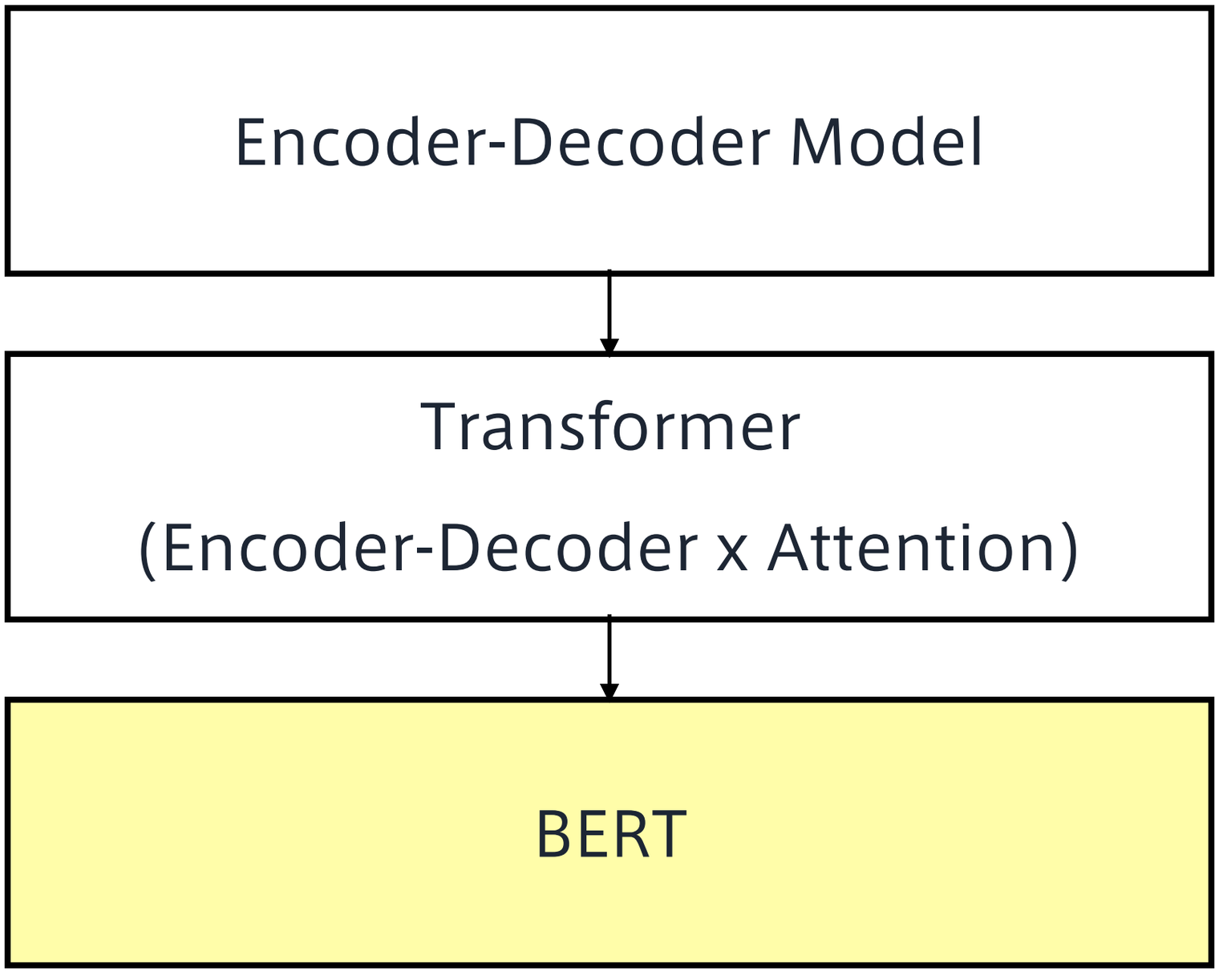
BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

2019/3/25

BERTまでのロードマップ

BERTを理解するために必要な材料



書誌情報

- Title
  - BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
- 投稿日：2018/10/11
- URL
  - ArXiv
- 著者
  - Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
  - Google AI Language
- 概要
  - Bidirectional Transformerをユニットにフルモデルで構成したモデル
  - 事前学習タスクとして、マスク単語予測タスク、隣接文判定タスクを与える
  - BERTからTransfer Learningを行った結果、8つのタスクでSOTA達成
  - Googleが事前学習済みモデルを公開済み / TensorFlow / PyTorch

アウトライン

1	背景
2	BERT
3	有効性

背景

様々な自然言語処理タスクにおいて事前学習が有効

- 文レベルのタスク：文同士の関係性が重要
    - 文章類似度
    - 言い換え
- トークンレベル：モデルはトークンレベルで良い出力が求められる
    - Named Entity Recognition
    - Q&A

