AI 特許紹介(20)

AI 特許を学ぶ！究める！

～Transformer 特許～

2020 年 9 月 10 日

河野特許事務所

所長　弁理士　河野英仁

「AI 特許紹介」シリーズは、注目すべき AI 特許のポイントを紹介します。熾烈な競争となっている第４次産業革命下では AI 技術がキーとなり、この AI 技術・ソリューションを特許として適切に権利化しておくことが重要であることは言うまでもありません。

AI 技術は Google, Microsoft, Amazon を始めとした IT プラットフォーマ、研究機関及び大学から毎週のように新たな手法が提案されており、また AI 技術を活用した新たなソリューションも次々とリリースされています。

本稿では米国先進 IT 企業を中心に、これらの企業から出願された AI 特許に記載された AI テクノロジー・ソリューションのポイントをわかりやすく解説致します。

1.概要

特許権者　Google

出願日　2018 年 6 月 28 日

登録日　2019 年 10 月 22 日

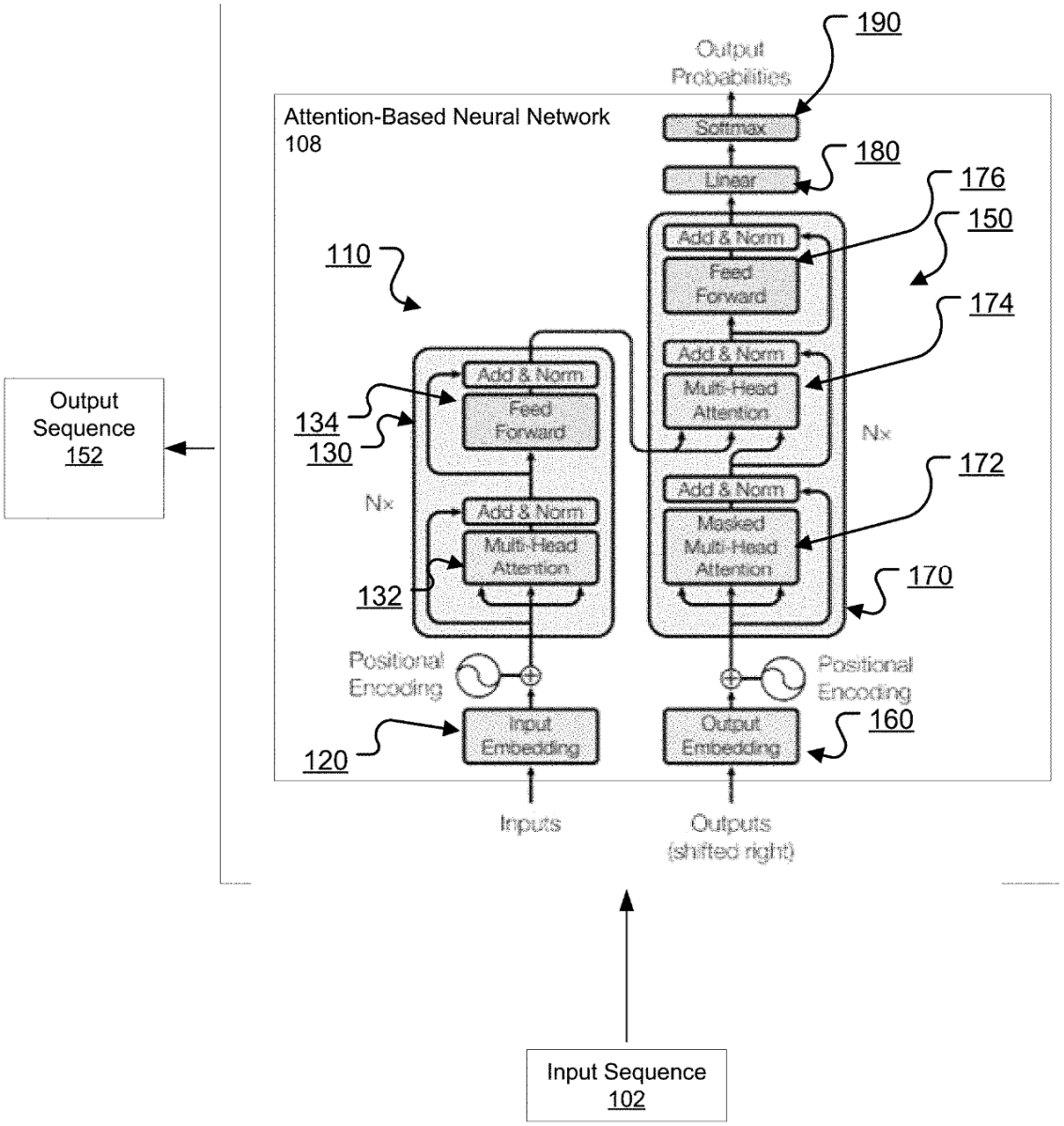
登録番号 US10452978

発明の名称　アテンションに基づくシーケンス変換ニューラルネットワーク

978 特許は、アテンションメカニズムを導入することにより、RNN 及び CNN を用いることなく、時系列データの予測処理を行うトランスフォーマー技術に関する。

2.特許内容の説明

図 1 は、ニューラルネットワークシステム 100 を示す。



ニューラルネットワークシステム 100 は、入力シーケンス 102 を受信し、入力シーケンス 102 を処理して、入力シーケンス 102 を出力シーケンス 152 に変換する。

入力シーケンス 102 は、入力順に複数の入力位置の各々においてそれぞれのネットワーク入力を有し、出力シーケンス 152 は、出力順に複数の出力位置の各々においてそれぞれのネットワーク出力を有する。ニューラルネットワークシステム 100 は、アテンションベースのシーケンス変換ニューラルネットワーク 108 を含み、このニューラルネットワーク 108 は、エンコーダニューラルネットワーク 110 およびデコーダニューラルネットワーク 150 を含む。

エンコーダニューラルネットワーク 110 は、入力シーケンス 102 を受信して、入力

シーケンス内の各々のネットワーク入力のそれぞれエンコードされた表現を生成するように構成される。

