概要

現在の大規模言語モデル (Language Models; LLM) は、高い性能を有する一方で、モデルのパラメタ数が膨大であるため、計算資源の限られた環境においては利用が困難である。このため知識蒸留と呼ばれるモデル圧縮手法が注目されている。知識蒸留は、大規模な言語モデル (教師モデル) と未学習の小規模モデル (生徒モデル) を用意し、訓練データのサンプルに対する教師モデルの内部状態を模倣することで、教師モデルの性能を維持しつつモデルのサイズを小さくする手法である。

一方、LLM による推論の過程を説明あるいは可視化することも重要である。例えば、推論時にモデルが注意している単語を示すことは、モデルによる推論の過程を説明するものとみなせる。しかしながら、生徒モデルが教師モデルと同じ単語に注意して推論を行っているとは限らないことが先行研究により指摘されている。これは生徒モデルの説明能力が教師モデルと比べて低下している可能性があることを意味する。金融分野や医療分野など、精度だけではなくモデルの推論根拠の提示も重要な分野においては、知識蒸留モデルの説明能力の低下は問題となりうる。

本研究は、下流タスクに対する性能を維持するだけでなく、推論時にどのような単語に注意するかといった教師モデルの特性も転移する知識蒸留の手法を提案する。既存の知識蒸留の手法を拡張し、教師モデルと生徒モデルが注意する単語を定量化し、これを新しい損失項として組み込むすることで、教師モデルの推論性能と説明能力の両方を継承する生徒モデルを学習することを狙う。

まず、教師モデルと生徒モデルの注意単語の違いを検証する実験を行った。自然言語処理のベンチマークデータセット GLUE を用い、BERT を教師モデルとし、8 つのタスクに対して既存知識蒸留手法である TinyBERT を用いて生徒モデルを学習した。教師モデルと生徒モデルのそれぞれについて、Integrated Gradients (IG) を用いて入力における各単語の重要度を計算し、これが大きい単語をモデルの注意単語として抽出した。検証の結果、同じ入力に対しても生徒モデルと教師モデルとで異なる単語に注意して推論を行っている傾向が強いことが確認された。また、BERT の Encoder レイヤー数を削減した小規模モデルである BERT-Medium、BERT-Mini、BERT-Tiny をファインチューニングしたモデルと比べて、TinyBERT による生徒モデルはタスクを解く能力が高い一方で、教師モデルとの注意単語の一致率は低くなることが確認された。

これを踏まえ、本研究では、教師モデルと生徒モデルの注意単語の一致率を高めるための新たな知識蒸留手法を提案する。具体的には、生徒モデルを学習する際、教師モデルが注意する単語に対して生徒モデルも高い注意を払うよう損失関数を設計する。まず、先に述べた分析と同じように、IG を用いて単語の重要度を計算する。次に、Softmax 関数を用いて入力における単語の重要度スコアの和が 1 になるように正規化し、単語の重要度の確率分布 (以下、「重要度分布」と呼ぶ) を得る。生徒モデルを学習する際には教師モデルと生徒モデルの重要度分布の差

を最小化する。本研究では Soft Jaccard 係数を用いて両モデルの重要度分布の類似度を測り、1 から Soft Jaccard 係数を引いた値を新しい蒸留損失項として定義する。この損失項と既存手法の TinyBERT における蒸留損失項の重み付き和を新しい損失関数として定義する。この損失関数を最小化するようにパラメタを更新することで、教師モデルの内部状態だけではなく推論時の注意単語も模倣するよう生徒モデルが学習される。

提案手法の評価実験を行う。教師モデルは BERT(Transformer 層が 12 層)、生徒モデルは事前学習時に知識蒸留されたモデルである TinyBERT-6L(6 層) または TinyBERT-4L(4 層) とし、GLUE の 8 つのタスクを対象に生徒モデルをファインチューングする。ベースラインの知識蒸留手法を TinyBERT とし、提案手法と比較する。評価指標として、タスクの性能の評価指標 (正解率など) に加え、注意単語の教師モデルとの一致度を測るために Jaccard 係数と Ranking 指標を用いる。Jaccard 係数は IG によって上位 K 個の重要単語を抽出したとき、その重要単語集合の Jaccard 係数である。一方、Ranking 指標は上位 K 個の重要単語が順位も含めて完全に一致している (1) か否か (0) を表す。モデルの評価時にはテストデータのサンプルに対する Jaccard 係数ならびに Ranking 指標の平均値を算出する。また、K の値を 1 から 10 まで変動させ、それぞれの値ごとに指標を算出する。

実験の結果、TinyBERT と提案手法とで下流タスクに対するモデルの性能に大きな差はなかった。これは本研究で提案する損失項を組み込んでもモデルの性能が大きく損われることがないことを意味する。Jaccard 係数については全てのタスクと K の値において改善が見られた。特に有効だったのは SST-2 タスク (感情分析タスク) であり、TinyBERT-6L については K の値によって 20〜24 ポイント、TinyBERT-4L については 12〜19 ポイント程度の改善が見られた。CoLA タスク (文の妥当性判定タスク) ではベースラインでも Jaccard 係数が高く、Top-9 では 0.86 以上であったが、提案手法では更に 3 ポイント向上した。Ranking 指標については、一部のタスクと K の組について提案手法がベースラインと比べて低かったが、全体的にはベースラインを上回った。Rankinig 指標は重要度スコアの順序も考慮するため、Jaccard 係数と比べて厳しい評価指標である。したがって全体的に Rankinig 指標の値は低かったものの、K が 1〜3 の場合については 10 ポイント程度の改善が見られるケースもあった。以上から、同じような単語を同じような順序で注意する能力を教師モデルから継承するという点での提案手法の有効性が確認された。