Attention is Not Only a Weight: Analyzing Transformers with Vector Norms

小林 悟郎[†]

1 はじめに

本稿では、EMNLP2020 に採択された筆者の論文 “Attention is Not Only a Weight: Analyzing Transformers with Vector Norms” (Kobayashi et al. 2020) について、分野の背景および今後の展望を交えて解説する。論文の解説に入る前に、まずは論文採択に至るまでの研究の過程について簡単に述べたい。

2 採択に至るまで

研究の発端は6月、研究室内で行われている深層学習に関する最新論文読み会にて1本の論文 (Reif et al. 2019) と出会ったことだった。この論文は多様なタスクで成功を収める言語モデル BERT (Devlin et al. 2019) が暗黙的に言語的性質を獲得しているかどうかを分析・検証する、いわゆる probing と呼ばれる研究に分類される。当時の私は研究室の新人研修の一環で BERT 論文の精読をしていたこともあり、その挙動や性質の解明に強い興味を持った。読み会で隣に座っていた先輩の栗林さん（第二著者）および横井さん（第三著者）と論文の内容について議論が盛り上がり、その流れで本研究がスタートした。

知らない用語だらけの英語論文を読んだり、先輩方の議論についていけなかったりと、初めての研究活動に戸惑いながらもサーベイや実験を行い、8月のNLP若手の会 (YANS2019) で中途段階の研究成果についてポスター発表を行った。たくさんいただいたフィードバックの中でも、「この研究の先にどんな嬉しい未来があるのかをよく考えるべき」とのコメントが本研究の意義や立ち位置を見つめ直すきっかけとなり、その一環として行ったサーベイを通じて採択論文の核となるアイデアにたどり着いた。3月の言語処理学会年次大会 (NLP2020) では論文を投稿し、口頭発表を行った。光栄にも最優秀賞という大変競争的な賞をいただくことができ、本研究の価値についての自信に繋がった。

NLP2020 への論文投稿を終えた1月から、同内容での国際会議への投稿を本格的に目指し始めた。初めての英語論文執筆ということを鑑み、本番の EMNLP2020 投稿前に論文の内容や英

[†] 東北大学

語表現についてコメントをいただくため、ACL Student Research Workshop (SRW) に、EMNLP とは二重投稿にならない形式 (non-archival) で投稿した。とても丁寧に査読していただいただけでなく、海外研究者の方がメンターとして論文や発表準備にアドバイスをくださった。また、SRW への投稿によって早い段階で英語論文の形が出来上がったことも論文の完成度向上に繋がったように思う。

振り返ってみると、国内から国外まで、論文への査読や発表へのコメント一つ一つが EMNLP 採択への大きな後押しとなったと思う。各会議の運営の皆様、査読者の皆様、当日質問・議論をしてくださった皆様にこの場を借りて心より感謝申し上げたい。

3 採択論文の概要

続いて、採択論文の解説に移る。

3.1 背景

近年の自然言語処理では、深層学習による進歩がめざましい。特に BERT を代表とする、Transformer (Vaswani et al. 2017) をベースとしたモデルが幅広いタスクで成功を収めている。その一方で、これらの巨大なニューラルネットワークのモデルは基本的にブラックボックスであり、望ましい出力をしていたとしてもそれに至った過程を知ることはできない。これは、システムとして実用する上での信頼性の観点で大きな問題となりうる。また、モデルの改善を試みる際に、新たな工夫を考えるための判断材料が限られてしまう。そこで本研究は大きな成功を収める Transformer モデル内部の挙動を明らかにすることに焦点を当てている。

Transformer モデルの成功要因の解明や更なる改善に向け、その性質や挙動を分析・検証する試みは盛んに行われている。モデル内部の挙動を調べる既存研究では、巨大な Transformer モデルに対して「アテンション重み」という非常に小さなスコープでの分析が典型的に行われてきた。Transformer は注意機構による「情報の混ぜ合わせ」を核とした構造をなしている。例えば言語モデルでは周囲の単語情報を、機械翻訳タスクでは原言語側の単語情報を混ぜ合わせて単語表現を更新していく。既存研究はこの注意機構による混ぜ具合に着目し、混ぜ合わせ時に割り振る重み（アテンション重み）を観察してきた。この方法は巨大なモデルを分析するにあたって、注意機構という一部にのみ注目して、他の大部分を無視してしまっている。さらに、その注意機構内でもアテンション重みは一部の作用に過ぎず、注意機構内の混ぜ具合ですら正しく反映できていない可能性がある。モデルのほんの一部にしか注目せず、スコープの小さい分析が行われている現状を問題視し、本研究ではモデル内の他の部分も考慮した、スコープの大きい分析方法を確立し、内部挙動の解明を目指す。採択論文ではその第一歩として、注意機構内で無視されてきた作用を考慮することで注意機構の混ぜ具合を精緻に分析する方法を提案した。

