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

2019/07/02

Hiroshi Nagaya

論文紹介

タイトル

Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”

採択会議

The 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2019)

BERT:
この子の名前にちなんでいる

Abstract

- BERTという，Transformerの双方向的な訓練を活用した汎用言語モデルの提案（by Google AI Language）.
- ラベル付けされていないテキストデータからdeep bidirectional representationsを事前学習することによって，言語推論問題への回答などを可能にしている.

何がすごいのか

- 11個のNLPタスクにおいて，SOTA（State-of-the-Art）を達成
- 事前学習（汎用）モデルがあれば，数時間の追加学習（Fine-tuning）で実現可能

タスク	概要	前SOTA	BERT
GLUE	8種の言語理解タスク	75.2	81.9
1. MNLI	2入力文の含意/矛盾/中立を判定	82.1	86.7
2. QQP	2質問文が意味的に等価か判定	70.3	72.1
3. QNLI	SQuADの改変．陳述文が質問文の解答を含むか判定	88.1	91.1
4. SST-2	映画レビューの入力文のネガポジを判定	91.3	94.9
5. CoLA	入力文が言語的に正しいか判定	45.4	60.5
6. STS-B	ニュース見出しの2入力文の意味的類似性をスコア付け	80.0	86.5
7. MRPC	ニュース記事の2入力文の意味的等価性を判定	82.3	89.3
8. RTE	2入力文の含意を判定	56.0	70.1
SQuAD	質疑応答タスク．陳述文から質問文の解答を抽出	91.7	93.2
CoNLL	固有表現抽出タスク．単語に人物/組織/位置のタグ付け	92.6	92.8
SWAG	入力文に後続する文を4つの候補文から選択	59.2	86.3

GLEU（8種のNLPタスク，スコアは平均）

Rank	Name	Model	URL	Score
1	Jacob Devlin	BERT: 24-layers, 1024-hidden, 1		80.4
2	Alec Radford	Singletask Pretrain Transformer OpenAI Transformer		72.8
3	Samuel Bowman	BiLSTM+ELMo+Attn ELMo		70.5

SQuAD（質疑応答タスク）

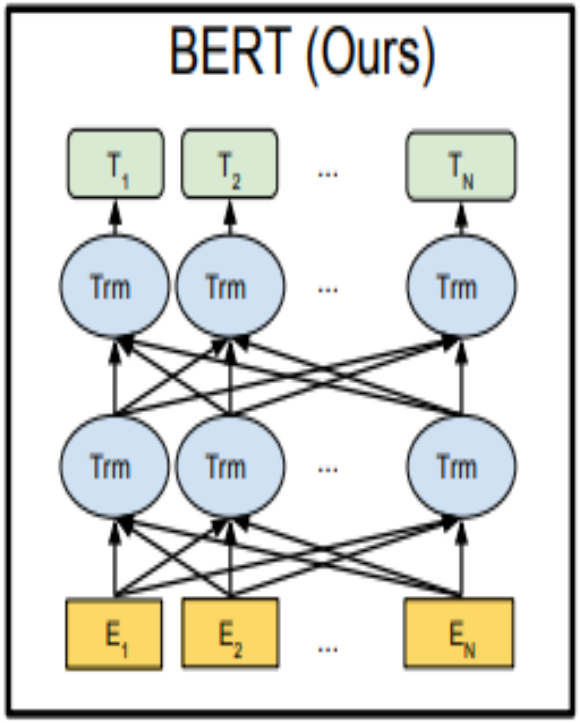
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	87.433	93.160
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677

