Attention mechanism

Tatsuma Furuya

Takemura Lab

Attention mechanism

- 入力されたデータのどこに注目すべきか，動的に特定する仕組み
- NLPやComputer Vision，その他の系列データに対して効果的
- NLPでSoTAを達成したTransformerモデルやCVでSoTAを達成したVision Transformerモデルでも使用

Attentionが効果的な分野

- CNNの画像認識
- seq2seqの機械翻訳
- Transformerの機械翻訳
- Vision Transformer

画像認識タスクで効果的な処理とは？

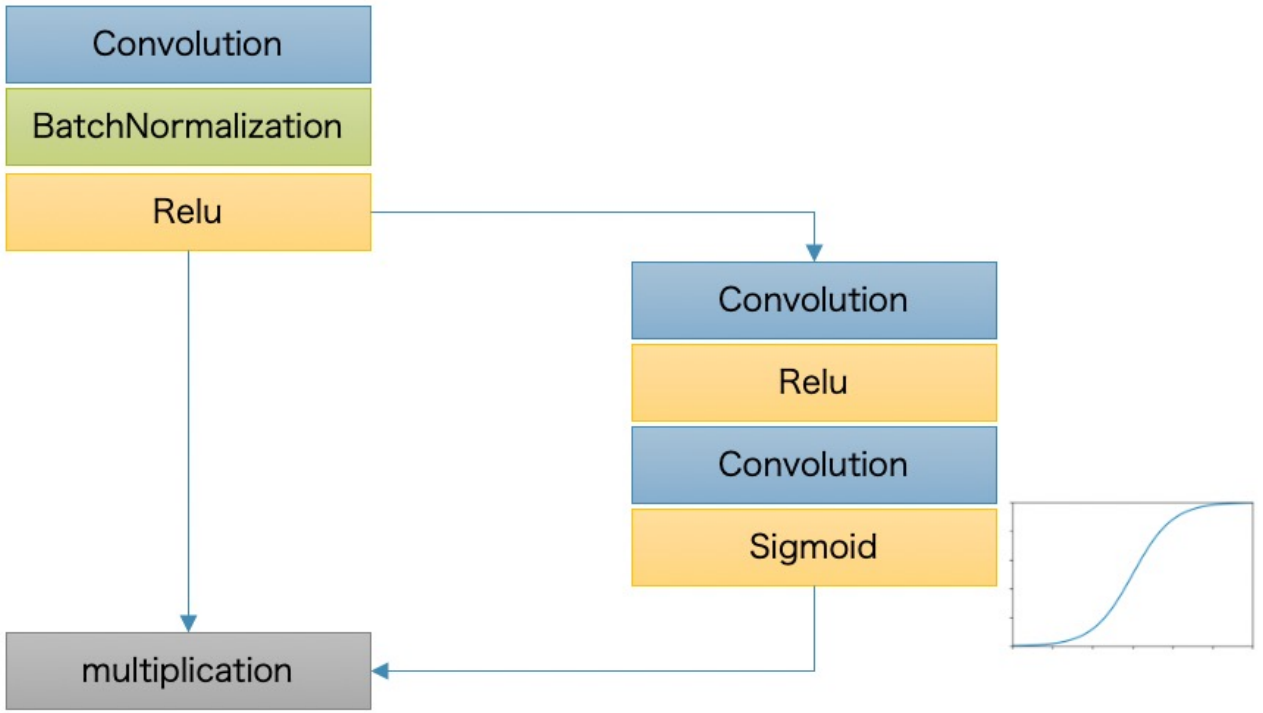
画像を全景と背景に分けて全景のみにフォーカス

問題点

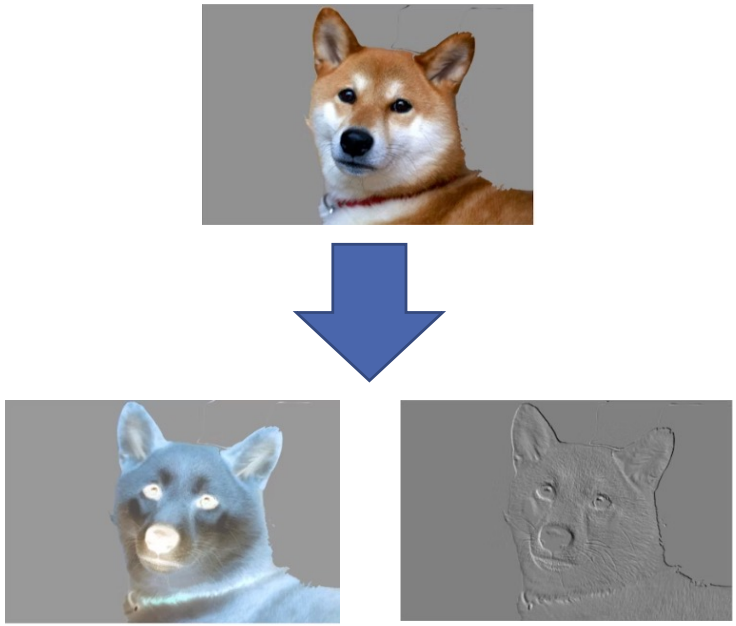
- 前処理で全景と背景を分けるモデルの精度に依存
- 計算コストが膨大

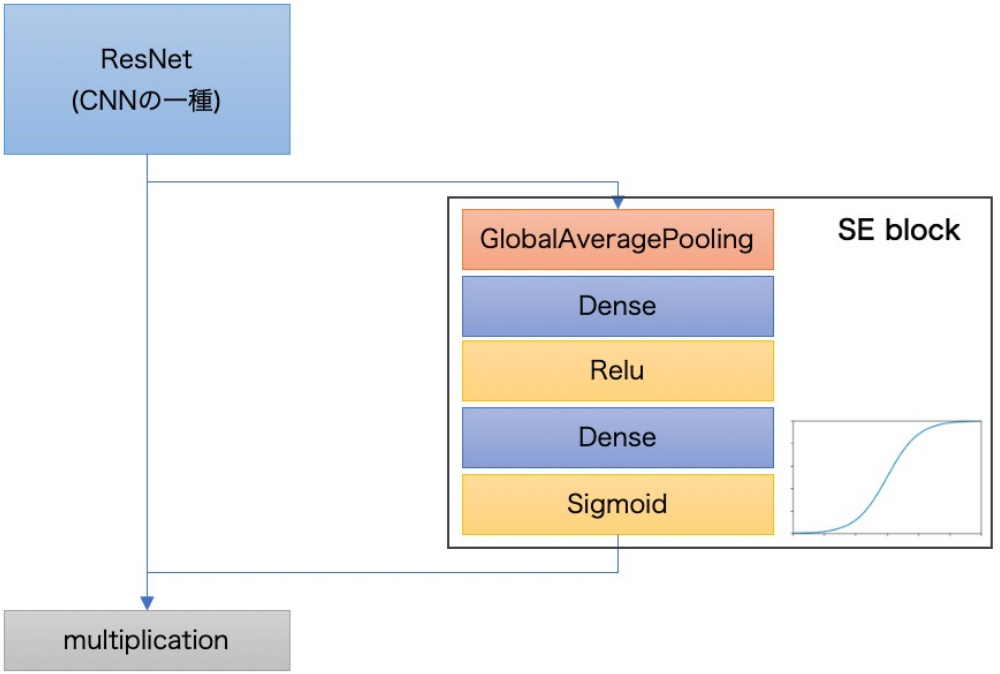
学習の中で注目すべき領域も学習(Attention mechanism)

最も単純な実装


Convolution
BatchNormalization
Relu
Convolution
Relu
Convolution
Sigmoid
multiplication

