TensorFlow NLP Modelling Toolkit

This codebase provides a Natrual Language Processing modeling toolkit written in TF2. It allows researchers and developers to reproduce state-of-the-art model results and train custom models to experiment new research ideas.

Features

- Reusable and modularized modeling building blocks
- State-of-the-art reproducible
- Easy to customize and extend
- End-to-end training
- Distributed trainable on both GPUs and TPUs

Major components

Libraries

We provide modeling library to allow users to train custom models for new research ideas. Detailed intructions can be found in READMEs in each folder.

- modeling/: modeling library that provides building blocks (e.g.,Layers, Networks, and Models) that can be assembled into transformer-based achitectures .
- data/: binaries and utils for input preprocessing, tokenization, etc.

State-of-the-Art models and examples

We provide SoTA model implementations, pre-trained models, training and evaluation examples, and command lines. Detail instructions can be found in the READMEs for specific papers.

1. BERT: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding by Devlin et al., 2018
2. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations by Lan et al., 2019
3. XLNet: XLNet: Generalized Autoregressive Pretraining for Language Understanding by Yang et al., 2019
4. Transformer for translation: Attention Is All You Need by Vaswani et al., 2017

Common Training Driver

We provide a single common driver train.py to train above SoTA models on popluar tasks. Please see docs/train.md for more details.

Pre-trained models with checkpoints and TF-Hub

We provide a large collection of baselines and checkpoints for NLP pre-trained models. Please see docs/pretrained_models.md for more details.