AI 特許紹介(54)
AI 特許を学ぶ！究める！
～LayoutLM 特許～

2023 年 7 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
特許出願人　Microsoft Technology Licensing
出願日　2020 年 6 月 12 日
公開日　2021 年 12 月 16 日
公開番号 WO2021248492
発明の名称　文書内のテキストの意味表現

492 特許は、文書中のテキスト要素と、テキスト要素の空間的配置を示すレイアウト情報とに基づき、複数のテキスト要素のそれぞれの意味論的特徴表現を生成する LayoutLM に関する。

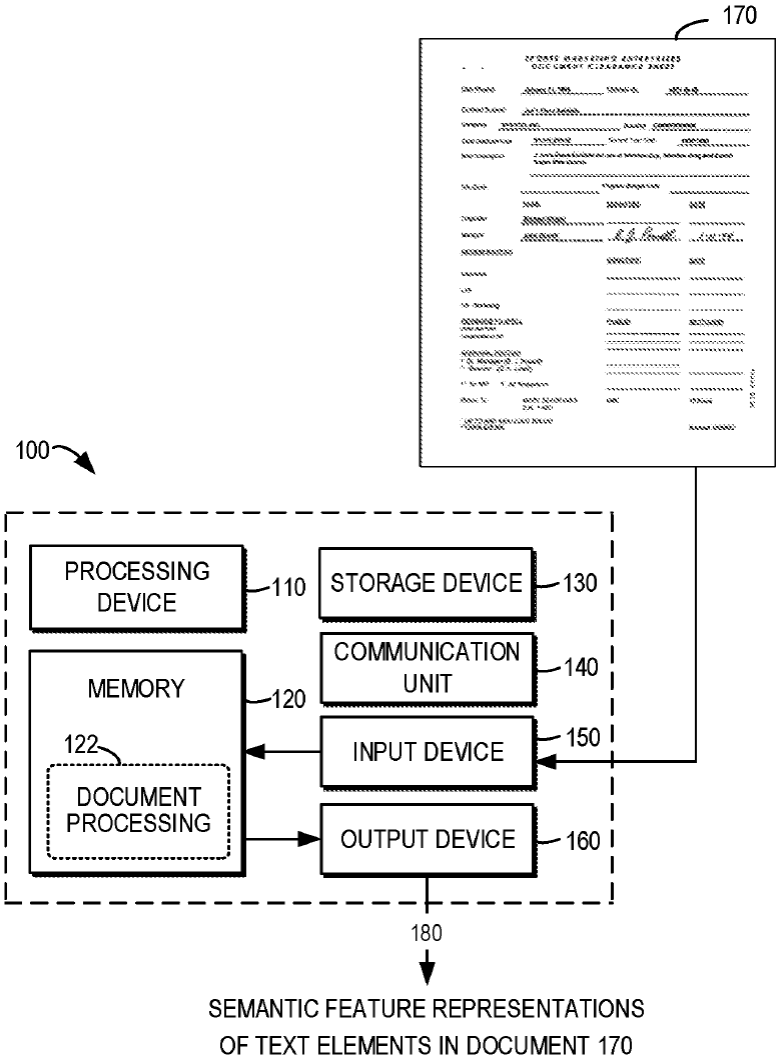
2.特許内容の説明
ドキュメント人工知能 (AI)は、ドキュメントを自動的に読み取り、理解し、分析する技術を使用するトレンドのアプリケーション分野である。既存のドキュメント AI モデルとアルゴリズムのほとんどは、光学式文字認識 (OCR)等の技術を使用してドキュメント画像からテキスト情報を認識し、次に、さまざまな NLP モデルを活用してテキス

ト情報のセマンティクスを調査することにより、純粋に自然言語処理（NLP）の観点からテキスト情報を処理する。

しかしながら、多くの NLP モデルはテキストレベルの操作のみに焦点を当てており、単一のテキストモーダルに基づいてトレーニングされているため、結果は文書画像の分類やフォームの理解など、文書画像に固有のタスクを実行するのには適していない。