SENet ~画像の特徴に対するAttention~

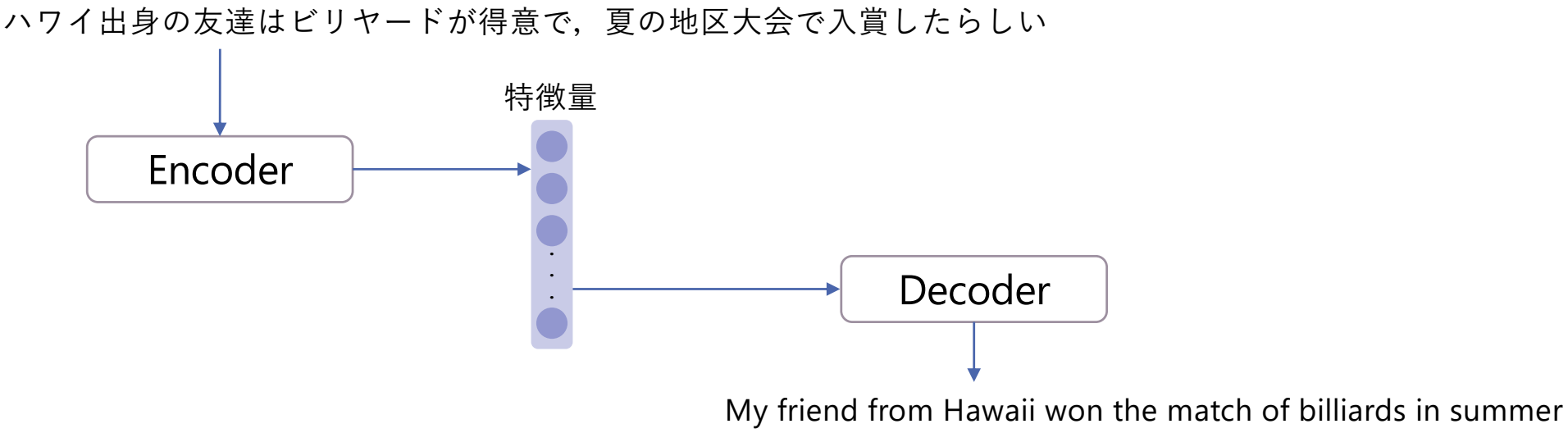




seq2seq

seq2seq

翻訳タスクのように系列データ(文章)から系列データ(文章)を推論するモデル.
エンコーダとデコーダから構成される.