BERT（Model Architecture）

- 右図のように，Transformerを活用したモデル

- 入力文字列をシーケンシャルに（つまり左→右 or 右→左）に読み込む指向的モデルとは違い，Transformerのエンコーダは文字列の単語を一度に読み込む

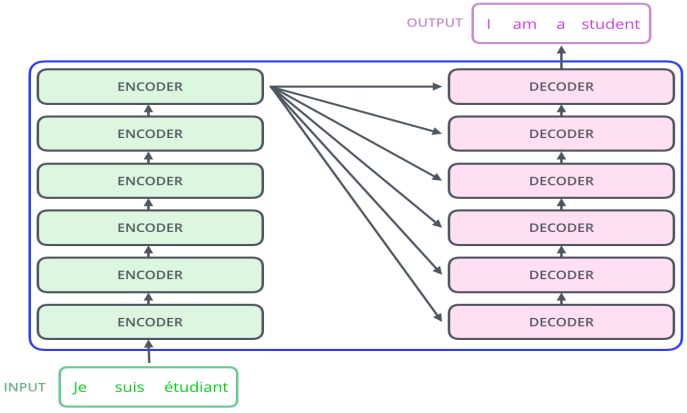
こうした非指向的な特徴によって，言語モデルが単語の周囲全体（単語の左と右）に基づいて文脈を学習可能に！



Transformer

- ひとつは入力文字列を読み込むエンコーダ
- 何らかのタスクのために予測を生成するデコーダ

互いに分離されたふたつのメカニズムを含むモデル

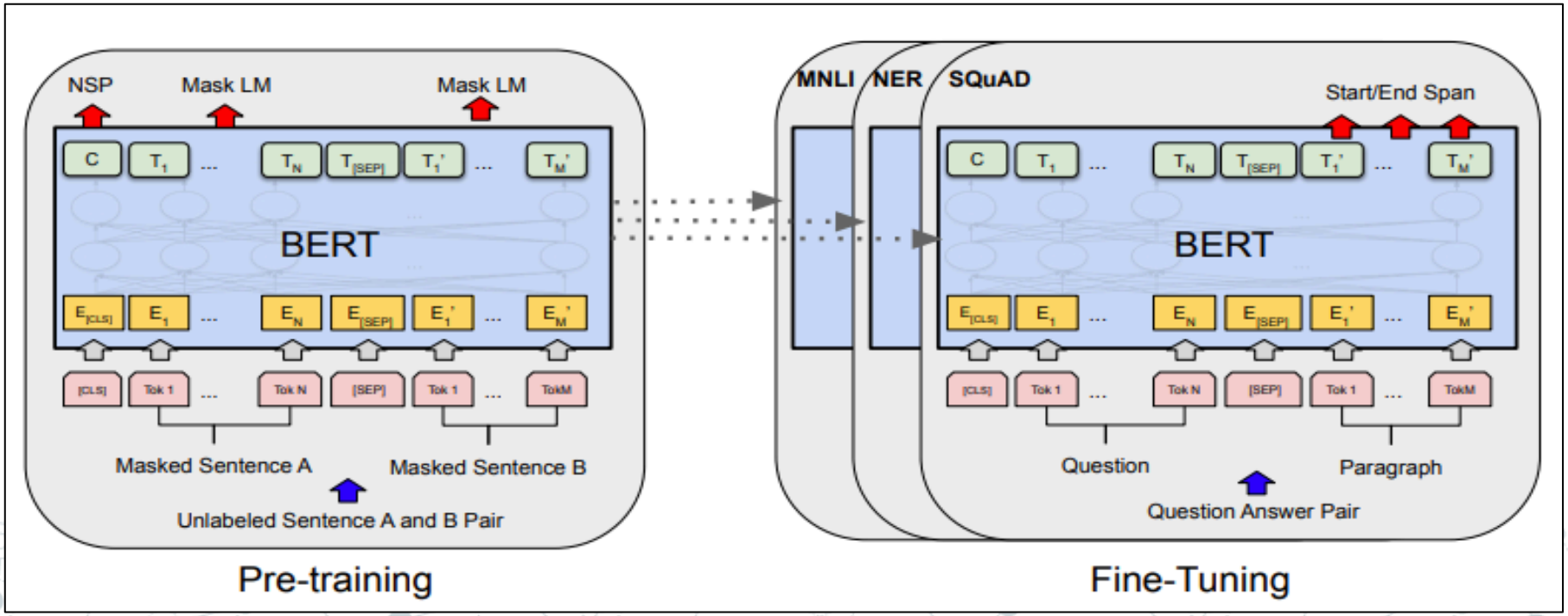


BERT（Learning）

Pre-trainingとFine-tuningの2つに学習パートが分かれている.