本発明では、テキスト要素のセットとレイアウト情報は、テキスト要素のそれぞれの意味論的特徴表現を生成するために共同して使用される。テキスト情報とレイアウト情報の両方を組み合わせて使用することにより、テキスト要素の豊富なセマンティクスを特徴表現に効果的に取り込むことができる。

下記図はハードウェア構成を示すブロック図である。



コンピューティングデバイス 100 は、入力デバイス 150 を介して、テキストが存在する文書 170 を受信する。図の例では、文書 170 はスキャンされた画像である。入力文書 170 は文書処理モジュール 122 に提供され、文書処理モジュール 122 は文書 170 から認識されたテキストを処理する。特に、文書処理モジュール 122 は、文書 170 内に提示されるテキスト要素のセットに対応する意味論的特徴表現 180 を生成する。

意味論的特徴表現（意味論的表現、意味論的特徴、またはテキスト埋め込みとも呼ばれる）は、自然言語のテキストシーケンス内のテキスト要素の直感的な意味または意味論を特徴付けたりエンコードしたりするために使用される。意味論的特徴表現は数値ベクトルの形式にすることができる。

意味論的特徴表現は、NLP の一連の言語モデリングおよび特徴学習技術を使用して決定でき、この技術では、語彙のテキスト要素が、その語彙内の他のテキスト要素に対する意味、使用法、およびコンテキストに基づいて実数ベクトルにマッピングされる。さらに、同様の意味を持つテキスト要素は同様のベクトルを持ち、ベクトル空間内で互いに近接している。

意味論的特徴表現を使用すると、テキスト要素または文書に関連する 1 つ以上の下流処理タスクを容易にすることができ、たとえば文書の理解を容易にすることができる。

レイアウト情報は、特定の文書内のテキスト要素間の空間的関係またはレイアウトをキャプチャすることができ、さらにテキスト要素の意味論に寄与する。たとえば、多くの形式の情報は、「DATE: 11/28/84」のようなキーと値のペアとして表示され、ここで、「DATE」という単語がキーで、記号文字列「11/28/84」が値である。一般に、キーと値のペアは、特定の形式で左から右、または、上から下の配置で配置される。

レイアウト情報を利用してセマンティックな特徴表現を抽出すると、テキスト要素の空間的配置とそのセマンティクスをより適切に調整するのに役立ち、財務レポート分析、領収書の理解、証明書/ライセンスの認識、注文書の認識など、ドキュメントの分析と理解における多くの実世界のアプリケーションに利益をもたらす。

文書処理モジュール 122 は、文書 170 全体にわたってテキスト情報およびレイアウト情報を共同処理して、テキスト要素の意味論的特徴表現を生成する。文書処理モジュール 122 は、深層学習モデルを利用して、テキスト情報およびレイアウト情報を意味論的特徴表現に符号化する。

