --logdir ckpts/
```

Inference:

```
python inference.py --config_path configs/default_config.yml --article_path <path_to_article>.txt --ckpt_dir ckpts/
```