事前学習には二種類のアプローチがある

- Feature-based
- Fine-tuning

Feature-based アプローチ

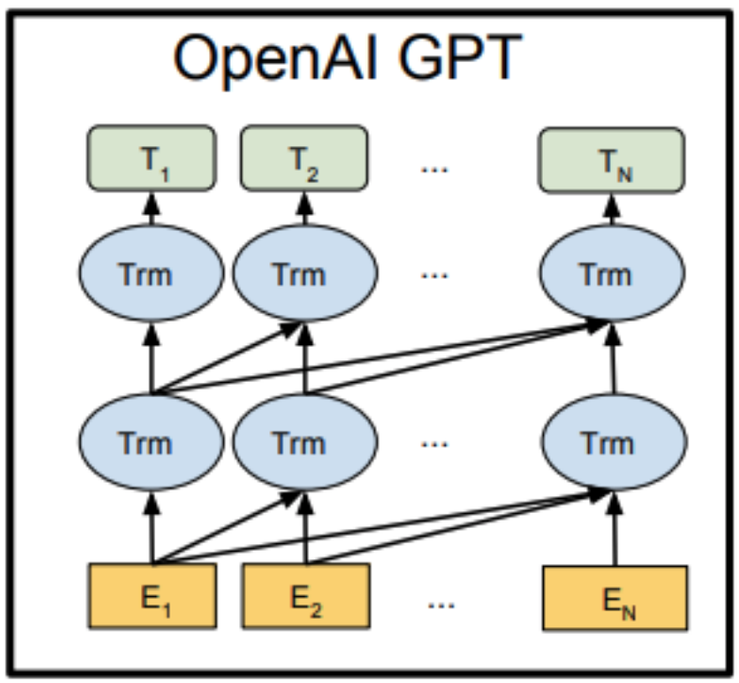
特徴量抽出機として活用するためのもの

- 様々なNLPタスクの素性として利用される
  - N-gramモデル[Brown et al., 1992]や Word2Vec[Malcov et al., 2013]など
    - 文や段落レベルの分散表現に拡張されたものもある
- 最近では、ElMo[Peters et al., 2017, 2018]が話題に
  - Context-Sensitiveな素性を抽出
  - 既存の素性にconcatして使うことで複数のNLPタスクでSOTA
    - QA, 極性分類, エンティティ認識

Fine-tuning アプローチ

言語モデルの目的関数で事前学習する

- 事前学習の後に、使いたいタスクの元で教師あり学習を行う（すなわち事前学習はパラメタの初期値として利用される）
- 最近では、OpenAI GPT[Ralford et al. 2018]など。
  - GLUEスコアでSOTA
  - TransformerのDecoderだけを利用する
    - 単方向接続モデル



アウトライン

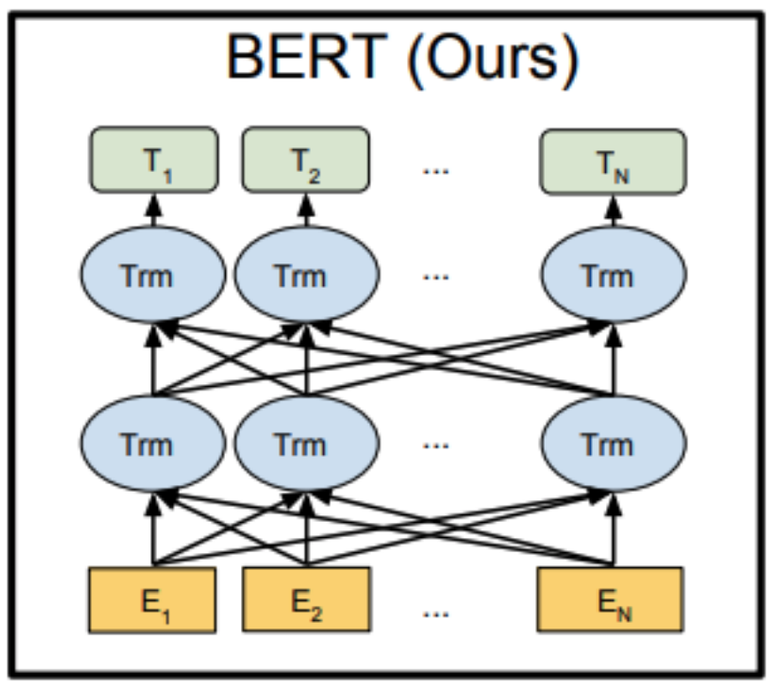
1	背景
2	BERT
3	有効性

BERT

Fine-tuningアプローチの事前学習に工夫を加えた

- 双方向Transformer

- tensorを入力としtensorを出力
- モデルの中に未来情報のリークを防ぐためのマスクが存在しない
  →従来のような言語モデル型の目的関数は採用できない(カンニングになるため)
  →事前学習タスクにおいて工夫する必要がある