下記図は、文書処理モジュール 122 の構造を示す。

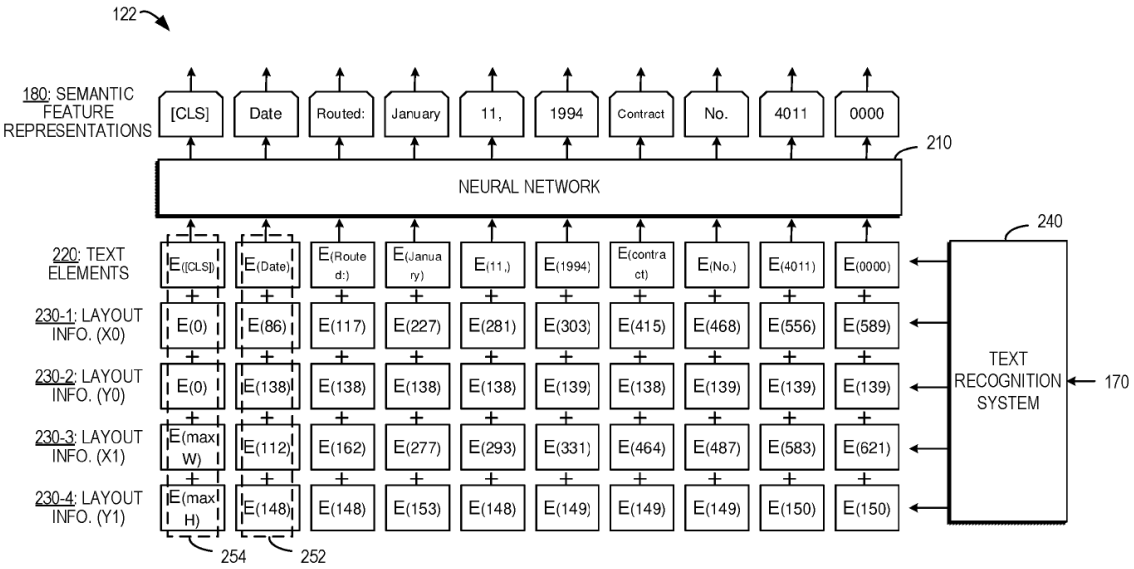


FIG. 2

図示のように、テキスト認識システム 240 は、文書 170 から一連のテキスト要素 220 およびレイアウト情報を含むテキスト情報を抽出する。図の例では、レイアウト情報は、４つのレイアウト情報要素 230-1, 230-2, 230-3, 及び 230-4 を含む。

テキスト要素 220 およびレイアウト情報 230 は、それぞれのテキスト要素 220 の意味論的特徴表現 180 を生成するためにニューラルネットワーク 210 に提供される。文書 170 が画像を含む場合、テキスト認識システム 240 は、画像内に表示されるテキストを認識する。

下記図はテキスト要素の抽出例を示す説明図である。

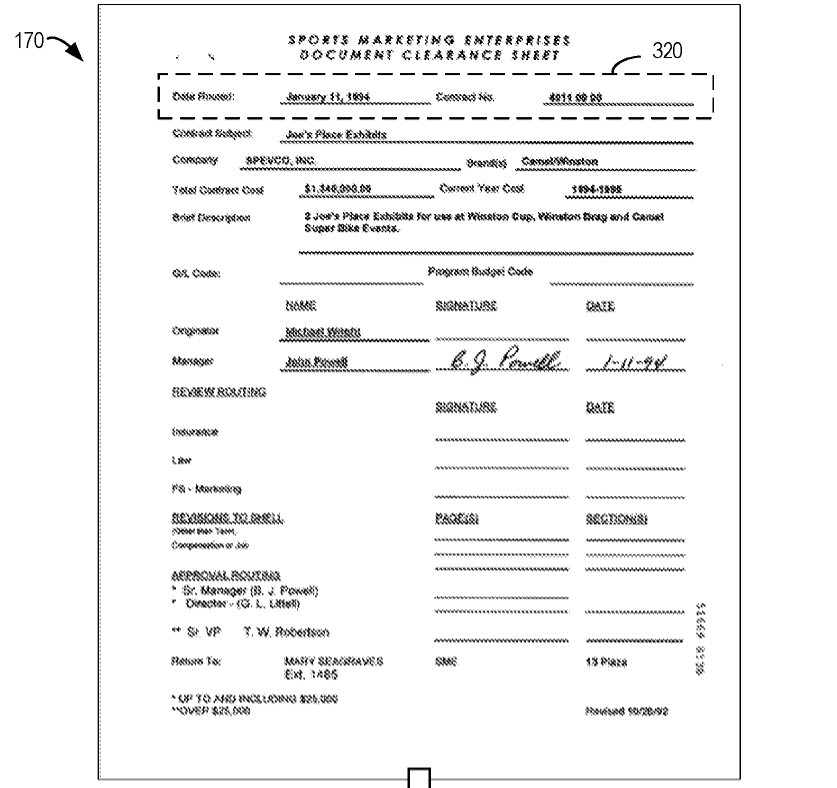


IMAGE BLOCK	TEXT ELEMENT	BOUNDING BOX (X0, Y0, X1, Y1)
Date	Date	86 138 112 148
Routed:	Routed:	117 138 162 148
January	January	227 138 277 153
11,	11	281 138 293 148
1994	1994	303 139 331 149
Contract	Contract	415 138 464 149
No.	No.	468 139 583 150
4011	4011	556 139 583 150
00 00	0000	589 139 621 150