Pre-training
ラベル無しデータを用いたいくつかの事前学習タスクによって学習.

Fine-tuning
はじめに事前学習パラメータによって初期化されたのち，ラベル付きデータを用いたdownstreamタスクによってそのパラメータをチューニング.



次の2つのタスクにおける損失関数の結合が最小化されるように事前学習モデルを学習

- Masked LM：穴埋め推測
- Next Sentence Prediction：隣接文予測

【使用データ】
- Books Corpus (800M words) とEnglish Wikipedia (2,500M words)を使用.
- Wikipediaからは本文を抜き出し，リストやヘッダーは除外．Billion Word Benchmarkのような文単位のデータよりも長くて連続的な文書単位のデータが効果的.

学習タスク1（Masked LM：穴埋め推測）

- 次に来る単語や前の単語ではなく，ランダムにmask（隠）された単語を予測．単語は以下の三つの規則に従って，15%ほどの単語が[MASK]に置換．

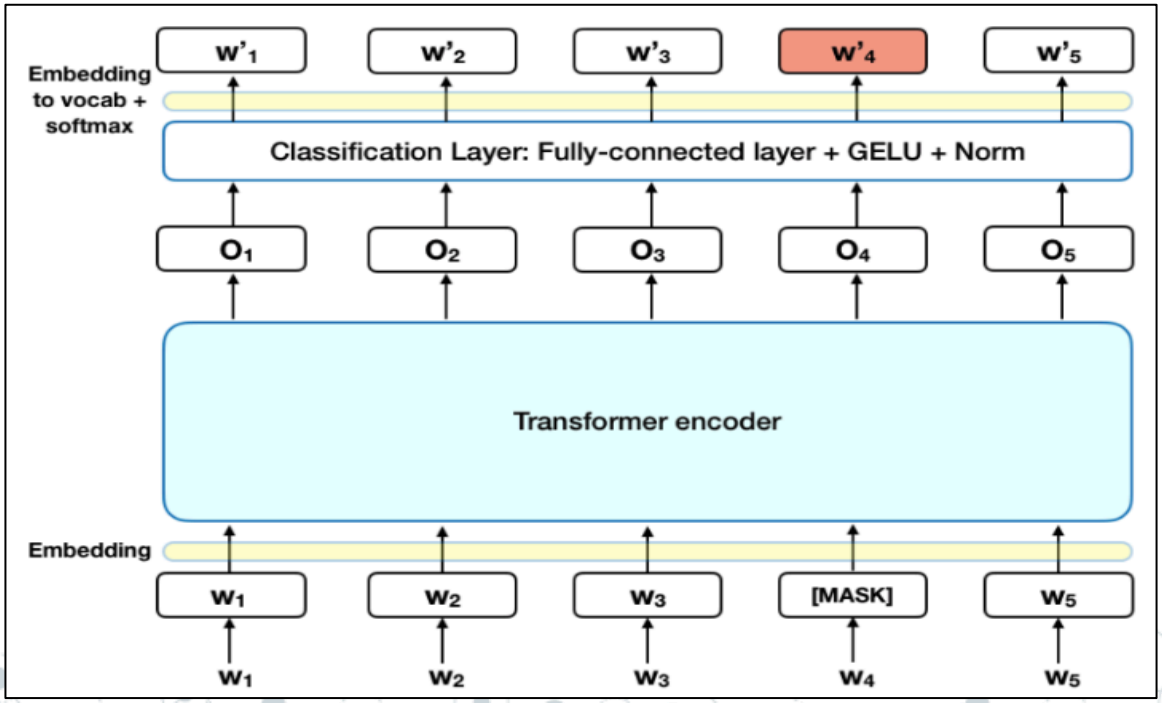
- 80% of the time: Replace the word with the [MASK] token, e.g., `my dog is hairy` → `my dog is [MASK]`
- 10% of the time: Replace the word with a random word, e.g., `my dog is hairy` → `my dog is apple`
- 10% of the time: Keep the word unchanged, e.g., `my dog is hairy` → `my dog is hairy`. The purpose of this is to bias the representation towards the actual observed word.