Transformer-Encoder

自己注意機構により文脈を考慮して各単語をエンコード

6層

$r_1^{(6)}$ $r_2^{(6)}$ $r_3^{(6)}$

自己注意機構 & Feed Forward NN

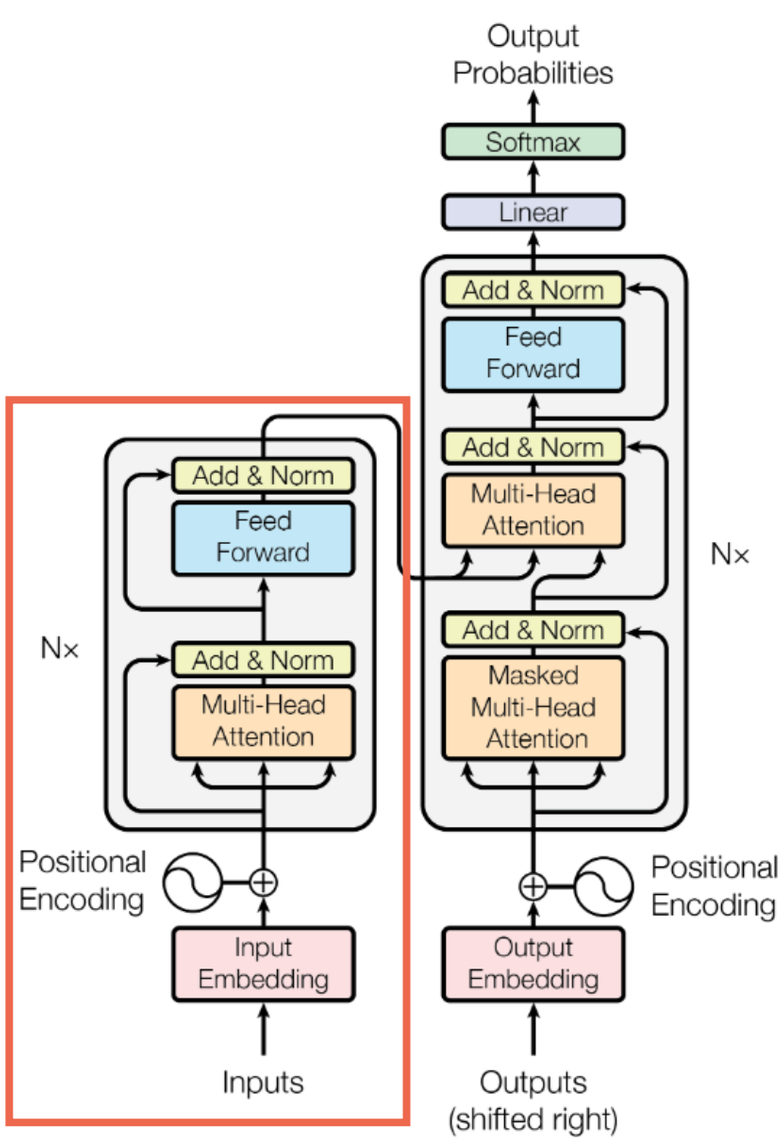
$r_1^{(5)}$ $r_2^{(5)}$ $r_3^{(5)}$

$r_1^{(1)}$ $r_2^{(1)}$ $r_3^{(1)}$

自己注意機構 & Feed Forward NN

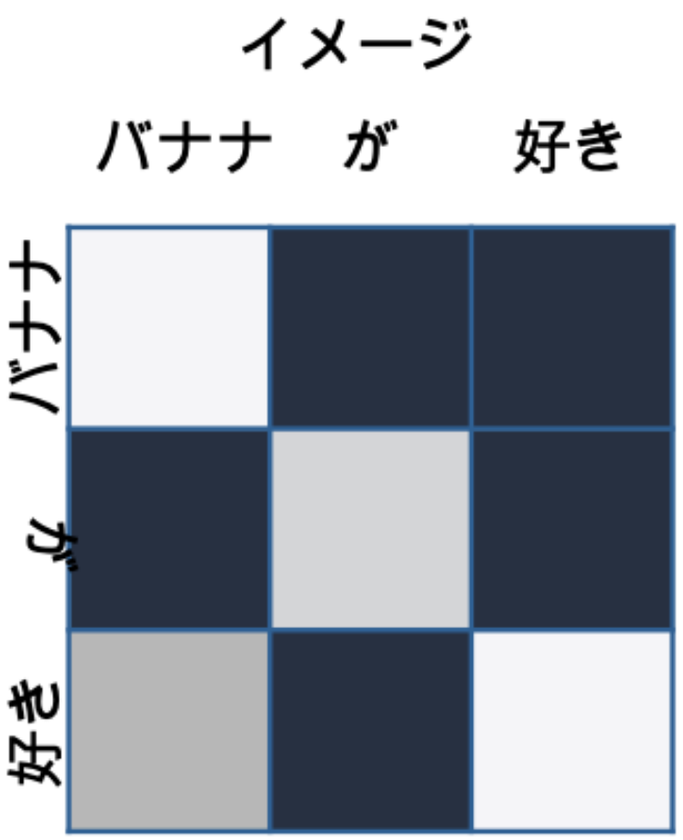
X_1 X_2 X_3

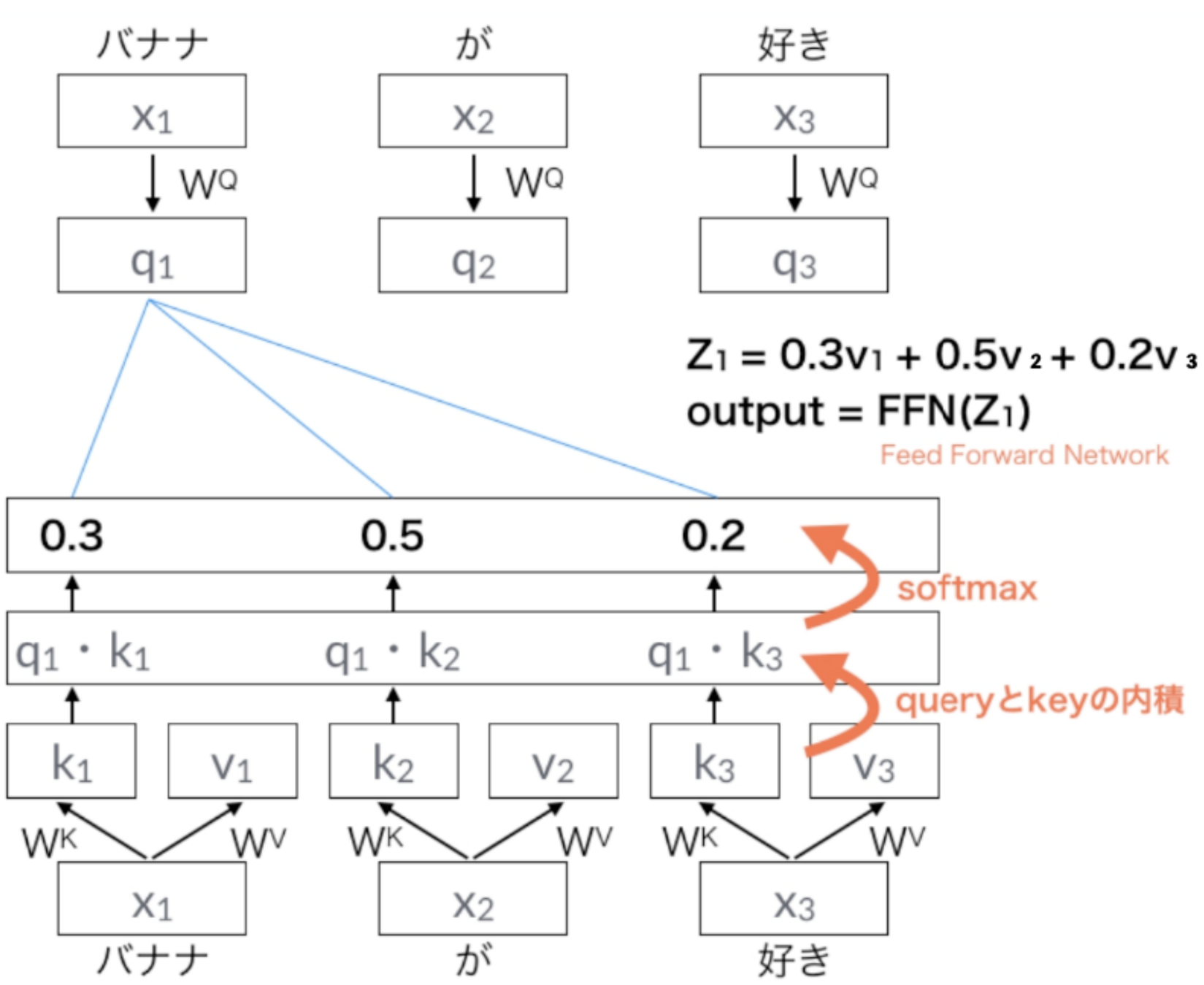
バナナ が 好き



Self-Attentionが肝

入力を全て同じにして学習的に注意箇所を決めていく





