C-02

BERT を用いた原文と要約文の分散表現の最適な統合手法の検討

創発ソフトウェア研究室　高山 裕成

研究背景

大規模言語モデル (Large Language Models, LLM) の進化

Transformer 構造を持つ BERT や
GPT (Generative Pre-trained Transformer) などを
活用した LLM の商業利用への需要拡大

⇒文章全体の適切な分散表現を得るための
プーリング戦略が重要

先行研究

CLS-Average Pooling (CAP) 層の導入 [大和, 2024][1]

学習可能なパラメータ $p,\ q\ (\geq 0)$ を用いて
$E_{[\mathrm{CLS}]}$ 及び, E_{Avg} の重み付き和

$$pE_{[\mathrm{CLS}]} + qE_{Avg}$$

を文の分散表現とする

⇒テキスト分類タスクにおいて, $E_{[\mathrm{CLS}]}$, E_{Avg} のみを
用いた場合よりも高い性能を発揮

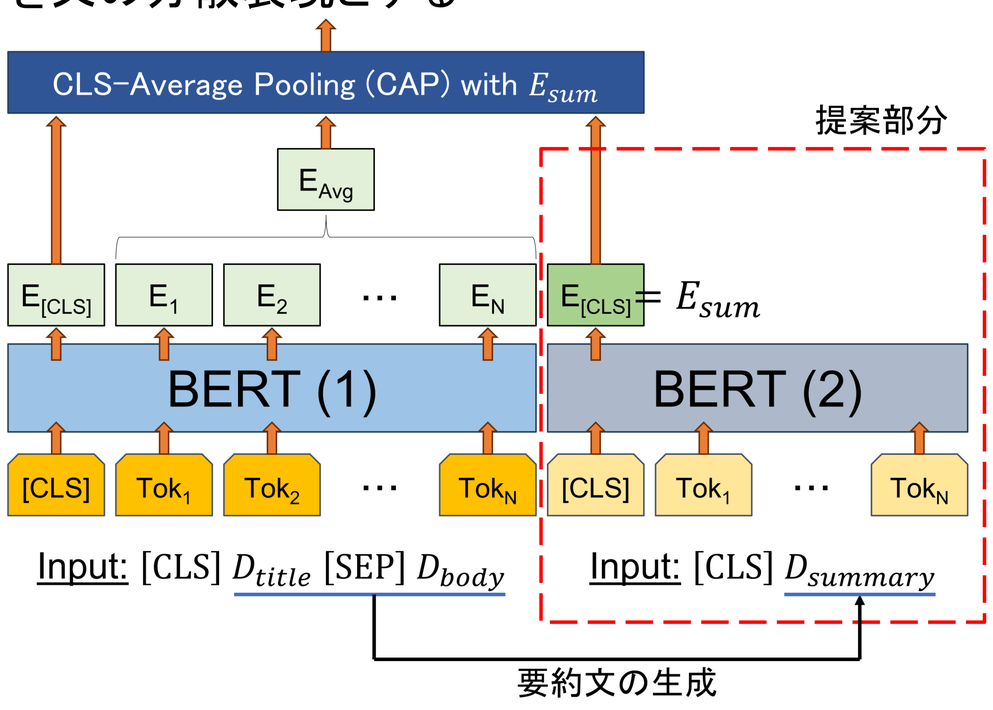
提案手法

CAP 層における要約文ベクトル項 E_{sum} の追加

学習可能なパラメータ $p,\ q,\ r(\geq 0)$ を用いて
$E_{[\mathrm{CLS}]}$, E_{Avg}, 及び E_{sum} の重み付き和

$$pE_{[\mathrm{CLS}]} + qE_{Avg} + rE_{sum}$$

を文の分散表現とする



1. 2 つの入力系列を独立した訓練済み日本語 BERT モデルへ入力
   日本語 BERT-base モデル (東北大学)
2. 各 BERT モデルの最終層の出力より $E_{[\mathrm{CLS}]}$, E_{Avg}, 及び E_{sum} を
   算出し, 提案手法であるプーリング層へ入力

実験設定

【livedoor ニュースコーパスデータセットを用いたテキスト分類】

データセットに含まれる各記事データ D は
記事タイトル D_{title}, 記事本文 D_{body},
カテゴリーラベル $D_{label} \in \{0, 1, \ldots, 8\}$ を持つ

