パターン認識
レポート課題

井上智裕

2021 年 1 月 16 日

1 選んだ論文

Alex Tamkin, Dan Jurafsky, and Noah Goodman. Language through a prism: A spectral approach for multiscale language representations, NeurIPS, 2020, https://arxiv.org/abs/2011.04823

2 導入

2.1 扱う問題

自然言語処理分野。特定の言語モデルに頼らない、自然言語におけるスケールの異なる構造（語彙、節、文書など）の発見・学習。

2.2 問題意識

単語の意味など語彙レベルの構造、節や文レベルでの構造、文書全体の主題構造や物語構造など、言語は異なるレベルでの構造を有している。従来の研究では、単語の分散ベクトル表現の学習 [1] や文の分散表現の学習 [2] といった個々のレベルでの構造を捉えるためのモデルの構築など、異なる階層の言語構造のそれぞれを明示的にモデル化する方法が示されてきた。そんな中、近年では様々なタスクに対応するための汎用事前学習モデルが作られるようになったが、これらのモデルでは特定の階層の情報を特定のニューロンが担うのではなく、ニューロン全体に情報が分散している。この論文では、従来手法のように文や節といった特定の構造レベルの言語モデルに頼ることなく、BERT[3] のような汎用モデルから各スケールの構造を抽出・学習するための方法を提案している。

3 理論

3.1 スペクトルフィルタ

言語における異なるレベルの構造の発見にあたり、信号処理分野などで広く使われているスペクトル分析の手法を導入する。入力系列に対し周波数領域で演算を行うためには、入力表現を周波数領域での表現に変換する必要がある。このスペクトル変換には、離散コサイン変換（DCT）を使用する。実数列 $\{x^{(0)}, \cdots, x^{(N-1)}\}$ に対して、DCT（各周波数の重み）は次式で得られる。

$$f^{(k)} = \sum_{n=0}^{N-1} x^{(n)} \cos\left[\frac{\pi}{N}(n+\frac{1}{2})k\right] \quad k = 0, \cdots, N-1 \tag{3.1}$$

この DCT を言語表現に適用するにあたり、文脈に応じた単語表現を対象とする。文脈に応じた単語表現とは、トークン（単語やサブワード単位）の入力系列に対し、自然言語処理モデルが各入力の前後の文脈を考慮に加えて処理することで生成されるベクトル系列である。文脈に応じた単語表現 $v_0, \cdots, v_{N-1}$ が与えられたとき、単語表現のあるニューロン（次元）i に沿った表現 $v_0[i], \cdots, v_{N-1}[i]$ に DCT を適用し、得られる系列

$f_0[i], \cdots, f_{N-1}[i]$ を第 i ニューロンのスペクトルとする。

この手法では、ニューロンのスペクトルに対して、特定の閾値を持つスペクトルフィルタを通すことによる特定成分の除去と逆離散コサイン変換（IDCT）による元のドメインへの変換を行うことにより、系列から特定のスケールの構造を取り出すことができる。この研究では、入力サイズを 512 として、スケールに応じて表 3.1 に示す 5 種類のフィルタを用意した。

表 3.1　スペクトルフィルタ

フィルタ	対応する言語構造	周期 [tokens]	DCT index k
HIGH	単語	1 – 2	130 – 511
MID-HIGH	節	2 – 8	34 – 129
MID	文	8 – 32	9 – 33
MI-LOW	段落	32 – 256	2 – 8
LOW	文書	256 – ∞	0 – 1

3.2　プリズム層

BERT[3] は近年提案された自然言語処理分野における汎用事前学習済みモデルである。BERT は、ランダムにマスクされた入力系列のマスクされた部分を予測するタスクである masked language modeling (MLM) タスクと、ある文 A に対して別のある文 B が実際に続く文か予測するタスクである next sentence prediction (NSP) タスクによって事前学習が行われる。これによって、BERT は前後の文脈を考慮した単語レベルでの表現学習や文同士の関係性の学習を行うことができるため、様々な階層のタスクに適応できる事前学習済みモデルとして利用できる。しかしながら、この事前学習の方法では特定のニューロンが特定の階層のタスクに特化するのではなく、各階層のタスクに関する情報がすべてのニューロンに分散している可能性があるといえる。そこで、この研究では学習においてスペクトルフィルタを利用することで、異なるスケールのタスクに対して異なるニューロンを使用する方法を提案している。BERT のユニットを 5 等分し、それぞれに表 3.1 の 5 種類のフィルタを対応させ、適用する。この操作をプリズム層として最後の BERT 層のあとに加えることにする。このプリズム層を加えた事前学習済み BERT モデルを MLM タスクを用いて学習させる。