学習タスク1（Masked LM：穴埋め推測）

1. エンコーダからの出力のすぐ上に分類レイヤーを追加

2. 出力として受け取ったベクトルと埋め込み行列をかけ合わせ，出力を単語次元（vocabulary dimension）に変換

3. ソフトマックス関数を使って出力されたそれぞれの単語の確率を計算



学習タスク2（Next Sentence Prediction：隣接文予測）

- 入力として文のペアを受け取り，ペアの文における2つ目の文Bが（学習データとなる）オリジナルの文書において後続の文になっているかどうかを予測するように学習
- Bは，50%の確率でコーパスからランダムに選択された文章）

Next Sentence Prediction The next sentence prediction task can be illustrated in the following examples.

Input = [CLS] the man went to [MASK] store [SEP]
he bought a gallon [MASK] milk [SEP]
Label = IsNext

Input = [CLS] the man [MASK] to the store [SEP]
penguin [MASK] are flight ##less birds [SEP]
Label = NotNext

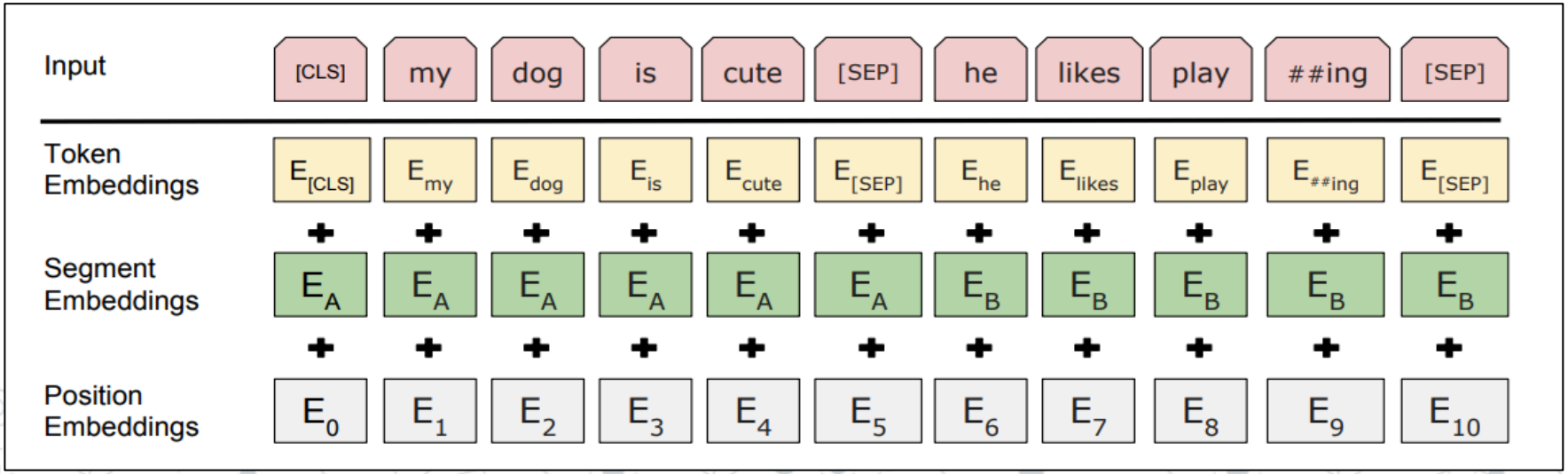
Pre-training

学習タスク2（Next Sentence Prediction：隣接文予測）