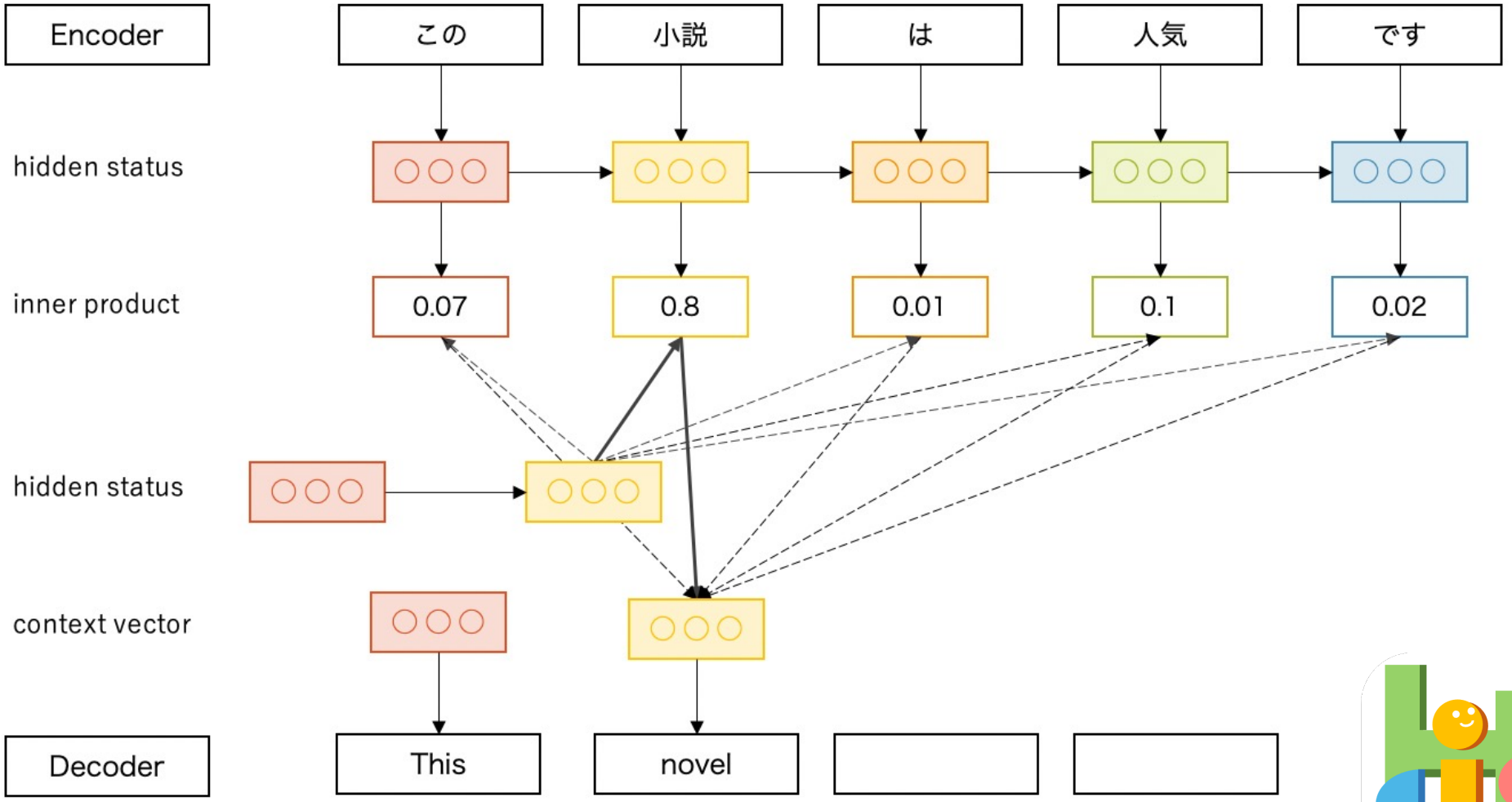


NLPにおけるAttentionとは

- **系列中の重要な情報（文中のある単語の意味を理解するために文中の単語のどれに注目すれば良いか）を直接的に用いる仕組み**
  - **RNNs**の系列の**位置情報**を捉える利点
  - **CNN**の**並列化**しやすいという利点

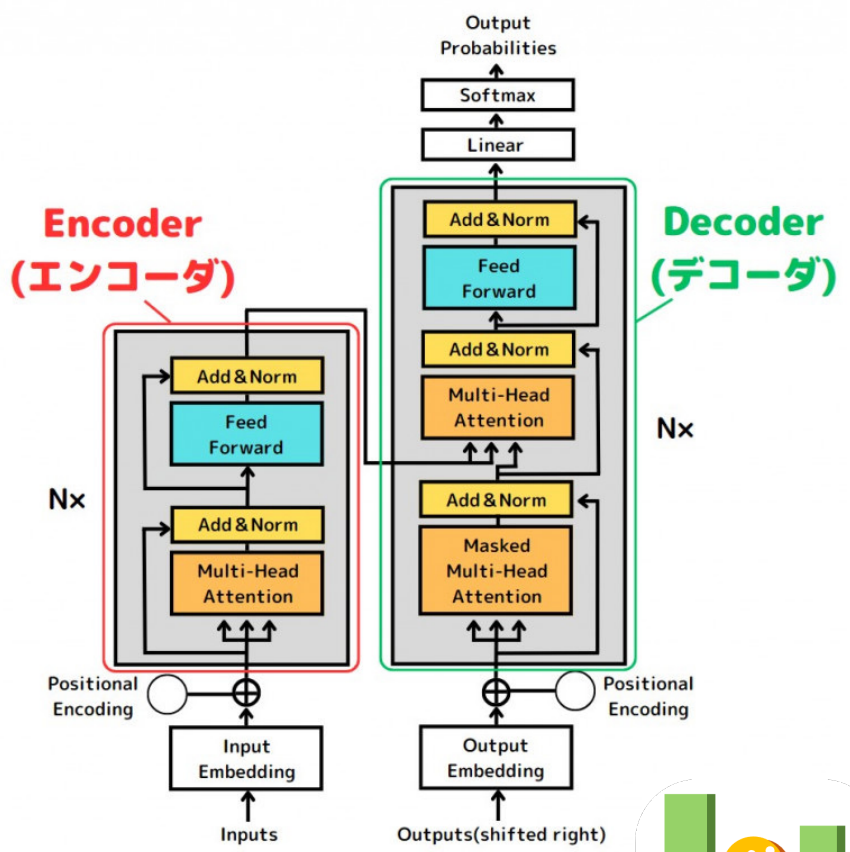
を兼ね備えている

簡単な実装(Soft Attention)