FIG. 3

テキスト情報は文書 170 から行ごとに抽出される。図に示されるように、テキスト行領域 320 が認識され、

“Date, ” “Routed: , ” “January, ” “11, ” “1994” “Contract, ” “No., ” “4011, ” “0000” を含む、この領域内のテキスト要素のセット 220 が決定される。

これらのテキスト要素は、さらなる処理のためにニューラルネットワーク 210 への入力として提供される。文書 170 の他の領域のテキスト要素 220 をさらに抽出して処理

する。

テキスト要素 220 の空間的配置を示すために、レイアウト情報 230 は、文書内のテキスト要素のそれぞれの位置を示す。多くの既存の言語モデルにおける入力単語シーケンス内の単語位置をモデル化する位置埋め込みとは異なり、レイアウト情報 230 は、文書 170 内の各テキスト要素（例えば、単語）の空間位置をモデル化することを目的とする。

文書 170 内のテキスト要素 220 のそれぞれの位置を表すために、文書 170 全体を、例えば左上の点を原点とする二次元（２Ｄ）座標系として考える。このような設定では、テキスト要素の位置を 2D 座標系内で 2D 位置として定義できる。文書 170 内のテキスト要素 220 を境界付けるそれぞれのバウンディングボックスを決定し、バウンディングボックスの位置を使用してテキスト要素 220 の位置を定義する。

バウンディングボックスは一般に、特に画像からのテキスト認識のプロセスにおいて、テキスト要素を認識するための関心領域（ROI）の位置を特定するために使用される。一例では、テキスト要素 220 を境界付けるバウンディングボックスの位置は、(x0, y0, x1, y1)によって定義される。ここで、(x0, y0) は、境界ボックスの左上の２Ｄ位置に対応し、(x1, y1) は、境界ボックス内の右下の 2D 位置を表す。

(x0, y0, x1, y1)によって定義される位置は、テキスト要素 220 の位置として直接考慮される。このような位置は、位置を定義するだけでなく、テキスト要素 220 のサイズも定義できる。テーブル 330 は、認識されたテキスト要素に対応する境界ボックスの位置のリストも示す。レイアウト情報 230 は、「x0」に対応するレイアウト情報要素 230-1、「y0」に対応するレイアウト情報要素 230-2、「x1」に対応するレイアウト情報要素 230-3、及び「y1」に対応するレイアウト情報要素 230-4 を含む。

それぞれのテキスト要素 220 の個々の位置に加えて、レイアウト情報 230 は、文書 170 内の全体の配置範囲の情報をさらに含む。配置範囲は、テキスト情報（各テキスト要素 220 を含む）の可能な位置の範囲を示す。位置決め範囲は、文書 170 の全体的なレイアウト情報を示すために使用される。位置決め範囲は、(0, 0, maxW, maxH)によって定義される。ここで、(0, 0) は、座標原点（例えば、文書 170 の左上の点）を示し、(maxW, maxH) は、原稿 170 の幅と高さの最大の座標値（例えば、原稿 170 の右下の点）を示す。

決定されたテキスト要素 220 およびレイアウト情報 230 は、ニューラルネットワー

ク 210 への入力として提供される。ニューラルネットワーク 210 における処理を可能にするために、テキスト要素 220 およびレイアウト情報 230 は、ニューラルネットワーク 210 への入力として使用される対応する埋め込みとして表現される。埋め込みとは、所定のサイズを有するテキスト要素またはレイアウト情報要素の数値表現である。