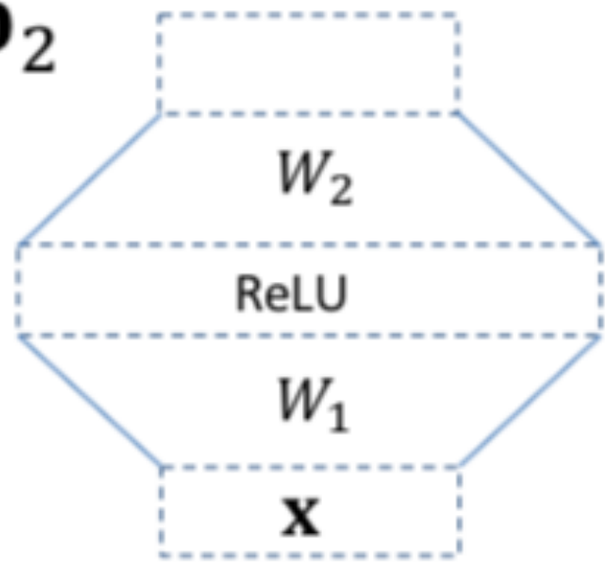
Position-Wise Feed-Forward Networks

位置情報を保持したまま順伝播させる

- 各Attention層の出力を決定
  - ２層の全結合NN
  - 線形変換→ReLu→線形変換

$$\mathrm{FFN}(\mathbf{x}) = \max(0, \mathbf{x}W_1 + \mathbf{b}_1)W_2 + \mathbf{b}_2$$

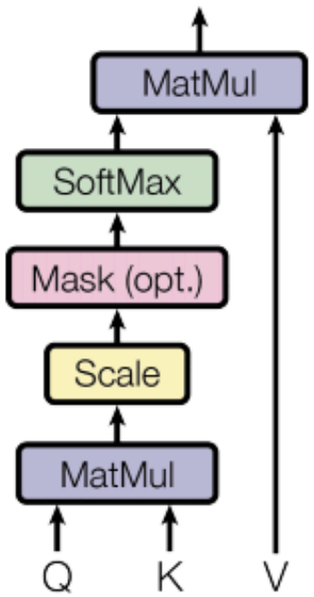
$$W_1 \in \mathbb{R}^{512\times2048} \quad \mathbf{b}_1 \in \mathbb{R}^{2048}$$
$$W_2 \in \mathbb{R}^{2048\times512} \quad \mathbf{b}_2 \in \mathbb{R}^{512}$$



Scaled dot product attention

全単語に関するAttentionをまとめて計算する

Scaled Dot-Product Attention



次元に応じてスケーリング

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$\begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} \begin{bmatrix} k_1 & k_2 & k_3 \end{bmatrix} = \begin{bmatrix} q_1 \cdot k_1 & q_1 \cdot k_2 & q_1 \cdot k_3 \\ q_2 \cdot k_1 & q_2 \cdot k_2 & q_2 \cdot k_3 \\ q_3 \cdot k_1 & q_3 \cdot k_2 & q_3 \cdot k_3 \end{bmatrix}$$