デコーダニューラルネットワーク 150 は、出力シーケンス 152 を生成するためにネットワーク入力のエンコードされた表現を使用するように構成される。エンコーダ 110 およびデコーダ 150 はいずれも、アテンションベースである。エンコーダまたはデコーダのいずれも、畳み込みレイヤまたは再帰型レイヤを含まない。

エンコーダニューラルネットワーク 110 は、埋め込みレイヤ 120、および複数(N 個)のエンコーダサブネットワーク 130 のシーケンスを含む。埋め込みレイヤ 120 は、入力シーケンス内の各ネットワーク入力について、ネットワーク入力を、埋め込みスペース内のネットワーク入力の数値表現に、たとえば埋め込みスペース内のベクトルに、マップするように構成される。次いで、埋め込みレイヤ 120 は、ネットワーク入力の数値表現を、エンコーダサブネットワーク 130 のシーケンス内の第 1 のサブネットワークに、つまり N 個のエンコーダサブネットワーク 130 の第 1 のエンコーダサブネットワーク 130 に提供する。

シーケンス内の最後のエンコーダサブネットワークにより生成されたエンコーダサブネットワーク出力は、ネットワーク入力のエンコードされた表現として使用される。各エンコーダサブネットワーク 130 は、エンコーダセルフアテンションサブレイヤ 132 を含む。エンコーダセルフアテンションサブレイヤ 132 は、複数の入力位置の各々についてサブネットワーク入力を受信し、入力順に各特定の入力位置ごとに、特定の入力位置においてエンコーダサブネットワーク入力から導き出されたクエリを使用して入力位置においてエンコーダサブネットワーク入力にわたりアテンションメカニズムを適用して、特定の入力位置のそれぞれの出力を生成するように構成される。

このアテンションメカニズムについては後述する。

エンコーダサブネットワーク 130 の各々は、エンコーダセルフアテンションサブレイヤの出力をエンコーダセルフアテンションサブレイヤへの入力と結合して、エンコーダセルフアテンション残余出力を生成する残余接続レイヤと、レイヤ正規化をエンコーダセルフアテンション残余出力に適用するレイヤ正規化レイヤとを含む(図 1 の Add & Norm)。

エンコーダサブネットワークの一部または全部はまた、それぞれ入力シーケンス内の各位置で動作するように構成される位置ごとのフィードフォワードレイヤ 134 を含む。

各入力シーケンス位置について、フィードフォワードレイヤ 134 は、入力位置において入力を受信し、入力位置において入力に変換シーケンス（例えば ReLU 活性化関数）を適用して入力位置の出力を生成する。