4　実験

4.1　言語表現に対するスペクトルフィルタの適用

スペクトルフィルタにより、異なるスケールの言語構造が取り出せるか評価を行うための実験を行う。スペクトルフィルタによりフィルタリングされた表現を用いて異なるスケールのタスクを行った時、スペクトルフィルタの選択が分類器の能力にどのように影響するか比較する。以下の各データセットに対して、$BERT_{BASE}$ モデル [3] を用いて 768 次元の単語表現を得る。次に、各次元に沿ってスペクトルフィルタを適用し、フィルタリングされた表現を用いて特定タスクを実行するソフトマックス分類器を学習させる。用いたデータセット、タスクは以下の通りである。

1. 文章タグ付け（単語レベル）。Penn Treebank データセットを使用する。タスクは、与えられたトーク

ン表現から品詞（例：動詞過去形、wh 代名詞、数詞）を予測することである。

2. 対話の分類（発話レベル）。Switchboard Dialog Speech Acts コーパスを使用する。このタスクは、与えられたトークン表現を含む発話の対話の分類（例：謝罪、言葉を濁す、感謝）を予測することである。
3. トピック分類（文書レベル）。20 Newsgroups データセット を使用する。与えられたトークン表現を含む文書のトピック（ニュースグループ；例：SCI.SPACE、COMP.GRAPHICS、REC.AUTOS）を予測する。
4. masked language modeling (MLM) タスク。単語の抜けがある文章の抜けを予測するタスク。BERT の事前学習で利用されるため、スペクトルフィルタ適用前の表現のターゲットタスクである。

各タスクの実行結果を図 4.1 に示す。図 4.1 より、単語レベルのタスクである文章タグ付けでは HIGH バンドを用いたときの精度が最も高い一方、フィルタリング前より性能が下がっている。これは、単語レベルの情報を主に使う一方で低周波数情報も必要としていることを示す。一方、文書レベルのタスクであるトピック分類タスクでは LOW バンドを用いたときの精度が最も高く、元の表現よりも高い性能を示した。これは、元の表現に存在する高い周波数変動がこのタスクにおいては、精度に悪影響を及ぼしている可能性があることを示している。また、発話を対象とした対話の分類タスクでは、MID フィルタを用いたときの精度がもっとも高い性能を示した。また、MLM タスクの結果は文章タグ付けの結果に最も類似しており、MLM タスクが局所的なタスクであることを確認できた。

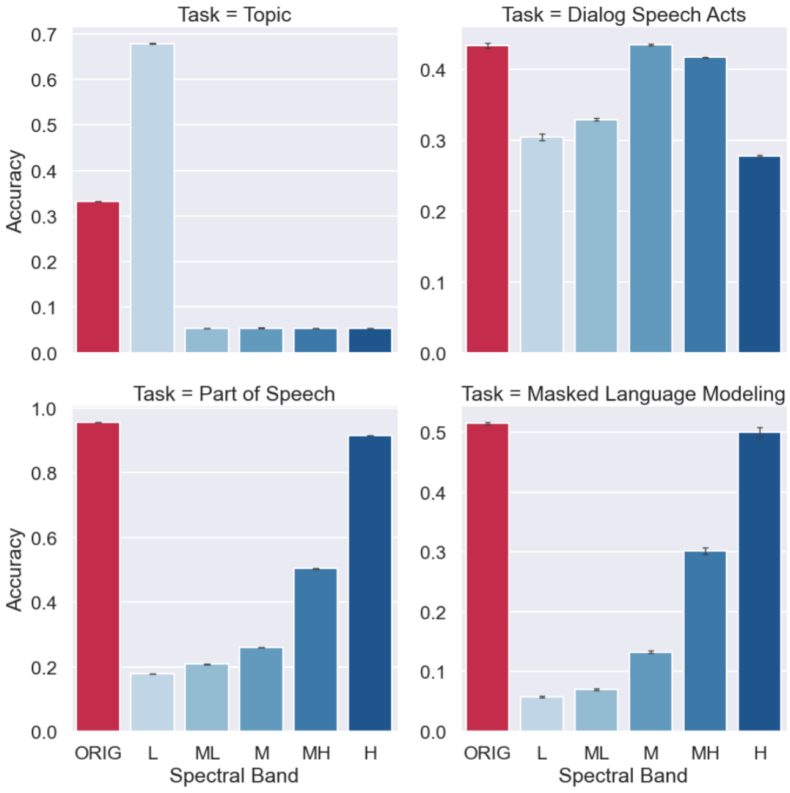

図 4.1　各タスクの実行結果

4.2 プリズム層ありのモデルを用いた実験

プリズム層を加えた事前学習済み BERT モデルを MLM タスクを用いて学習させる。学習は WikiText-103 データセットに対し、Adam のデフォルトパラメータを用いて、バッチサイズ 8、5 万ステップで学習を行う。比較のために、プリズム層なしの事前学習済みの BERT モデルについても同様に学習を行う。結果は表 4.1 に示すようにプリズム層を加えたモデルはそうでないモデルと比較して、文章タグ付けのタスクで高い精度を維持しながら、トピック分類や対話の分類において大きく改善した。