Q K

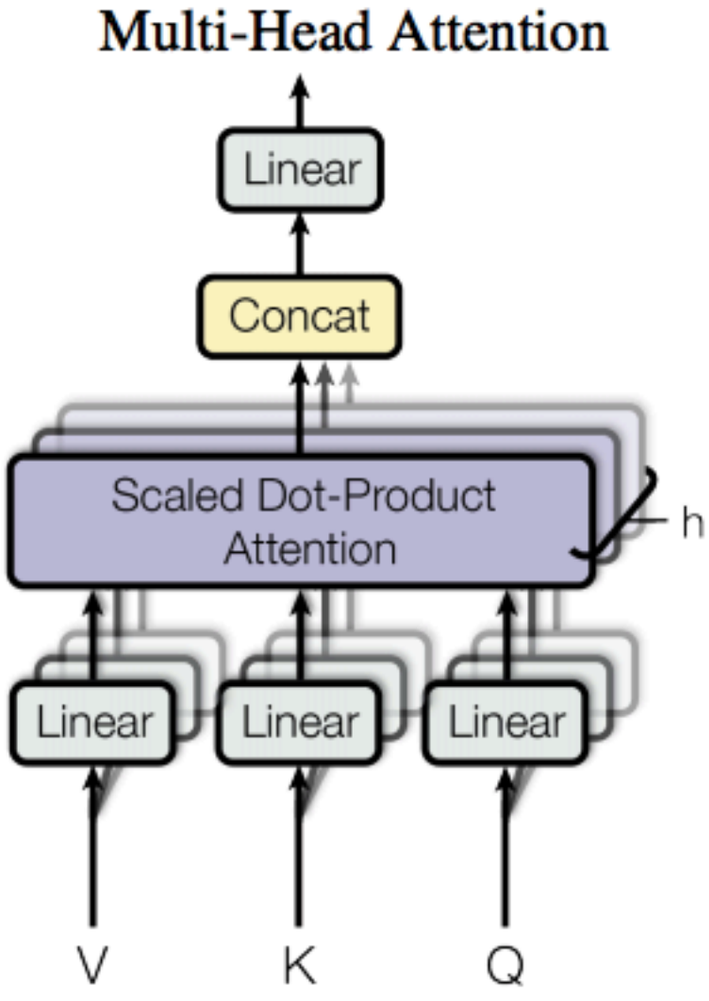
$$\begin{bmatrix} 0.1 & 0.2 & 0.7 \\ 0.4 & 0.3 & 0.3 \\ 0.8 & 0.1 & 0.1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \begin{bmatrix} 0.1v_1 + 0.2v_1 + 0.7v_1 \\ 0.4v_2 + 0.3v_2 + 0.3v_2 \\ 0.8v_3 + 0.1v_3 + 0.1v_3 \end{bmatrix}$$

$softmax(QK)$ V

【再掲】

Multi-Head attention

重みパラメタの異なる８個のヘッドを使用

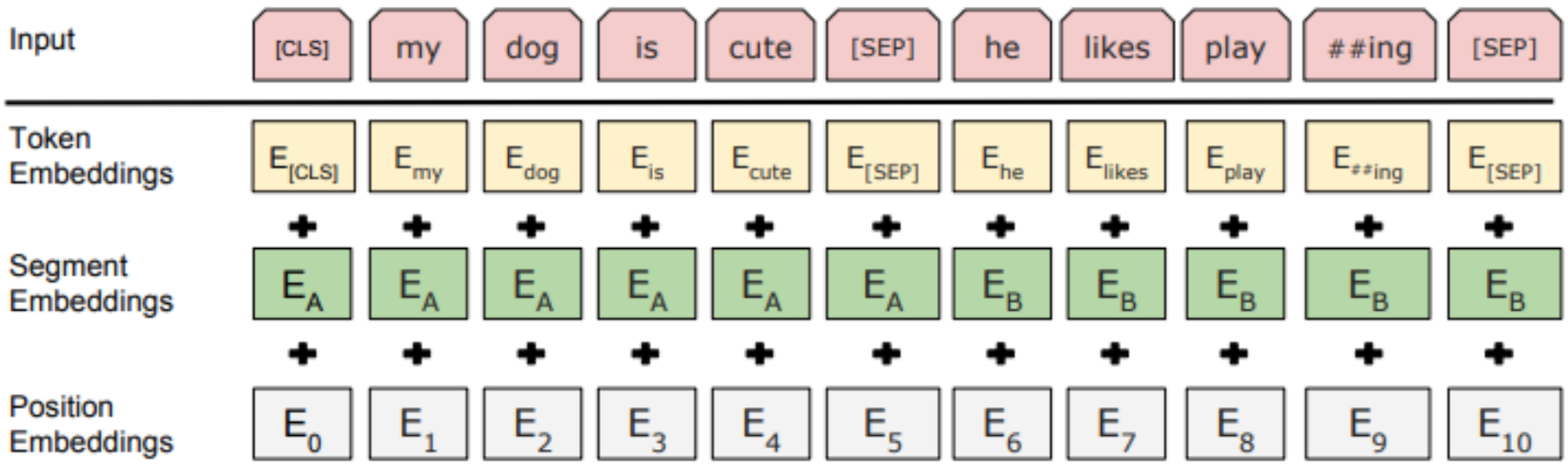


- ８個のScaled Dot-Product Attentionの出力をConcat
- それぞれのヘッドが異なる種類の情報を収集

入力表現

3種類のEmbeddingのSumを入力とする

- 文章のペア or ひとつの文章を入力にする (事前学習の際は必ずペア)
- 以下3種類のEmbeddingを使用する
  - トークン埋め込み：**WordPiece**でTokenizationしたものをEmbedding
  - 単語位置埋め込み：系列長1～512の表現
  - 文区別埋め込み：一文目、二分目の区別



