進捗報告

1　やったこと

- 大和さんの修論読みと内容理解
- T5 といった新しめのモデルについて調べた
- 大和さんの修論内容の実装 (途中です．すみません でした.)

2　BERT から得られる分散表現に対するアプローチについて考えよう

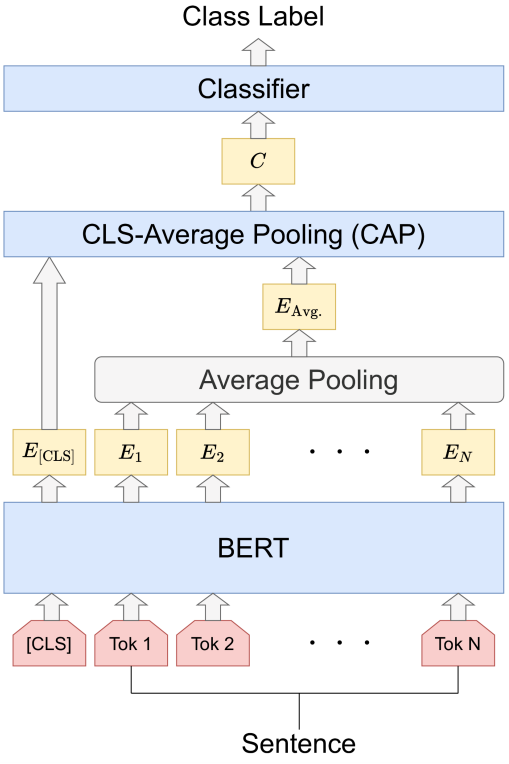

図 1: 大和さんモデル

大和さんの研究では, 文の分散表現のプーリング手法として一般的に用いられている 2 つの手法である [CLS] トークンに対応する分散表現を使う方法と [CLS] トークン以降の分散表現の平均を取る手法 (Average Pooling) を組み合わせることで分類精度がどうなるかについて述べている.

3　改善策 (方向性の提案)

3.0.1　[PAD] トークンの扱い

BERT では最大長に合わせてパディングされるため, 入力文の長さのばらつきによっては殆どの入力が [PAD] トークンとなることがあり得る．Average Pooling をする際に多くの [PAD] トークン分を加味することに合理性を感じられないと思ったため, Average Pooling 層に入れる前のデータに attention_mask をかけることで [PAD] トークンに対応する分散表現をゼロベクトルとして扱った方がいいのではと考えた.

3.0.2　CAP 層の付け替え

そもそも, BERT の出力をさらに LSTM などに突っ込むなど拡張性があるが, キリがない気がする.

3.0.3　入力文の次元削減

T5 や Chat GPT などを用いて入力文の次元数を段階的に減らす (要約文の生成) ことで分類精度と学習時間にどのような影響が出るか, といったアプローチができる.

4　やること優先順

- とりあえず動かせる状態にする (できれば水曜日までにやります　逐次報告します)
- 上の改善策を色々試してみる
- ポスター発表に向けて必要ならば改めてデータセットの選定や, 数値実験の課題を早急に決めたい

5　自戒用

これ以上逃げないように逃げ道を潰していく・コミュニケーションを絶やさない・とりあえずできることを全部やってから考える