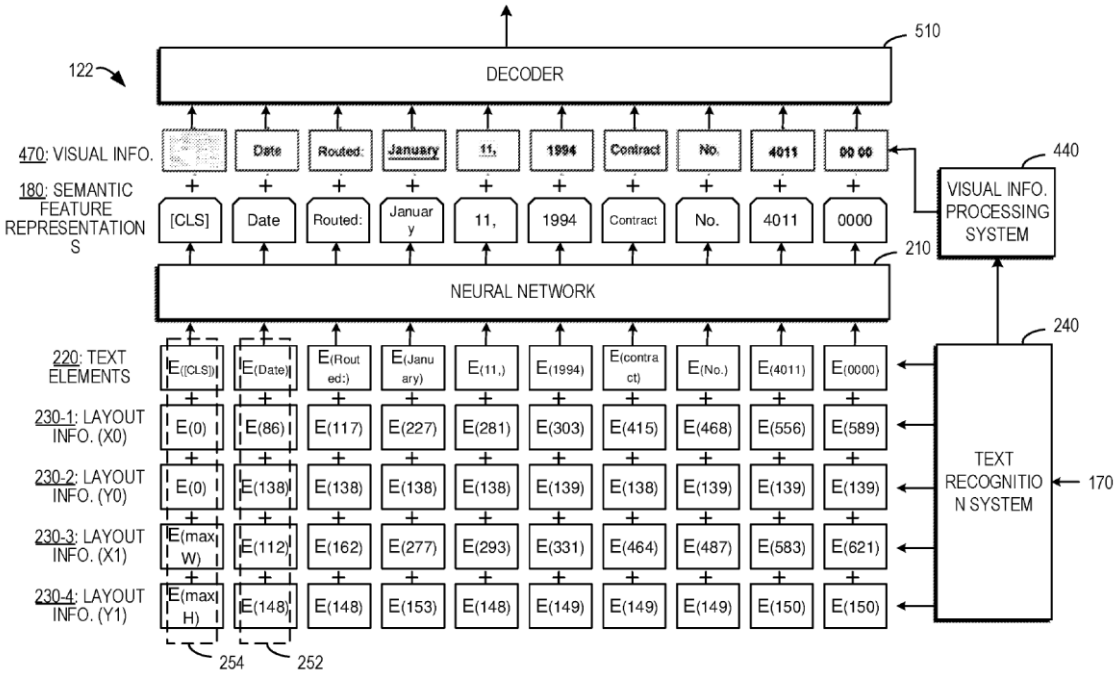
テキスト要素の埋め込みは、図に示す「E (x) 」として表され、「x」はテキスト要素を示す。テキスト要素 220 のテキスト要素埋め込みへのマッピングは、テキスト要素とテキスト要素埋め込みとの間の所定のマッピングテーブルに基づいて実行される。

文書 170 内に提示されるテキスト要素 220 に加えて、「CLS」として表される特別なマーカーが、テキスト要素 220 のシーケンスに、例えばシーケンスの先頭にさらに含まれる。マーカー「CLS」の埋め込みは、予め定められており、E ([CLS]) と表される。レイアウト情報 230 は、レイアウト情報要素に分割することができ、各レイアウト情報要素は、レイアウト情報埋め込みとも呼ばれる埋め込みにマッピングされる。

テキスト要素の位置が (x0, y0, x1, y1)によって定義される例では、4 つの埋め込み層が組み込まれ、レイアウト情報要素 “x0” 230-1, “y0” 230-2, “x1” 230-3, 及び “y1” 230-4 は、２D位置埋め込みと呼ばれる。レイアウト情報の埋め込みは、図に示すように、“E (x0) , ” “E (y0) , ” “E (x1) , ”及び “E (y1) ”で表され、“x0, ” “y0, ” “x1, ” 及び “y1”は、テキスト要素 220 の位置における特定の座標値を示す。(0, 0, maxW, maxH) などの位置決め範囲も“E (0) , ” “E (0) , ” “E (maxW) , ” 及び “E (maxH) で示されるように、2D 位置埋め込みにマッピングすることができる。