入力のペアとなっているふたつの文を区別するために，予め以下のように処理

Token Embedding
…冒頭に[CLS]トークンを挿入、またそれぞれの文の末尾に[SEP]トークンを挿入

Sentence Embedding
…文A（最初の文）あるいは文B（後続の文）であることを指し示す情報を追加

Positional Embedding
…文における位置を指し示すために、各トークンに位置的埋め込み情報を追加

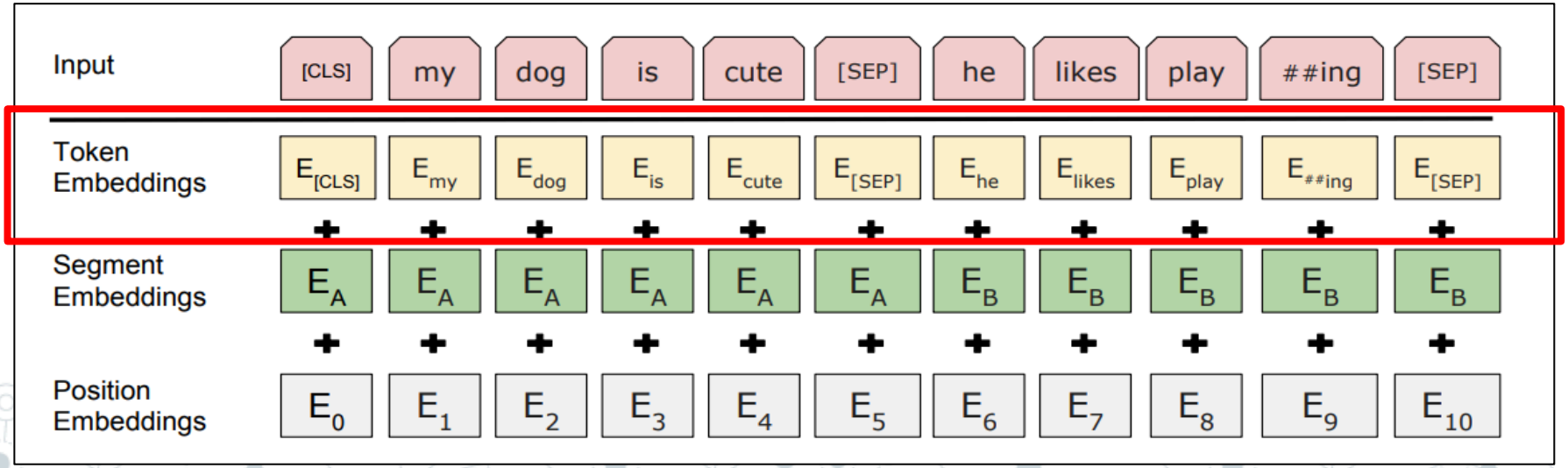


学習タスク2（Next Sentence Prediction：隣接文予測）
入力のペアとなっているふたつの文を区別するために，予め以下のように処理

Token Embedding
…冒頭に[CLS]トークンを挿入、またそれぞれの文の末尾に[SEP]トークンを挿入

Sentence Embedding
…文A（最初の文）あるいは文B（後続の文）であることを指し示す情報を追加

Positional Embedding
…文における位置を指し示すために、各トークンに位置的埋め込み情報を追加

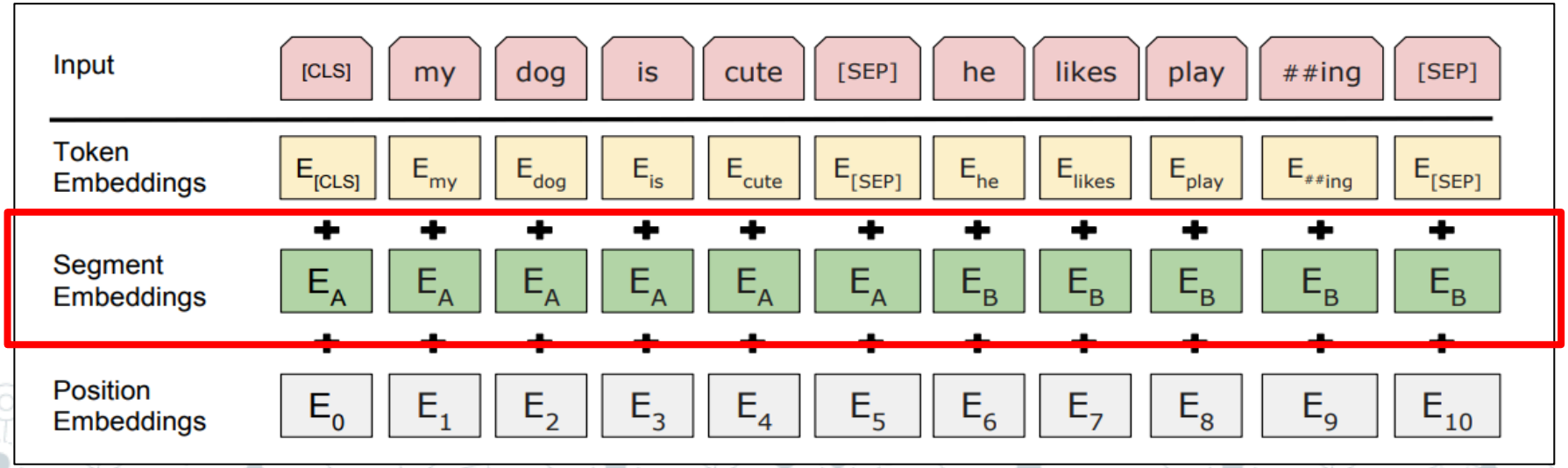


学習タスク2（Next Sentence Prediction：隣接文予測）
入力のペアとなっているふたつの文を区別するために，予め以下のように処理

Token Embedding
…冒頭に[CLS]トークンを挿入、またそれぞれの文の末尾に[SEP]トークンを挿入

Sentence Embedding