表 4.1　スペクトルフィルタ

タスク	モデル	精度 (%)	標準偏差 (%)
トピック分類	BERT	32.21	0.08
	BERT+ プリズム層	51.01	0.14
対話分類	BERT	47.09	0.33
	BERT+ プリズム層	54.02	0.61
文章タグ付け	BERT	95.86	0.02
	BERT+ プリズム層	94.41	0.02

また、入力の中央の連続した 100 個の単語がマスクされた MLM タスクにおける性能を比較したところ、予測精度は図 4.2 のようになった。これより、プリズム層がある場合はそうでない場合に比べて、欠落したトークンを推測できる確率が高いことが分かる。これより、BERT+ プリズムモデルが遠くの文脈を考慮してトークンを予測するより優れた長距離言語モデルであるといえる。

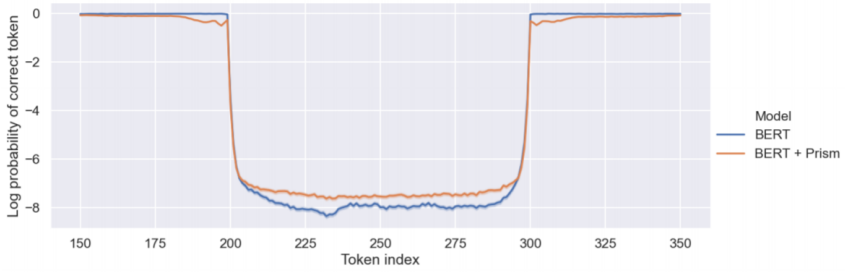

図 4.2　MLM タスクの実行結果

5 結論

スペクトル分析の技術を自然言語処理モデルに導入することで言語表現を言語の異なる階層ごとに適した表現に分離することができることが分かった。また、プリズム層を用いた学習を行うことでスケールの異なるタスクにおけるパフォーマンスを向上させたり、長距離の文脈を考慮した推定の精度を向上させたりすることができた。

6 論文に対する考察

この研究のユニークな点は、時系列の入力に対して深層学習モデルが生成した表現系列を信号処理的に分析する点である。表現系列を次元ごとに周波数領域に移すことで利用される情報を時間軸のスケールごとに分けるアプローチは大変興味深く思った。今回の分析では自然言語を対象としていたが、この手法は他の領域であっても適用できる範囲が広いと思う。例えば、動画は画像の時間的な連続データであるから、動画に対してもスペクトルフィルタを用いることができる。この場合、各画素の表現や畳み込みを行ったあとの情報をプリズム層に通せば各領域の時間的な周期性を分析することができる。あるいは、各画像に対しても空間的な周期を捉えるものとしてスペクトルフィルタの技術は利用できると考えられる。こうした応用により、今回の研究で BERT を特定の階層のタスクに適応させたように、既存の事前学習済モデルを特定の階層のタスクでパフォーマンスが上がるようにチューニングすることができると考えられる。

また、スペクトルフィルタの応用には、モデルのパフォーマンスを改善できる可能性があるだけではなく、あるタスクにはどの階層に着目したモデルが適切なのか考える指針を与える可能性がある。今回の研究では、3 つのタスクを扱っていたが、これにより、トピック分類では高周波情報が学習の妨げとなっていることや文章タグ付けでは高周波情報だけではなく周期の大きい文脈情報も必要となることが分かった。このようにあるタスクに対する周波数ごとに分けたモデルの精度を比較することがモデルを学習させる際に必要なデータの前処理やモデルの構築に役立つ可能性がある。

また、今回の研究では最後の層のあとにプリズム層を加え、スケールの異なる情報に特化したニューロンを作っていたが、途中の層にフィルタリングを利用するという応用も考えられる。例えば、途中の複数の層の後にフィルタを挟み、層を経るごとに高い周波数からカットしていけば、前の層ほど局所的な構造を見て、後の層ほど全体を見るというような CNN に近いモデルを作ることができる可能性がある。あるいは、途中にフィルタを加えたうえでカットする周波数も学習するパラメタに加えて学習を行えば、層やニューロンごとにタスクに適した周波数になるように学習させることができると考えられる。これを行うことでタスクに対する精度の向上が見込めるだけではなく、学習されたフィルタの周波数からモデルがどのようにデータを分析しているか可視化することができるという利点もある。

以上のように、スペクトルフィルタを導入することで既存のモデルの改善のみならず、タスクやモデルの分析にもつながるため、研究の余地があると考えられる。

参考文献

[1] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality, 2013.

[2] Felix Hill, Kyunghyun Cho, and Anna Korhonen. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1367–1377, San Diego, California, June 2016. Association for Computational Linguistics.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.