次いでデコーダニューラルネットワーク 150 について説明する。デコーダニューラルネットワーク 150 は、複数の生成時間ステップの各々において、(i)エンコードされた表現、および(ii)出力順に出力位置に先行する出力位置におけるネットワーク出力に条件付けられた対応する出力位置のネットワーク出力を生成することにより、出力シーケンスを生成する。

デコーダニューラルネットワーク 150 は、埋め込みレイヤ 160、デコーダサブネットワーク 170 のシーケンス、線形レイヤ 180、およびソフトマックスレイヤ 190 を含む。

各デコーダサブネットワーク 170 は、デコーダセルフアテンションサブレイヤ 172 およびエンコーダ-デコーダアテンションサブレイヤ 174 という 2 つの異なるアテンションサブレイヤを含む。

各デコーダセルフアテンションサブレイヤ 172 は、各生成時間ステップにおいて、対応する出力位置に先行する各出力位置について入力を受信し、特定の出力位置の各々について、特定の出力位置において入力から導き出された 1 つまたは複数のクエリを使用して対応する位置に先行する出力位置において入力にわたりアテンションメカニズムを適用して、特定の出力位置の更新された表現を生成する。すなわち、デコーダセルフアテンションサブレイヤ 172 は、出力シーケンス内の現在の出力位置に先行する位置にはない任意のデータをアテンションまたは処理しないようにマスクされるアテンションメカニズムを適用する。

一方、各エンコーダ-デコーダアテンションサブレイヤ 174 は、各生成時間ステップにおいて、対応する出力位置に先行する各出力位置について入力を受信し、出力位置の各々について、出力位置の入力から導き出された 1 つまたは複数のクエリを使用して入力位置においてエンコードされた表現にわたりアテンションメカニズムを適用して出力位置の更新された表現を生成する。

このように、エンコーダ-デコーダアテンションサブレイヤ 174 は、エンコードされた表現にわたりアテンションを適用するが、エンコーダセルフアテンションサブレイヤ 172 は、出力位置において入力にわたりアテンションを適用する。