埋め込み変換後、テキスト要素ごとに、対応するテキスト要素の埋め込みとレイアウト情報の埋め込みを結合（例えば、合計）して、ニューラルネットワーク 210 に入力する。例えば、文書 170 から認識されたテキスト要素「日付 Date」については、結果として生じる埋め込みの組み合わせ 252 が得られる。他のテキスト要素の埋め込みの組み合わせも同様に取得できる。ニューラルネットワーク 210 は、テキスト要素 220 の意味論を抽出するために、テキスト要素 220 およびレイアウト情報 230、より具体的にはそれらの埋め込みを処理する。

その他、文書 170 内のテキスト要素 220 の個々の視覚的外観および文書 170 全体の全体的な視覚的外観も重要なヒントであり、文書 170 内のテキスト要素 220 の意味論に寄与する。下記図は、そのような実装形態による文書処理モジュール 122 2 のさらなる構造例を示す。



この例では、視覚情報 470 が文書 170 からさらに決定され、意味論的特徴表現 180 を生成するために利用される。視覚情報 470 は、文書 170 内に提示されるテキスト要素 220 のそれぞれの視覚的外観を示す情報を含む。また視覚情報 470 は、文書 170 の全体的な視覚的外観を示す情報を含む。

個々のテキスト要素 220 および文書 170 の視覚的外観は、異なるフォント、サイズ、テキストの方向、タイプ、斜体、色、下線、その他ハイライト、余白、ページの向き、インデント、間隔、またはその他の適用可能な形式など、テキスト要素 220 または文書 170 に適用される異なるフォーマットにより変化する。

3.クレーム

492 特許のクレーム 1 は以下の通りである。

1.電子装置において、

プロセッシングユニットと、

プロセッシングユニットに接続され、命令が格納されているメモリと、

命令がプロセッシングユニットによって実行されると、デバイスに以下を含む動作を実行させ、

文書内に提示される複数のテキスト要素を含むテキスト情報を決定し、

文書内に提示される複数のテキスト要素の空間的配置を示すレイアウト情報を決定

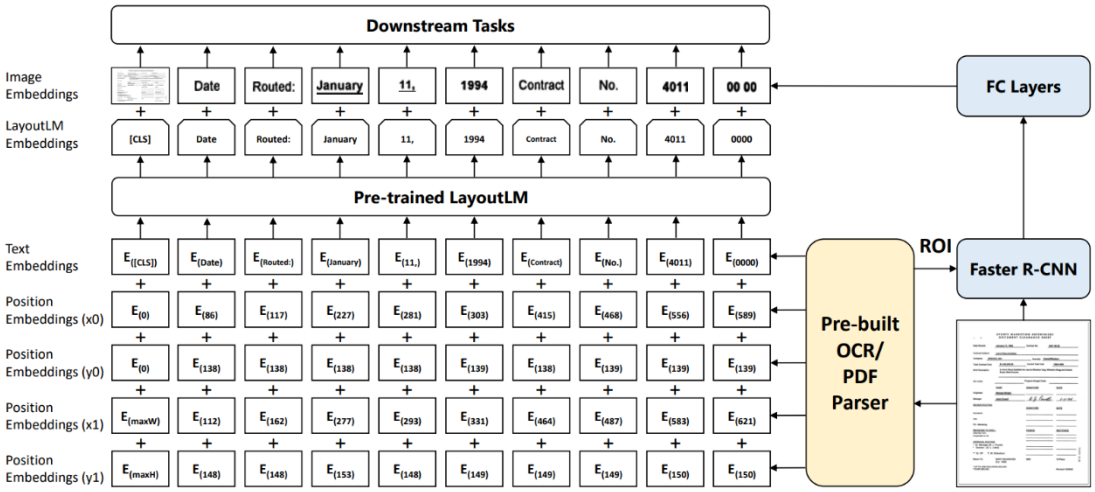
し、

複数のテキスト要素およびレイアウト情報に少なくとも部分的に基づいて、複数のテキスト要素のそれぞれの意味論的特徴表現を生成する。

4. 本特許に関連する論文

本特許に関する論文“LayoutLM: Pre-training of Text and Layout for Document Image Understanding”[1]が、Yiheng Xu 氏らにより公表されている。