・要約文 $D_{summary}$ の生成

$$\begin{cases} D_{summary} = \mathcal{L}(D_{prompt}) \\ D_{prompt} = \left[\begin{array}{l} \text{“}role\text{”} : \text{“system”}, \text{“}content\text{”} : \text{“あなたは優秀な AI アシスタントです。”} \\ \text{“}role\text{”} : \text{“user”}, \text{“}content\text{”} : \text{“以下は「}D_{title}\text{」というタイトルの} \\ \qquad\qquad\qquad\qquad\qquad \text{記事の本文です。内容を要約し、} \\ \qquad\qquad\qquad\qquad\qquad \text{結果のみを出力してください。”} \\ \text{“}role\text{”} : \text{“user”}, \text{“}content\text{”} : \text{“}D_{body}\text{”} \end{array} \right] \end{cases}$$

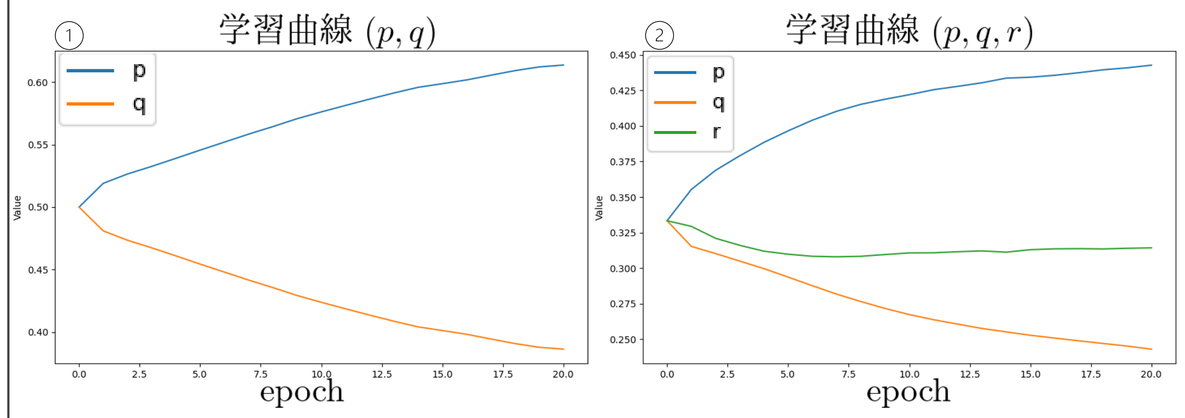
$\mathcal{L}$: 大規模言語モデル API (e.g., ChatGPT API, PLaMo API)

学習パラメータ	値
エポック数	20
バッチサイズ	16
最適化手法	Adam
学習率	3.0×10^{-5}
損失関数	Cross Entropy Loss
分類器	全結合層

	重み初期値
従来手法	$(p, q) = (\frac{1}{2}, \frac{1}{2})$
提案手法	$(p, q, r) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

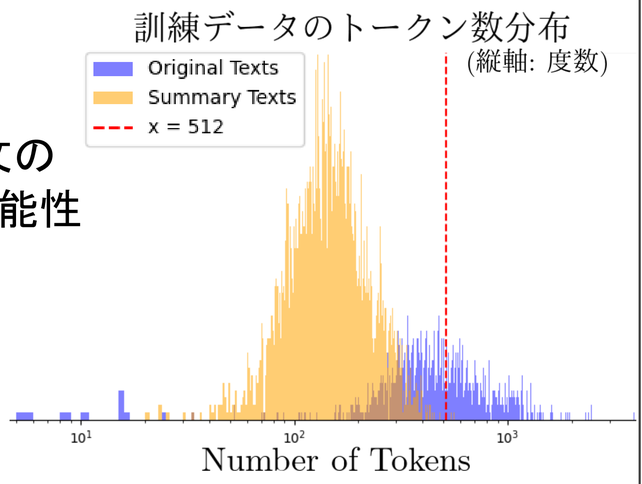
実験結果

	Accuracy	F1 (weighted)
① 従来手法 (CAP)	0.9565	0.9560
② 提案手法 (CAP with E_{sum})	<u>0.9687</u>	<u>0.9683</u>



- 提案手法の有効性を確認
- E_{sum} による, 欠損した原文の
  重要な文脈情報の補完可能性

	トークン数	
	≥ 512	< 512
Original	3747	2148
Summary	11	5884



今後の課題

- 要約文の妥当性, 生成手法改善の検討
- 他のデータセットを用いた提案手法の有効性の確認
- 最適な学習パラメータやアーキテクチャの探索

[1] 大和秀徳. BERT の分散表現に対する学習可能な重みを持つプーリング手法の提案,
大阪公立大学大学院情報学研究科 基幹情報学専攻 知能情報分野 修士論文, 2024.