A Bilingual BERT Model Ensemble for English-based Multitask Fine-tuning

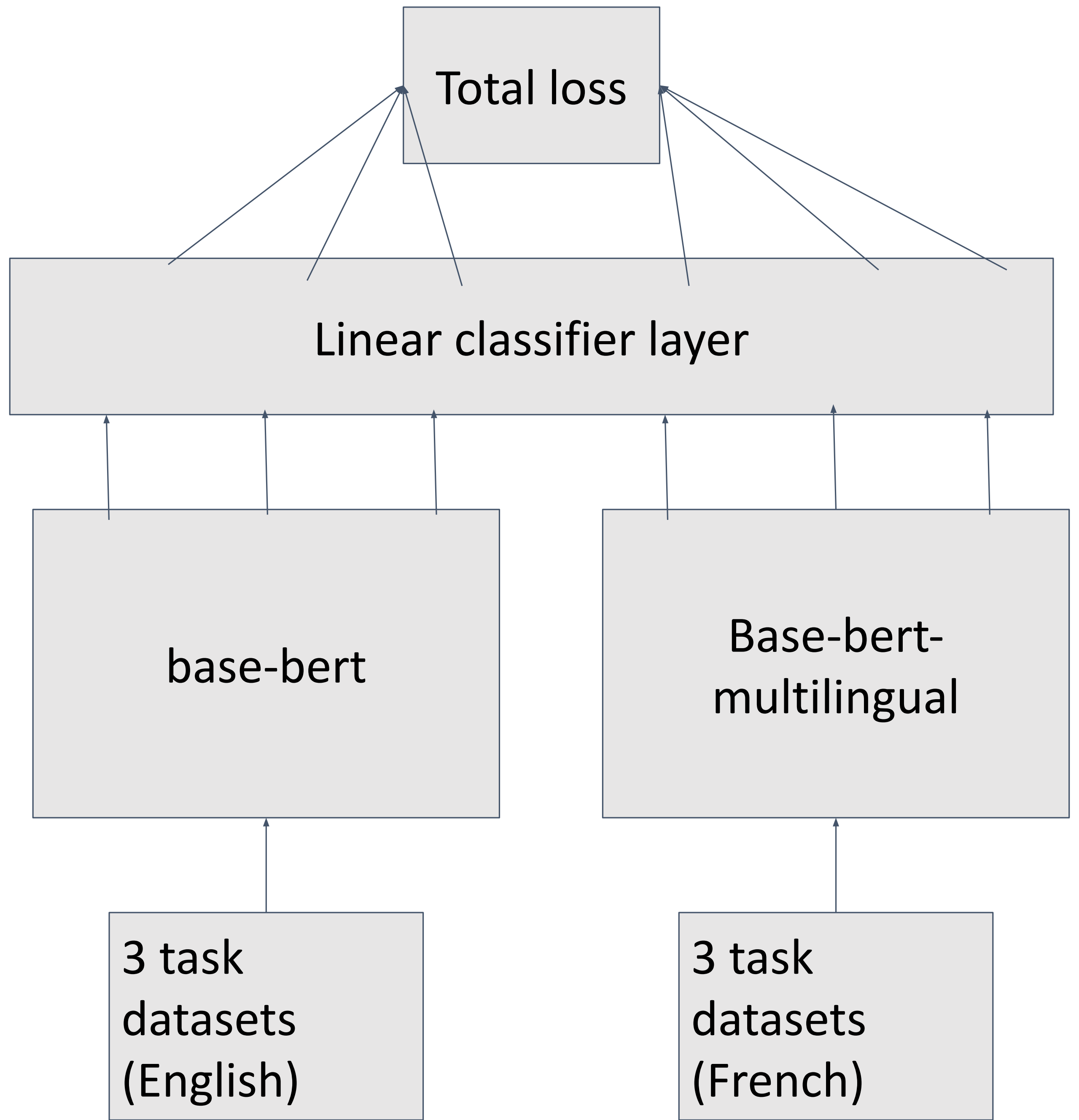
Ryan Dwyer, MS in Computer Science, Stanford University

CS 224N Default Final Project Extension, WInter 2024

Stanford
Computer Science

Project Overview

- Goal
  - Fine-tune a model to perform well on 3 English-based tasks:
    - Sentiment Classification
    - Paraphrase Detection
    - Semantic Similarity
- Theory
  - Transfer Learning through ensemble
    - Models can make up for each other's mistakes
  - Same task(s), different languages = diverse grammatical structures
    - Better understanding, better accuracy?
- Approach
  - English-pretrained + Multilingual-pretrained base BERT
    - English + French datasets, same 3 tasks similar language roots could help
  - Architecture/parameter tuning



Total loss

Linear classifier layer

base-bert

Base-bert-multilingual

3 task datasets (English)

3 task datasets (French)

Methods & Experiments

- Preliminary structure
  - Build underlying BERT structure (minBERT)
    - Pretrain/Finetune with SST and CFIMDB for sentiment (movie reviews)
- Extension Baseline
  - Just 3 main English datasets (no CFIMDB)
    - Each epoch loop through each dataset in batches
    - Sum together and average training loss across 3 tasks
  - Direct call to BERT layer in forward to get embedding
    - Followed by dropout (lower bias)
    - Followed by linear activation function to generate logits
      - Diff for sentiment (multilogit output) and paraphrase/similarity (single logit output)
- Multilingual Extension
  - English + French datasets
    - Same epoch loop format, just 3 more for French
      - Different preprocessing due to HuggingFace/dataset particularities
  - Bert-base-multilingual-uncased for French, bert-base-uncased for English
  - Run main dev file "multilingual.py" with finetune option