…文A（最初の文）あるいは文B（後続の文）であることを指し示す情報を追加

Positional Embedding
…文における位置を指し示すために、各トークンに位置的埋め込み情報を追加

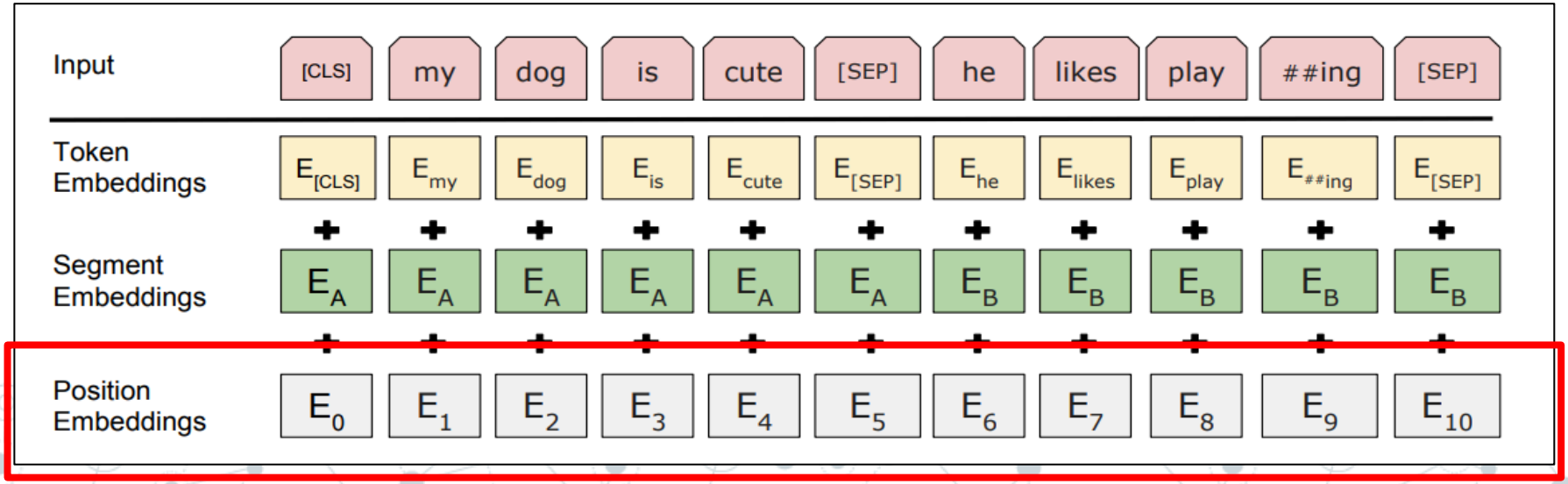


学習タスク2（Next Sentence Prediction：隣接文予測）
入力のペアとなっているふたつの文を区別するために，予め以下のように処理

Token Embedding
…冒頭に[CLS]トークンを挿入、またそれぞれの文の末尾に[SEP]トークンを挿入

Sentence Embedding
…文A（最初の文）あるいは文B（後続の文）であることを指し示す情報を追加

Positional Embedding
…文における位置を指し示すために、各トークンに位置的埋め込み情報を追加



学習タスク2（Next Sentence Prediction：隣接文予測）
予測は，次のようなステップで実行

1. 入力となるシーケンスのすべてを，Transformerモデルに渡す

2. （BERTにおける重みとバイアスを表す行列について学習した）単純分類レイヤーを使って、出力における[CLS]トークンを，2 x 1のベクトルに変換

3. ソフトマックス関数を使って、IsNextSequence（その文が次文であるかどうかを表す値）の確率を計算

Fine-tuning

タスクごとに特有な入力と出力に与えて，事前学習モデルのパラメータを調整
BERTモデルのコアに小さい層を追加するだけで，幅広い種類のNLPタスクに利用可能

