Attention Is All You Need

（2017）Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

どんなもの？

入力文章から別の文章で出力するモデル（Sequence Transduction Model）は、エンコーダー・デコーダを含む複雑なRNNやCNNが主流であった。本論文ではRNNやCNNを用いずAttentionのみを用いたモデル、Transformerを提案：
・再帰や畳み込みはいずれも使用しない
・並列化が可能で学習時間が大幅に短縮（既存最良結果よりも2 BLEU以上向上）

どうやって有効だと検証した？

WMT2014英独と WMT2014の2つのデータセットで検証：
・WMT 2014 英独翻訳タスクにおいて28.4 BLEUを達成し、アンサンブルを含む既存の最良の結果よりも2 BLEU以上向上
・WMT 2014英仏の翻訳タスクでは8つのGPUを用いて3.5日間の学習を行った結果、41.8という最新の単一モデルのBLEUスコアを獲得

技術の手法や肝は？

・単語の位置関係を捉えられる再帰や畳み込みを使っていないため、位置エンコードを導入。（一番最初に単語の分散表現を入力するときに単語位置に一意の値を各分散表現に加算する⇒Transformerは単語の位置に一意の値を与えてくれるsin関数とcos関数のパターンもしっかりと学習するため、位置の依存関係も学んでくれる）
・ Self-Attentionの使用：計算量が小さく、並列計算が可能。広範囲の依存関係を学習可能で、高い解釈可能性を有する。

議論はある？

無し

先行研究と比べて何がすごい？

RNNとエンコーダ-デコーダモデルは逐次的に単語を処理するがゆえに訓練時に並列処理ができない。また長文に対してはAttentionが使われていたが、そのAttentionはほぼRNNと一緒に使われていた。 Transformer はRNNを一切使わずにAttentionだけを使うことで、入力と出力の文章同士の広範囲な依存関係を捉えられる。
TransformerはCNNや逐次的なRNNを一切使わずAttentionのみを用いた一番最初のトランスダクションモデルである。

次に読むべき論文は？

・Junyoung Chung, Çaglar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR, abs/1412.3555, 2014.
・Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. arXiv preprint arXiv:1705.03122v2, 2017.

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

（2018） Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova

どんなもの？

BERTはTransformerのEncoderを使っているモデルで、文章を双方向から学習することによって文脈情報も学習。ラベルのついていない文章から表現を事前学習するように作られたもので、出力層を付け加えるだけで簡単にファインチューニングが可能。
事前学習としてMLM(=Masked Language Modeling)とNSP(Next Sentence Prediction)という2つの手法同時進行で学習させることで爆発的に精度向上⇒11個のNLPタスクで圧倒的SoTAを達成し、大幅にスコアを塗り替えた。

どうやって有効だと検証した？

モデルの性能はGLUE（The General Language Understanding Evaluation）という指標で評価。全てのデータセットにおいてOPEN AIなどの既存モデルよりもBERTの方が高いスコアを出している。「SQuAD 1.1」（スタンフォード大学の文章読解ベンチマーク）では、人工知能で初めて人間の平均の精度を超える結果を叩き出した。

技術の手法や肝は？

ラベルが付与されていない、つまり名前がついていない分散表現を事前学習としてMLM（入力の15%のトークンを[Mask]トークンでマスクし、元のトークンを当てるタスク）とNSP（「その2文が隣り合っているか」を当てるよう学習）という2つの手法同時進行で学習させ、双方向（「左から右」と「右から左」）で学習

議論はある？

マスクを表す[MASK]という単語は事前学習にしか存在せず、ファインチューニング時には出現しないので、ミスマッチが起こる。

先行研究と比べて何がすごい？

・従来の自然言語処理モデルでは、文章を単一方向からでしか処理できなかったのに対し、 BERTはMLM手法で双方向のTransformerによって学習させ、精度が向上し、文脈情報も学習する。
・ さらに、NSP手法によって単語だけでなく、文全体の表現、文の関係性も学習させ、より広範的な自然言語処理モデルとして機能できる。
・初めて文章レベル、トークンレベルの膨大なタスク群でのSoTAスコアの達成。

次に読むべき論文は？

・Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. In NAACL.
・Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI.

XLNet: Generalized Autoregressive Pretraining for Language Understanding

（2019） Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le

どんなもの？

BERTの強力な特徴である双方向Transformerという仕組みを残しつつ、 Masked Language Modelingを使わずに、今まで使われていたAutoreguressiveな言語モデルを使って事前学習をすることで、 BERTや RoBERTa（BERTと同じアーキテクチャで、データ数を増やしたりすることにより性能向上を図ったモデル）に比べて精度向上を実現。

どうやって有効だと検証した？

BooksCorpusとEnglish Wikipedia、Giga5、ClubWeb 2012-B、Common Crawlといったデータセットで検証。 BERTと同じ事前学習データセットを使って、BERTと比較した結果、いずれもBERTを大きく上回っている。 RoBERTaに対しても、精度を超えている。特に、RACEでは、かなり大きな差を開けて精度が改善された。