次にアテンションメカニズムについて説明する。下記図 2 の左がアテンションサブレイヤであり、右がマルチヘッドアテンションである。

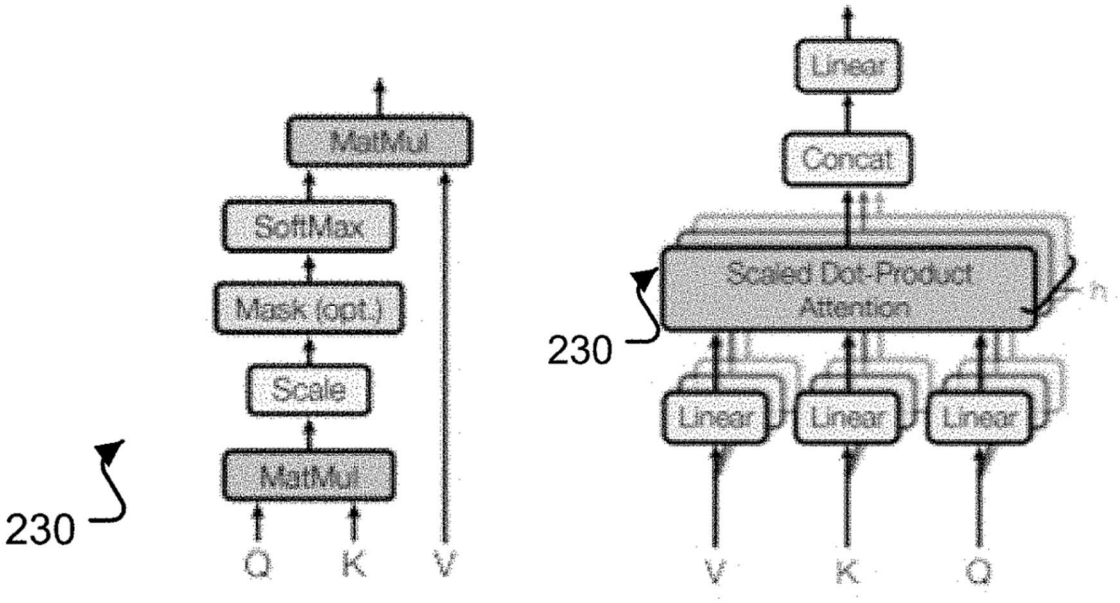


図 2

一般に、アテンションメカニズムは、クエリ、キー及びバリューのペアのセットを出力にマップする。このクエリ、キー、およびバリューはすべてベクトルである。出力は、バリューの加重合計として計算され、ここで各バリューに割り当てられる重みは、対応するキーとのクエリの適合関数によって計算される。処理対象が文章である場合、対象となる単語がクエリであり（Query 単語）、これに関連する他の単語がキーとなる（Key 単語）。

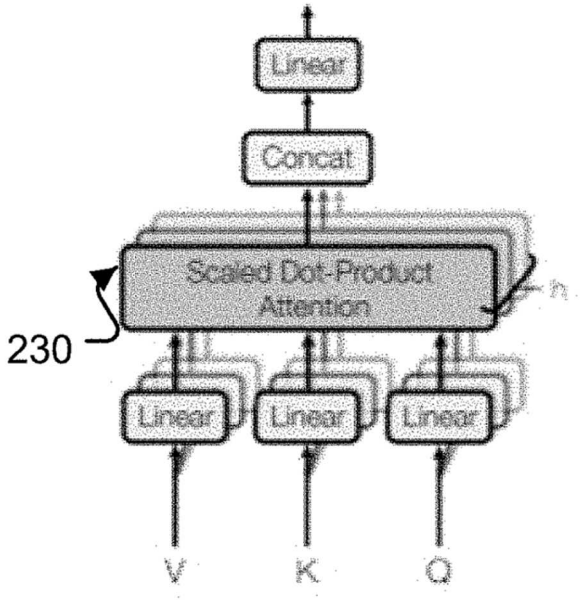
各アテンションサブレイヤは、縮小付き内積アテンションメカニズム 230 を適用する。縮小付き内積アテンションにおいて、所与のクエリについて、アテンションサブレイヤは、キーのすべてとのクエリの内積を計算し、内積の各々を、倍率により、たとえばクエリとキーの次元の平方根により、除算して、縮小付き内積にわたりソフトマックス関数を適用してバリューへの重みを取得する。次いで、アテンションサブレイヤは、これらの重みに従ってバリューの加重合計を計算する。したがって、縮小付き内積アテンションの場合、適合関数は内積であり、適合関数の出力は、倍率によりさらに縮小される。

動作中、アテンションサブレイヤは、クエリのセットにわたり同時にアテンションを計算する。特に、アテンションサブレイヤは、クエリを行列 Q にパックし、キーを行列 K にパックし、バリューを行列 V にパックする。ベクトルのセットを行列にパックするために、アテンションサブレイヤは、行列の行としてベクトルを含む行列を生成する

ことができる。

次いで、アテンションサブレイヤは、行列Qと行列Kの転置の間に行列乗算(MatMul)を実行して、適合関数出力の行列を生成する。そしてアテンションサブレイヤは、適合関数出力行列を縮小する。つまり倍率により行列の各要素を除算する。次いで、アテンションサブレイヤは、縮小付き出力行列にソフトマックスを適用して、重みの行列を生成し、重み行列と行列Vの間に行列乗算(MatMul)を実行して、バリューごとのアテンションメカニズムの出力を含む出力行列を生成する。

次にマルチヘッドアテンションについて説明する。アテンションサブレイヤが、さまざまな位置においてさまざまな表現サブスペースから情報に共同でアテンションする必要がある場合、マルチヘッドアテンションを採用する。



マルチアヘッドアテンションは、h個の異なるアテンションメカニズムを並行して適用し、同じアテンションサブレイヤ内の各々のアテンションレイヤが同じ元のクエリQ、元のキーK、および元のバリューVを受け取る。

各アテンションレイヤは、元のクエリ、キー、およびバリューを、学習された線形変換を使用して変換して、アテンションメカニズム230を変換されたクエリ、キー、およびバリューに適用する。各アテンションレイヤは、一般に、同じアテンションサブレイヤ内の相互のアテンションレイヤからさまざまな変換を学習する。