センチメント分析
[CLS]トークンのためにあるTransformerの出力のうえに分類レイヤーを追加することによって、次文分類と同じように実行

質疑応答タスク（SQuAD v1.1）
質問を入力となる文字列シーケンスを受け取り，回答のはじめとおわりの目印となるふたつのベクトルを学習することで可能

固有表現抽出（Named Entity Recognition：NER）
文字列シーケンスを受け取ってから，文字列中に現れる（人名、組織名、日付などのような）様々な種類の固有表現に目印をつけるように学習する．各トークンの出力ベクトルを固有表現抽出ラベルを予測する分類レイヤーに入力として与えることにより実行

Fine-tuning

タスクごとに特有な入力と出力に与えて，事前学習モデルのパラメータを調整
BERTモデルのコアに小さい層を追加するだけで，幅広い種類のNLPタスクに利用可能

センチメント分析
[CLS]トークンのためにあるTransformerの出力のうえに分類レイヤーを追加することによって、次文分類と同じように実行

質疑応答タスク（SQuAD v1.1）
質問を入力となる文字列シーケンスを受け取り，回答のはじめとおわりの目印となるふたつのベクトルを学習することで可能

固有表現抽出（Named Entity Recognition：NER）
文字列シーケンスを受け取ってから，文字列中に現れる（人名、組織名、日付などのような）様々な種類の固有表現に目印をつけるように学習する．各トークンの出力ベクトルを固有表現抽出ラベルを予測する分類レイヤーに入力として与えることにより実行

