SENTIMENT ANALYSIS

Group 13

M11015802 Eng Tze Qian
M11015080 湯傑堯
M11015018 陳彥家

Outline

- Introduction
- Data Pre-processing
- Model
- Experiment
- Conclusion

Introduction

- Dataset and Explain

- Sentiment Labelled Sentences Data Set / ACL Imdb
  - It is the best movie I ever seen.
  - It is so boring.

- Dataset Split
  - Train : 80% | Val : 10% | Test : 10%
  - ACL IMDB / 50000
  - Amazon / 1000, IMDB / 1000, Yelp / 1000

Data Preprocessing

- Standard pre-processing techniques
  - Lowercasing
  - Stopwords Removal
  - Stemming and Lemmatization
  - Words embeddings (TF-IDF)

- Pre-processing Model
  - Tokenizer
  - Sequence Padding
  - Attention Mask

ULMFiT

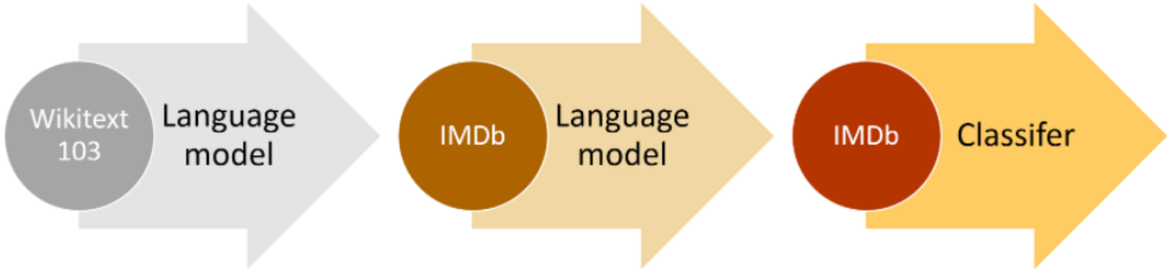
Why Universal Language Model Fine-Tuning for Text Classification (ULMFiT)

- Transfer Learning technique used in various NLP tasks
  - Has been state-of-the-art in NLP technique
  - Perform well even on small and medium datasets

ULMFiT

- Fine-tuning the pretrained language model
  - The language model is trained on wikitext-103
  - Wikipedia English is slightly different from the IMDb English
  - Whole model is freezed except the word embeddings

- Fine-tuning the classifier
  - Discriminative learning rate
  - Gradual unfreezing



XLNet

- Why we use this model ?
  - XLNet also incorporates the current optimal AR model Transformer-XL.
  - Overcome the shortcomings of BERT with its characteristics of AR.
  - Let the language model decompose sentences from sequential to random.

XLNet

Clean Text

Remove Punctuation, Numbers

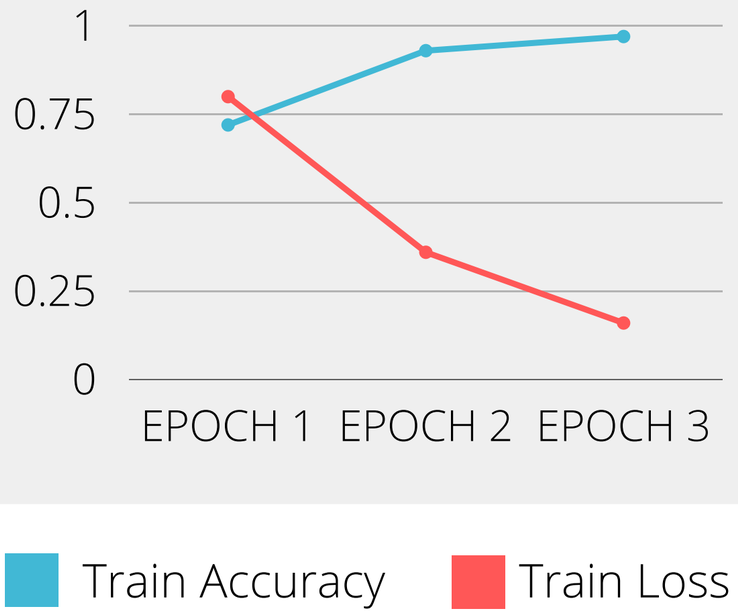
Pre Trained Model

Load Model from Google

Tokenizer & Encode

'Hello' → [830]

Training & Fine-Tune



Output & Predict

Experiment

- Sentiment Labelled Analysis

	Amazon	IMDb	Yelp	Combined
SVM	81.9%	74.5%	78.5%	81.5%
Naive Bayes	78.9%	78.1%	80.0%	82.2%
CNN	71.0%	81.0%	74.0%	75.0%
ULMFiT	77.8%	74.2%	76.8%	84.4%
XLNet	89.5%	95.5%	89.3%	92.7%

(Accuracy)

Experiment

- ACL IMDb

	Train	Val	Test
SVM	98.0%	87.0%	89.0%
LSTM	98.9%	80.9%	85.0%
NN	99.9%	86.2%	85.0%
CNN	98.1%	83.2%	84.0%
BERT	97.7%	93.9%	93.8%
ULMFiT	94.3%	94.1%	91.6%
XLNet	96.3%	94.1%	93.9%

(Accuracy)

Conclusion

- Transformer model (Bert, XLNet) is the current state-of-the-art approach
- XLNet has Best accuracy , but High cost.
  - For training 3000 datas wtih batch size 4 need up to 10 GB.
- Basic Neural Networks have serious over-fitting problem.
  - Epoch 1 has 61% accuracy and get 99% accuracy in Epoch 2.
- A well training model for IMDb may not be good in other environments.