3.2 提案

まずは、注意機構の混ぜ具合を分析するにあたって、アテンション重みのみを観察する既存の方法が起こしうる問題について述べる。注意機構は入力ベクトルの変換と変換されたベクトルの重み付き和により、入力を混ぜながらベクトルを更新する（図1）。この重み部分に着目して混ぜ具合を観察することは一見自然な方法に思える。しかし、重み付けされるベクトルが全く考慮されておらず、入力ベクトルと変換による作用が無視されてしまっている。よって、小さいベクトルに大きな重みを割り振ったり、大きなベクトルに小さな重みを割り振るような場合には、重みの観察からは実際の混ぜ具合とは大きく違った分析が得られ、誤った認識へ導いてしまう恐れがある。

本研究では、無視されていた入力ベクトルと変換の作用も考慮に入れた、ベクトルノルムに基づく分析方法を提案した。方法はとても単純で、変換して重み付けした後の実際に足されるベクトルに着目し、そのベクトルの長さを注意機構が混ぜた量として測るというものである。

3.3 実験

実験では、アテンション重みを観察する既存の分析方法とベクトルノルムに基づく提案した方法で、注意機構の混ぜ具合の見え方がどのように変わるのかを BERT と機械翻訳モデルで調査した。既存手法では、特殊トークンや文の終端トークンなどのあまり意味的情報を多く含まないと思われる部分に大きな重みを割り振る傾向が観察されていた（図2）。一方で提案手法を使うと、それらから足されるベクトルは大きくないことが分かり、重みを大きく割り振りなが

図1 注意機構の概略図。既存研究はアテンション重み（青色部）を観察しており、採択論文では実際に足されるベクトル（赤色部）に注目した分析を提案する。

図2 事前学習済み BERT base の層ごとのアテンション重みの傾向 (既存手法による結果). 特殊トークンである [CLS] や [SEP] に高い重みが割り振られている.

図3 事前学習済み BERT base の層ごとのベクトルノルムの傾向 (提案手法による結果). [CLS] や [SEP] から足されるベクトルは特に大きいわけではない.

らも実際にはあまり強く混ぜていない挙動を明らかにした (図3). 論文内では, 手法間の結果の違いに関する詳細な分析や, あまり混ぜないベクトルに対して大きな重みを割り振る理由についての考察も行っている.

3.4 今後の展望

提案したベクトルノルムに基づく手法によって, Transformer モデル内部の各注意機構の混ぜ合わせを精緻に分析することが可能となった. しかし先述した通り, 注意機構は「混ぜ合わせ」という核となる機能を担うが, その一つ一つは巨大なモデルから見るとほんの一部の作用に過ぎない. 特に各層には, Feedforward 層や残差結合 (Residual connection) などの作用が存在し, 注意機構での混ぜ具合は上書きされうる. そこで次のステップとして, 分析のスコープを層全体まで広げ, 層内の様々な機構を考慮した分析方法の確立を目指している.

さらに, モデル入力の混ざり具合は前の層までの混ざり具合を次の層へと引き継いでいくため, 更に分析のスコープを広げることを考えると, 前の全ての層を考慮した分析が必要となる. 最終的には Transformer の入力から出力までの情報の流れを追跡し, 「どんな情報がどこで多く混ぜられたか」を調査できるような分析の実現を見据えている. 具体的には, 「各種タスクに fine-tuning した BERT がどんな単語に注目して高性能を発揮しているのか」などを明らかにできる未来を考えている. これが実現すれば, Transformer モデルの挙動を説明する方法として使うことや, Transformer モデルを改善に導くような知見を獲得することが期待できる.

4 おわりに

幸いにも、採択論文で提案した方法は既にいくつかの関連研究で使われ始めている。一例として、同じく EMNLP2020 に採択された論文 (Prasanna et al. 2020) では、各種タスクに fine-tuning した BERT においても、特殊トークンに大きな重みを割り振るが大きなベクトルは足し合わさらないという現象が起きていることを示唆する結果を示している。

最近、国内外へ研究を発信する目的で Twitter のアカウントを作成し、日英両方での投稿を始めた。EMNLP 開催以降、海外の複数の研究者から我々の論文を紹介するツイートをしてもらえるなど、少しずつ本研究が認知され始めていることを実感している。

このような国内外からの反応が大きな励みとなっている。今後も Transformer モデル内部の挙動解明という目標の達成に向け、引き続き精力的に研究を進めていきたい。

参考文献

- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171–4186.
- Kobayashi, G., Kuribayashi, T., Yokoi, S., and Inui, K. (2020). “Attention is Not Only a Weight: Analyzing Transformers with Vector Norms.” In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7057–7075.
- Prasanna, S., Rogers, A., and Rumshisky, A. (2020). “When BERT Plays the Lottery, All Tickets Are Winning.” In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3208–3229.
- Reif, E., Yuan, A., Wattenberg, M., Viegas, F. B., Coenen, A., Pearce, A., and Kim, B. (2019). “Visualizing and Measuring the Geometry of BERT.” *Advances in Neural Information Processing Systems 32 (NIPS)*, pp. 8594–8603.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). “Attention is All you Need.” In *Advances in Neural Information Processing Systems 30 (NIPS)*, pp. 5998–6008.

略歴

小林 悟郎：2020 年東北大学工学部電気情報物理工学科卒業。現在、東北大学大学院情報科学研究科博士前期課程学生。