Fine-tuning

タスクごとに特有な入力と出力に与えて，事前学習モデルのパラメータを調整
BERTモデルのコアに小さい層を追加するだけで，幅広い種類のNLPタスクに利用可能

センチメント分析
[CLS]トークンのためにあるTransformerの出力のうえに分類レイヤーを追加することによって、次文分類と同じように実行

質疑応答タスク（SQuAD v1.1）
質問を入力となる文字列シーケンスを受け取り，回答のはじめとおわりの目印となるふたつのベクトルを学習することで可能

固有表現抽出（Named Entity Recognition：NER）
文字列シーケンスを受け取ってから，文字列中に現れる（人名、組織名、日付などのような）様々な種類の固有表現に目印をつけるように学習する．各トークンの出力ベクトルを固有表現抽出ラベルを予測する分類レイヤーに入力として与えることにより実行

Fine-tuning

タスクごとに特有な入力と出力に与えて，事前学習モデルのパラメータを調整
BERTモデルのコアに小さい層を追加するだけで，幅広い種類のNLPタスクに利用可能

センチメント分析
[CLS]トークンのためにあるTransformerの出力のうえに分類レイヤーを追加することによって、次文分類と同じように実行

質疑応答タスク（SQuAD v1.1）
質問を入力となる文字列シーケンスを受け取り，回答のはじめとおわりの目印となるふたつのベクトルを学習することで可能

固有表現抽出（Named Entity Recognition：NER）
文字列シーケンスを受け取ってから，文字列中に現れる（人名、組織名、日付などのような）様々な種類の固有表現に目印をつけるように学習する．各トークンの出力ベクトルを固有表現抽出ラベルを予測する分類レイヤーに入力として与えることにより実行

検証

11個のNLPタスクにおいて，SOTA（State-of-the-Art）を達成

タスク	概要	前SOTA	BERT
GLUE	8種の言語理解タスク	75.2	81.9
1. MNLI	2入力文の含意/矛盾/中立を判定	82.1	86.7
2. QQP	2質問文が意味的に等価か判定	70.3	72.1
3. QNLI	SQuADの改変．陳述文が質問文の解答を含むか判定	88.1	91.1
4. SST-2	映画レビューの入力文のネガポジを判定	91.3	94.9
5. CoLA	入力文が言語的に正しいか判定	45.4	60.5
6. STS-B	ニュース見出しの2入力文の意味的類似性をスコア付け	80.0	86.5
7. MRPC	ニュース記事の2入力文の意味的等価性を判定	82.3	89.3
8. RTE	2入力文の含意を判定	56.0	70.1
SQuAD	質疑応答タスク．陳述文から質問文の解答を抽出	91.7	93.2
CoNLL	固有表現抽出タスク．単語に人物/組織/位置のタグ付け	92.6	92.8
SWAG	入力文に後続する文を4つの候補文から選択	59.2	86.3

GLEU（8種のNLPタスク，スコアは平均）

Rank	Name	Model	URL	Score
1	Jacob Devlin	BERT: 24-layers, 1024-hidden, 1		80.4
2	Alec Radford	Singletask Pretrain Transformer OpenAI Transformer		72.8
3	Samuel Bowman	BiLSTM+ELMo+Attn ELMo		70.5

SQuAD（質疑応答タスク）

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	87.433	93.160
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677

研究にどう活かすか

- BERTや関連モデル自体の改良を研究対象にするわけではない
- ただ，汎用性が高く様々なタスクに応用可能な点で魅力的なので，あくまでツールとして，事前学習モデルの活用とタスクに応じたFine-tuningという形で利用

【参考】

BERTの事前学習モデルと実装コードの配布（by Google）と，
- https://github.com/google-research/bert

そのファイル群の日本語解説
- https://qiita.com/uedake722/items/63a1ea32fe84886c02f9