Discussions & Future Research

Discussions:

- Potential downfall of multitask modeling and model ensemble = gradient conflict
  - Transfer learning can occur, but if learning is sometimes not complimentary can actually cause harm
- However, methodology still shows potential if more measures taken to counteract downfalls/more powerful ensemble structure is used

Future Research:

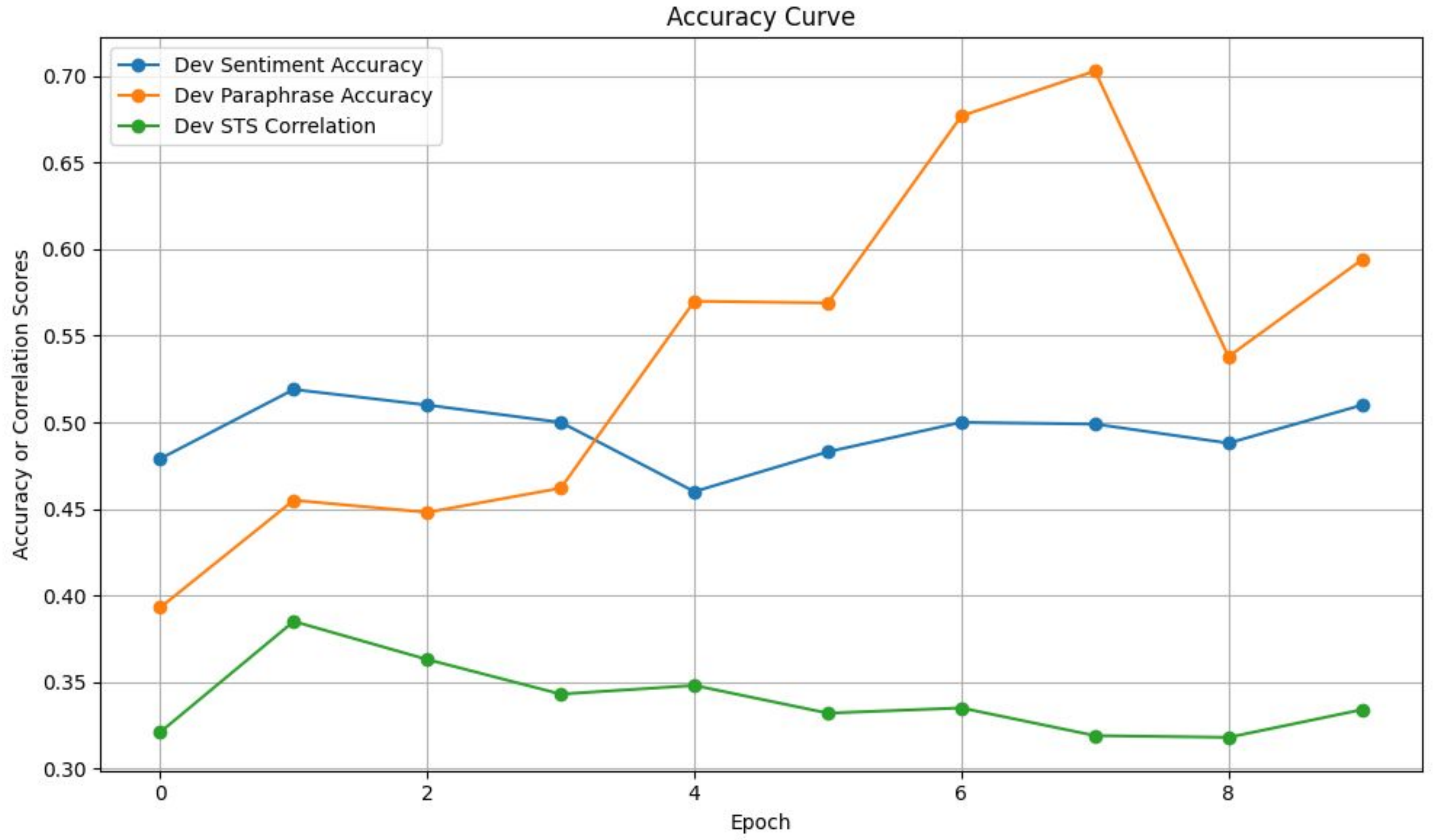
- With more compute, want to try more models (other papers used > 20) to ensemble to see more significant improvement
- Incorporate techniques like gradient surgery for better transfer learning, cosine embedding loss for better similarity comparisons, multiple rankings loss for others

Datasets & Architecture

- English datasets
  - Stanford Sentiment Treebank
  - Quora Paraphrase
  - SemEval Similarity
- French datasets (HuggingFace)
  - Book Review Sentiment
  - PAWS-X French Paraphrase
  - STS bank French similarity
- 1e-5 learning rate
- AdamW optimizer
- Cross-entropy Loss Function
- Batches of size 8

Results

- **Better SST performance than baseline**
- **Worse paraphrase/STS**



	Dev Accuracy			Test Accuracy		
	SST	Quora	STS	SST	Quora	STS
Baseline Fine-tune	0.477	0.753	0.347	0.476	0.755	0.284
Bilingual Fine-tune	0.510	0.467	0.334	0.526	0.466	0.2780