技術の手法や肝は？

BERTのMasked Language Modeling（マスクを使った言語モデルの学習）の代わりに、単語の順番をバラバラにして学習（元の文章の順番は変えずに、attentionを向ける先を限定して元の文章で自分より先に出てきた単語にもattentionを向けられるように）させることにより、双方向の言語モデルを実現。
Transformer-XLの仕組みをうまく導入して、長い文章に対してもうまく対応できるようにしている。

議論はある？

BERTと同じく、XLNetも部分的な予測を行うことになっている。
すなわち、シーケンス内のトークンのサブセットのみを予測する。

先行研究と比べて何がすごい？

BERTのMasked Language Modelingでは15%をマスクすることにより、マスクされた単語間の依存関係が無視されている。マスクを表す[MASK]という単語は事前学習にしか存在せず、ファインチューニング時には出現しない問題に対し、Permutation Language Modelingを利用して単語の順番を変えて学習させることにより、マスクを使わず双方向の言語モデルを学習。

次に読むべき論文は？

Zihang Dai, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860, 2019.

ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators

（2020） Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher D. Manning

どんなもの？

・ELECTRA:Efficiently Learning an Encoder that Classifies Token Replacements Accurateの略⇒置き換えた単語を適切に分類する方法を効率的に学習するエンコーダー
・ BERTの事前学習の問題点（文章中の単語のの15%しか学習しないなど）を克服
・事前学習にGAN(Generative Adversarial Networks; 敵対的生成ネットワーク)のアイデアを取り込み、事前学習の質と効率を高め、計算時間の短縮と、同じモデルサイズでも下流タスクでの性能が大幅に改善された

どうやって有効だと検証した？

GLUE（The General Language Understanding Evaluation）と「SQuAD 1.1」（スタンフォード大学の文章読解ベンチマーク）で検証。
GLUEスコア⇒小さなモデル：BERT-SmallとELECTRA-Smallでは同じパラメーター数でELECTRA-Smallの方が4.8%精度が改善；大きいモデル：ELECTRA-400kはRoBERTaやXLNetの1/4以下の計算量で同等の精度を出した。 SQuADスコア⇒同等以上の精度が確認された。

技術の手法や肝は？

“replaced token detection”という考え方を取り入れ、①Generatorが文章中の15%の単語を別の単語に置き換え、②Discriminatorが置き換えた単語かどうかを学習（ Generatorは小さめのモデルを使い、ある程度Discriminatorが本物か区別できるようにしている）⇒文章の15%だけでなく、文章中の全単語について学習することができる

議論はある？

無し

先行研究と比べて何がすごい？

・BERT の Masked Language Modelingが文章中のマスクした15%しか学習できないのに対し、トークンの真偽判定の形とすることで全ての入力トークンを学習に利用できるようになり、計算効率が向上。
・Discriminator へ “[MASK]” が入力されることがなくなり、BERT に存在した事前学習時とファインチューニング時のミスマッチを解消。
・GANのアイデアを取り込み、事前学習の質と効率を高める

次に読むべき論文は？

Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. Language GANs falling short. arXiv preprint arXiv:1811.02549, 2018.

ERNIE: Enhanced Representation through Knowledge Integration

（2019） Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, Hua Wu

どんなもの？

ERNIE (Enhanced Representation through kNowledge IntEgration). BERTと同じく双方向のTransformerを採用しているが、BERTのように単語にマスキングするに加えて、フレーズ単位とエンティティ(固有名詞)単位のマスキング（ knowledge masking strategies：知識マスキング戦略）を行うことにより、BERTが中国語（日本語も）を扱う上での問題点を解決したモデル。

どうやって有効だと検証した？

中国語の5つのタスク(Natural Language Inference、Semantic Similarity、Named Entity Recognition、Sentiment Analysis、Retrieval Question Answering)で精度を検証：いずれもBERTより精度が向上し、フレーズ単位、エンティティ単位のマスキングは相応の効果があるとされた。
Dialogue Language Model(DLM)や固有名詞部分を予測するタスクも改善が見られた

技術の手法や肝は？

・ERNIEは従来のBERTでは文字ごとにマスキングを行っていたのに対し、単語として連続する文字列はすべてマスクするように改良（Phrase-Level MaskingとEntity-Level Maskingを導入）⇒英語の時のような文法規則の学習効果が期待できる
・対話を記録したテキストデータを用いて事前学習をできるように、Dialogue Language Model(DLM)という手法を採用

議論はある？

無し

先行研究と比べて何がすごい？

・BERTをはじめとする最先端の自然言語処理では英語などのアルファベットを用いて表記する言語を主に取り扱っているため、中国語や日本語といったアジアの言語特有の問題（単語分割が明確ではない、文字そのものが意味を持つ表意文字等）に対処できていない⇒ ERNIEは単語として連続する文字列はすべてマスクするように改良
・BERTはWikipediaやニュースなどの単一話者によるデータしか学習対象にできなかったことに対し、 DLMを採用し、対話データを用いた学習も可能

次に読むべき論文は？

Yoshua Bengio, Rejean Ducharme, Pascal Vincent, and ´ Christian Jauvin. 2003. A neural probabilistic language model. Journal of machine learning research, 3(Feb):1137–1155.