本論文では、スキャンされた文書画像全体にわたるテキストとレイアウト情報との間の相互作用を共同でモデル化するための LayoutLM を提案している。下記図は LayoutLM モデルの構造を示す。



上記 LayoutLM は、2 次元レイアウトと画像埋め込みがオリジナルの BERT アーキテクチャに統合されている。LayoutLM 埋め込みと、Faster R-CNN から LayoutLM の後段で入力される画像埋め込みは、ダウンストリームタスクで連携して機能する。

下記テーブルに評価結果を示す。

[1] Yiheng Xu et al. “LayoutLM: Pre-training of Text and Layout for Document Image Understanding” arXiv:1912.13318v5 [cs.CL] 16 Jun 2020

Modality	Model	Precision	Recall	F1	#Parameters
Text only	$\text{BERT}_{\text{BASE}}$	0.5469	0.671	0.6026	110M
	$\text{RoBERTa}_{\text{BASE}}$	0.6349	0.6975	0.6648	125M
	$\text{BERT}_{\text{LARGE}}$	0.6113	0.7085	0.6563	340M
	$\text{RoBERTa}_{\text{LARGE}}$	0.678	0.7391	0.7072	355M
Text + Layout MVLM	$\text{LayoutLM}_{\text{BASE}}$ (500K, 6 epochs)	0.665	0.7355	0.6985	113M
	$\text{LayoutLM}_{\text{BASE}}$ (1M, 6 epochs)	0.6909	0.7735	0.7299	113M
	$\text{LayoutLM}_{\text{BASE}}$ (2M, 6 epochs)	0.7377	0.782	0.7592	113M
	$\text{LayoutLM}_{\text{BASE}}$ (11M, 2 epochs)	0.7597	0.8155	0.7866	113M
Text + Layout MVLM+MDC	$\text{LayoutLM}_{\text{BASE}}$ (1M, 6 epochs)	0.7076	0.7695	0.7372	113M
	$\text{LayoutLM}_{\text{BASE}}$ (11M, 1 epoch)	0.7194	0.7780	0.7475	113M
Text + Layout MVLM	$\text{LayoutLM}_{\text{LARGE}}$ (1M, 6 epochs)	0.7171	0.805	0.7585	343M
	$\text{LayoutLM}_{\text{LARGE}}$ (11M, 1 epoch)	0.7536	0.806	0.7789	343M
Text + Layout + Image MVLM	$\text{LayoutLM}_{\text{BASE}}$ (1M, 6 epochs)	0.7101	0.7815	0.7441	160M
	$\text{LayoutLM}_{\text{BASE}}$ (11M, 2 epochs)	**0.7677**	**0.8195**	**0.7927**	160M

Table 1: Model accuracy (Precision, Recall, F1) on the FUNSD dataset

FUNSD データセットでフォーム理解タスクを評価する。LayoutLM モデルを、SOTA で事前トレーニングされた 2 つの NLP モデル(BERT および RoBERTa)と比較する。BERT と比較して、RoBERTa はより多くのエポックを持つより大きなデータを使用してトレーニングされるため、このデータセットではるかに優れたパフォーマンスを発揮する。

LayoutLM には 4 つの設定があり、6 エポックで 500K ドキュメントページ、6 エポックで 1M、6 エポックで 2M、および 2 エポックで 11M である。LayoutLM モデルは、既存の SOTA 事前トレーニングベースラインを大幅に上回るパフォーマンスを示していることがわかる。

BASE アーキテクチャを使用すると、1,100 万のトレーニングデータを含む LayoutLM モデルは F1 スコアで 0.7866 を達成する。これは、同様のサイズのパラメータを持つ BERT や RoBERTa よりもはるかに高くなる。

さらに、事前トレーニングステップで MDC(Multi-label Document Classification)損失も追加すると、FUNSD データセットに大幅な改善がもたらされる。最後に、LayoutLM モデルは、テキスト、レイアウト、画像情報を同時に使用した場合に、0.7927 という最高のパフォーマンスを達成する。

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 3」、「ブロックチェーン 3.0(共著)」がある。