事前学習 (Pre-training) タスク

空欄語予測と隣接文予測の２タスクを行う

空欄語予測と隣接文予測を同時に事前学習する

=> 単語分散表現と文章分散表現を同時に獲得できる

事前学習 (Pre-training) タスク

空欄語予測と隣接文予測の２タスクを行う

空欄語予測 (Masked Language Prediction)

```
Input: the man went to the [MASK1] . he bought a [MASK2] of milk.
Labels: [MASK1] = store; [MASK2] = gallon
```

文章中の単語のうち15%がMASK対象に選ばれる

=> 選ばれた15%の単語の位置いはフラグを立てておく

選ばれた単語の内、80%が[MASK]に置き換えられ、10%が他の単語に置き換えられ、残り10%は置き換えられずにおかれる

=> **文章を入力としてフラグが付いている位置のオリジナルの入力単語が何であるかを出力する** (15%対象外の単語に関しては扱わない)

事前学習 (Pre-training) タスク

空欄語予測と隣接文予測の２タスクを行う

空欄語予測 (Masked Language Prediction)

```
Input: the man went to the [MASK1] . he bought a [MASK2] of milk.
Labels: [MASK1] = store; [MASK2] = gallon
```

双方向モデルは、left-to-rightモデルや双方向にconcatしたものよりも一般的に強力

- しかし一般的な条件付き言語モデルでは学習できない
  - 特定の単語予測： $P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i | w_1, \ldots, w_{i-1})$
  - 双方向だと間接的に複数層の文脈から自分自身を見てしまう
- そこで、発想を転換し次の単語ではなくランダムに抜かれた単語を予測するモデルにする(CBOWの発想)
- 一方でこの手法を採用すると単語全体の15%しか学習材料に使えないため学習に時間がかかってしまう

事前学習 (Pre-training) タスク

空欄語予測と隣接文予測の２タスクを行う

隣接文予測 (Next Sentence Prediction)

```
Sentence A: the man went to the store .
Sentence B: he bought a gallon of milk .
Label: IsNextSentence
```

```
Sentence A: the man went to the store .
Sentence B: penguins are flightless .
Label: NotNextSentence
```

- ２つの文章の関係性理解も重要
- 言語モデルのタスクのみでは捉えきれない

２つの連なる文章ペアに対して、隣接文を50%の確立でシャッフルする

=> **二つの文章を入力として、隣接文であるかのT/Fを出力する**

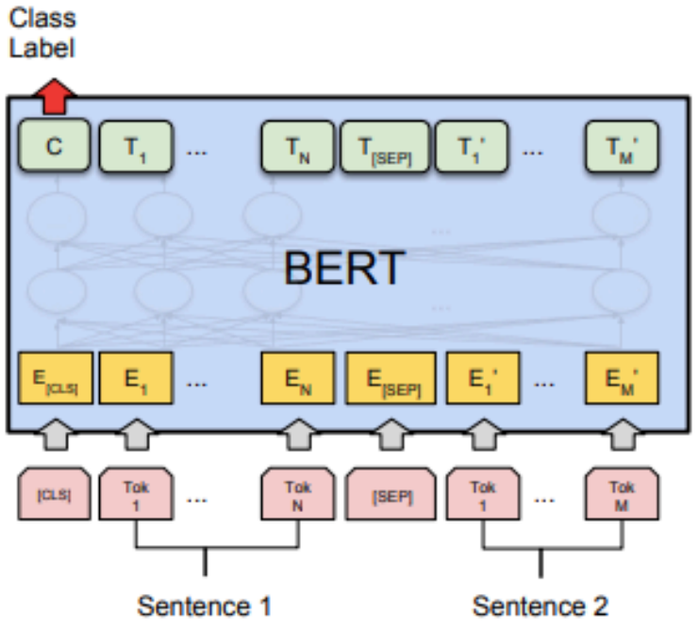
Pre-training 手続き

- データセット：**BooksCorpus**(800MB) + **English Wikipedia**(2500MB)
- 入力文章の合計系列長が512以下になるように2つの文章をサンプリング
  - Next Sentence Predictionのため、文章1と文章2の組み合わせは50%の確率で変わる
  - MLMのためWordPieceトークンに分けられた後マスクされる
- バッチサイズ：256 (= 256x512系列 = 128,000単語/バッチ)
  - **1,000,000**ステップ=33億の単語を40エポック学習
- Adam：LR=1e-4, L2weight_decay=0.01
- Dropout：0.1
- 活性化関数：**GeLu**