特に、各アテンションレイヤは、学習されたクエリ線形変換を各元のクエリに適用して、各元のクエリのレイヤ固有のクエリを生成し、学習されたキー線形変換を各元のキ

ーに適用して、各元のクエリのレイヤ固有のキーを生成し、学習されたバリュー線形変換を各元のバリューに適用して、各元のバリューのレイヤ固有のバリューを生成する。

次いで、アテンションレイヤは、これらのレイヤ固有のクエリ、キー、およびバリューを使用し、アテンションメカニズムを適用して、アテンションレイヤの初期出力を生成する。そして、アテンションサブレイヤは、アテンションレイヤの初期出力を結合して、アテンションサブレイヤの最終出力を生成する。図 2 に示されるように、アテンションサブレイヤは、アテンションレイヤの出力を連結(concat)し、学習された線形変換を連結された出力に適用して、アテンションサブレイヤの出力を生成する。

3.クレーム

978 特許のクレーム 1 及び 2 は以下の通りである。

1. 1 つまたは複数のコンピュータと、1 つまたは複数のコンピュータによって実行されたときに 1 つまたは複数のコンピュータに、シーケンス変換ニューラルネットワークを実装させる命令を格納する 1 つまたは複数の記憶装置と、を含むシステムであり、

前記シーケンス変換ニューラルネットワークは、入力順序の複数の入力位置のそれぞれにそれぞれのネットワーク入力を有する入力シーケンスを、出力順序の複数の出力位置のそれぞれにそれぞれのネットワーク出力を有する出力シーケンスに変換し、以下を含む：

入力シーケンスを受け取り、入力シーケンス内の各ネットワーク入力のそれぞれのエンコードされた表現を生成するように構成されたエンコーダニューラルネットワークを備え、

各エンコーダサブネットワークは、複数の入力位置のそれぞれに対するそれぞれのエンコーダサブネットワーク入力を受け取り、複数の入力位置のそれぞれに対してそれぞれのサブネットワーク出力を生成するように構成されており、各エンコーダサブネットワークは以下を含む：

複数の入力位置のそれぞれについて、および入力順序の特定の入力位置ごとにサブネットワーク入力を受け取るように構成されたエンコーダセルフアテンションサブレイヤーを備え、

特定の入力位置のそれぞれの出力を生成すべく、複数の入力位置でエンコーダサブネットワーク入力にセルフアテンションメカニズムを適用し、該セルフアテンションメカニズムの適用は以下を含み、

特定の入力位置でサブネットワーク入力からクエリを決定し、複数の入力位置でサブネットワーク入力から派生したキーを決定し、複数の入力位置でサブネットワーク入力から派生したバリューを決定し、決定したクエリ、キー、及びバリューを用いて、