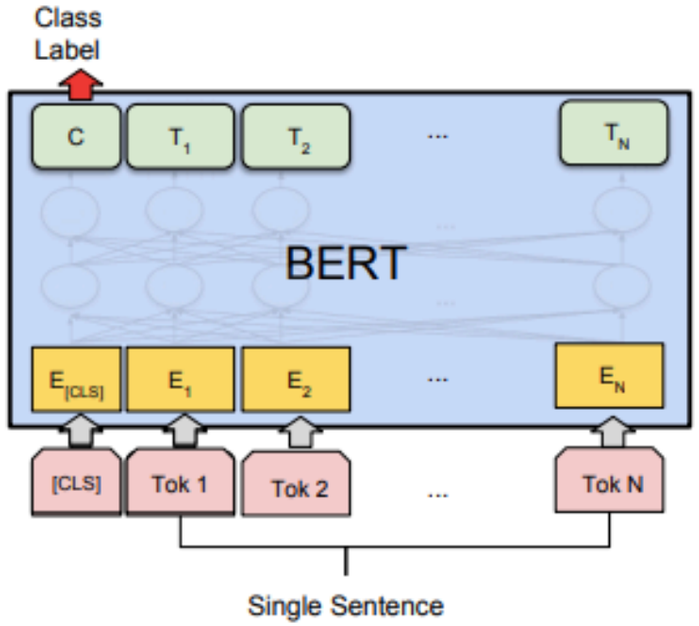
Pre-training 手続き

- 系列レベルの分類問題
  - 固定長の分散表現は最初の[CLS]トークンの内部表現から得られる
  - 新しく追加する層は分類全結合層+ソフトマックス層のみ
  - BERTも一緒に学習させる
- トークンレベルの分類問題 (NERなど)
  - 各タスクの仕様に合わせて学習させる
- Fine-tuning自体は高速にできるので、ハイパーパラメタ探索も可能

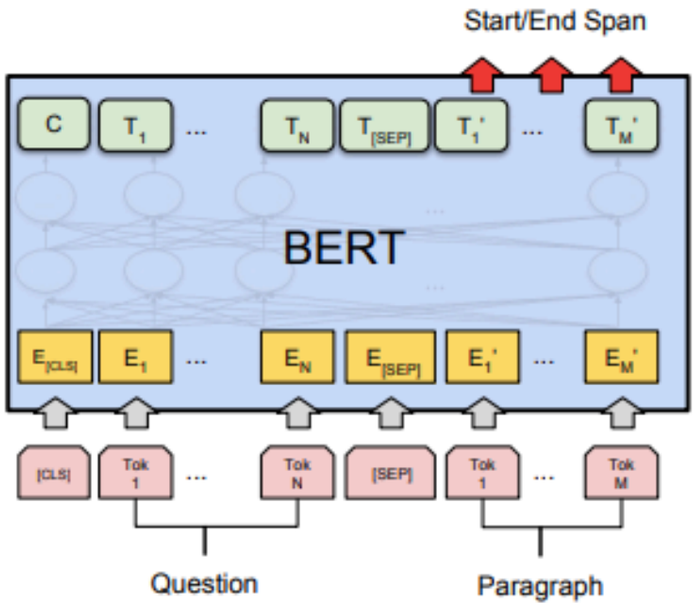
Fine-tuning概要



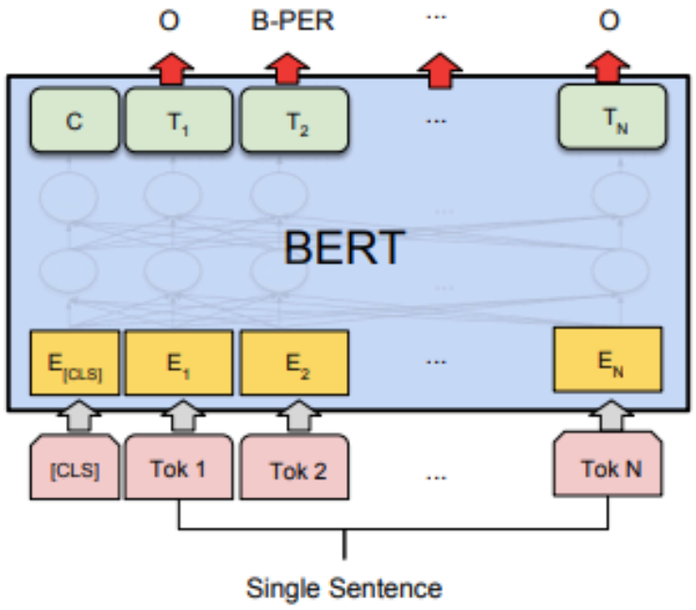
(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG



(b) Single Sentence Classification Tasks:
SST-2, CoLA



(c) Question Answering Tasks:
SQuAD v1.1



(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

アウトライン

1	背景
2	BERT
3	有効性

8個のNLPベンチマークタスク

タスクspecificなアーキテクチャを組むことなく一気にSOTA

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

<implementation>