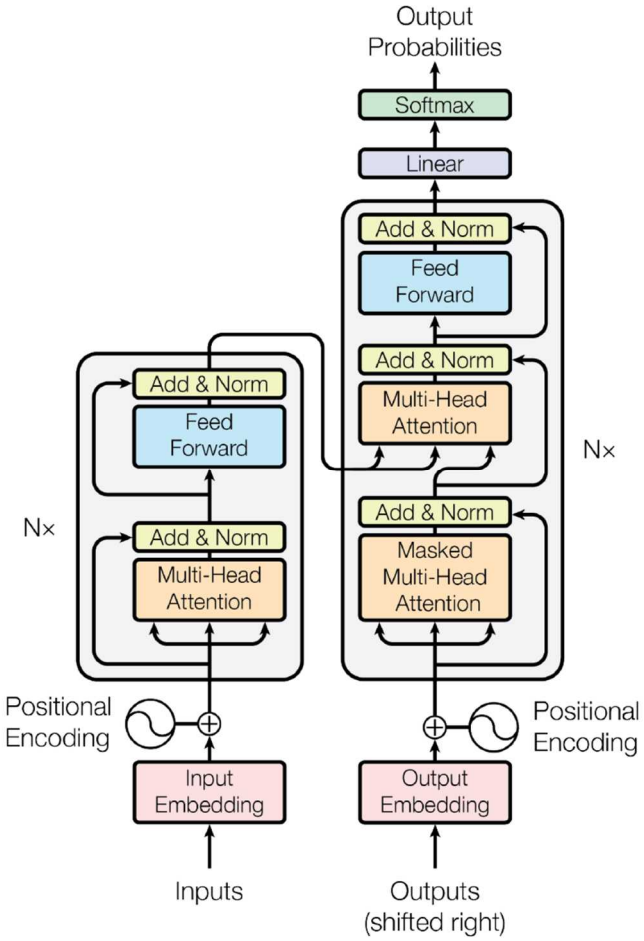
特定の入力位置のそれぞれの出力を生成し、

エンコードされた表現を受け取り、出力シーケンスを生成するデコーダニューラルネットワークとを備える。

4. Transformer に関する論文

アテンションメカニズムである Transformer に関する論文[1]が Google の Ashish Vaswani 氏らにより発表されている。

論文に開示されているネットワーク構成は、978 特許の内容と同じである。



下記テーブルは実験結果を示している。

[1] Ashish Vaswani “Attention Is All You Need” arXiv:1706.03762v5 2017 年 12 月 6 日

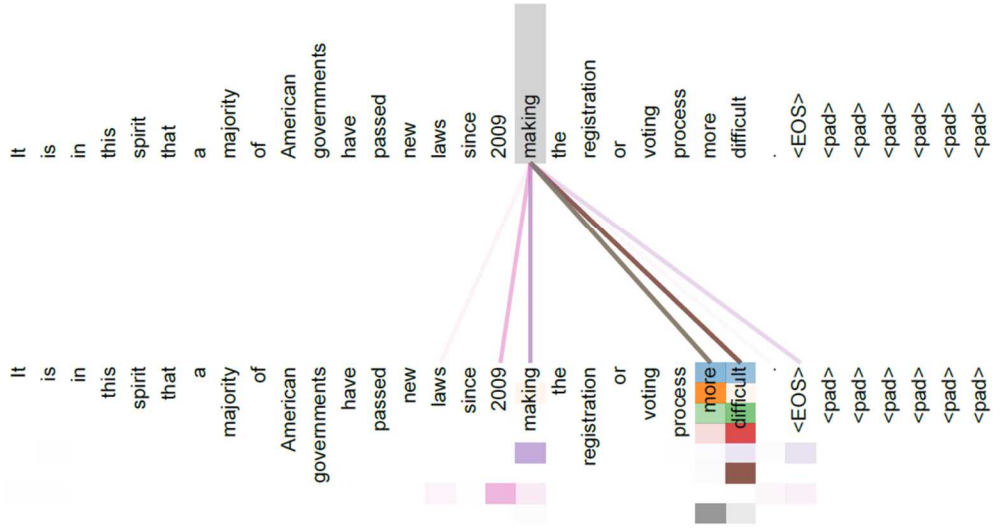
Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

Transformer は、英語からドイツ語、および、英語からフランス語への newstest2014 テストにおいて、過去の最先端モデルよりもわずかなトレーニングコストで、より優れた BLEU スコアを達成している。

WMT 2014 の英語からドイツ語への翻訳タスクでは、Big Transformer モデル（表 2 の最下段）の BLUE スコアは 28.4 であり、以前に報告された最良のモデル（アンサンブルを含む）よりも 2.0 優れている。

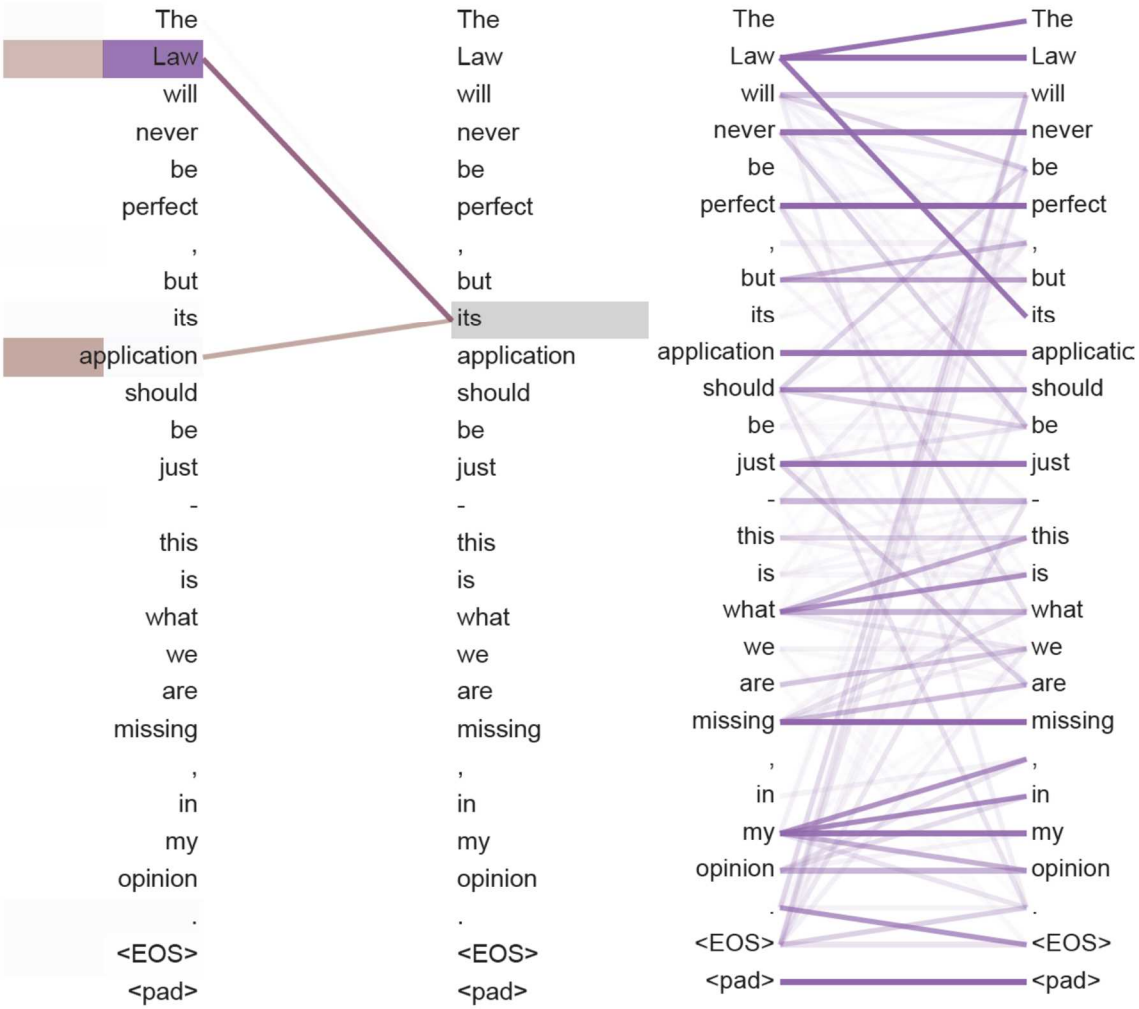
WMT 2014 の英語からフランス語への翻訳タスクでは、Big Transformer モデルの BLEU スコアは 41.8 であり、以前に公開されたすべての単一モデルよりも優れており、またトレーニングコストは 1/4 未満である。

Attention Visualizations



上記図は、6 の内、レイヤ 5 のエンコーダのセルフアテンションにおける長距離依存性に続くアテンションメカニズムの例を示している。図の例では making というクエリに対し、下側にキー単語が示されている。キー単語に付されている各色はそれぞれのヘッドの Attention Weight である。

多くのアテンションヘッドは、動詞「making」の遠方への依存に注意を払い、「making...more difficult」というフレーズを完成させる。



図の右側はヘッド 5 のフルアテンションであり、左側はアテンションヘッド 5 と 6 の単語「its」だけからアテンションを分離したものである。レイヤ 5 にある 2 つのアテンションヘッドは、首句反復(Law と its)の解決に明らかに関与している。

以上

著者紹介

河野英仁

河野特許事務所、所長弁理士。立命館大学情報システム学博士前期課程修了、米国フランクリンピアースローセンター知的財産権法修士修了、中国清華大学法学院知的財産夏季セミナー修了、MIT(マサチューセッツ工科大学)コンピュータ科学・AI 研究所 AI コース修了。

AI 特許コンサルティング、医療 AI 特許コンサルティングの他、米国・中国特許の権利化・侵害訴訟を専門としている。著書に「世界のソフトウエア特許(共著)」、「FinTech 特許入門」、「AI/IoT 特許入門 2.0」がある。