Transformer

- 数多くのNLPタスクでSoTAを達成したネットワークアーキテクチャ
- Attention構造を中心としたエンコーダデコーダモデル



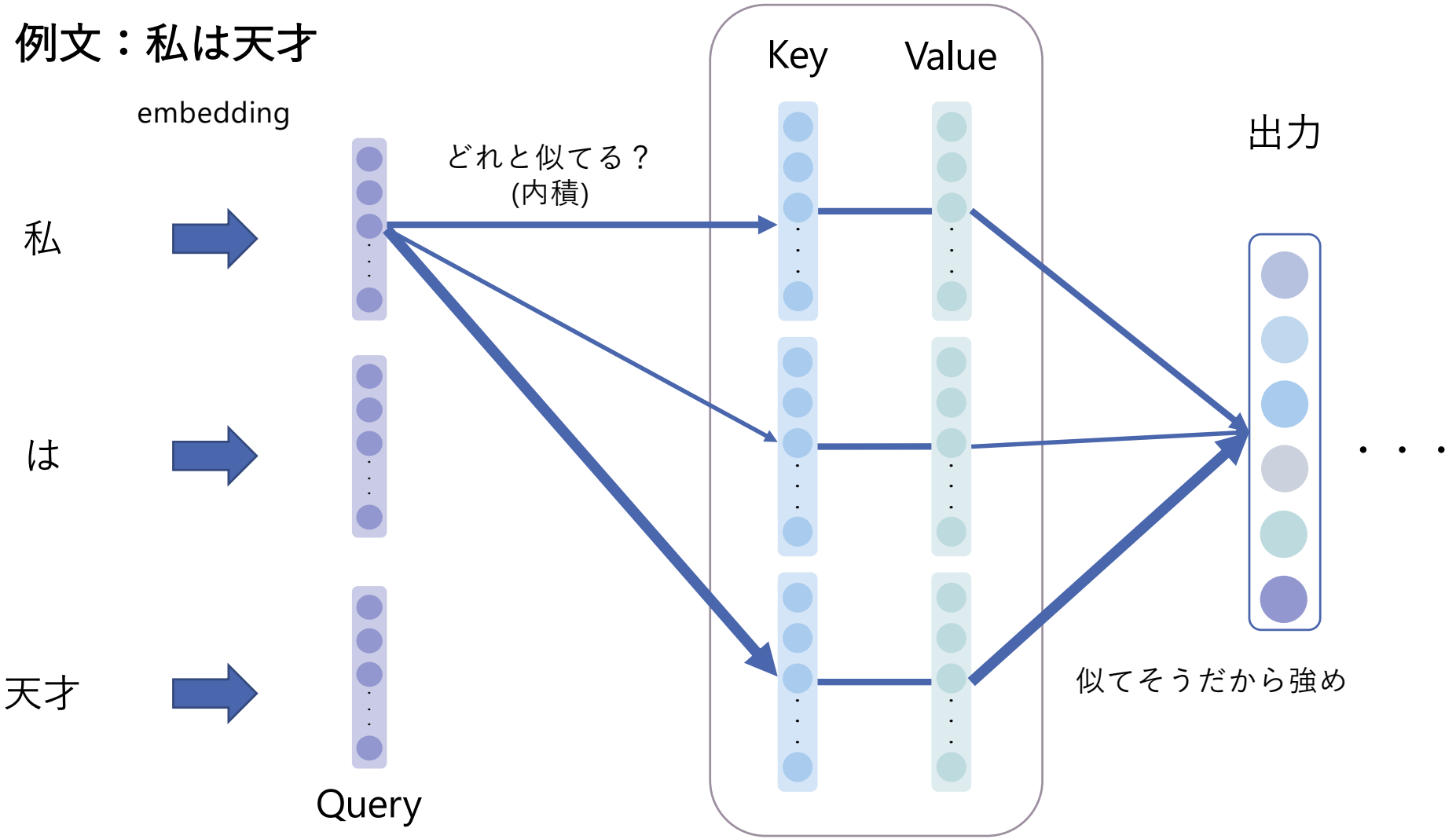
Scaled Dot Product Attention

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

- QueryとKeyとValueから構成
- 入力された検索キーワード(Query)にもっとも近しいキー(Key)を選び，対応するバリュー(Value)を得る機構

Scaled Dot-Product Attentionの雰囲気

例文：私は天才



Source-Target-Attention

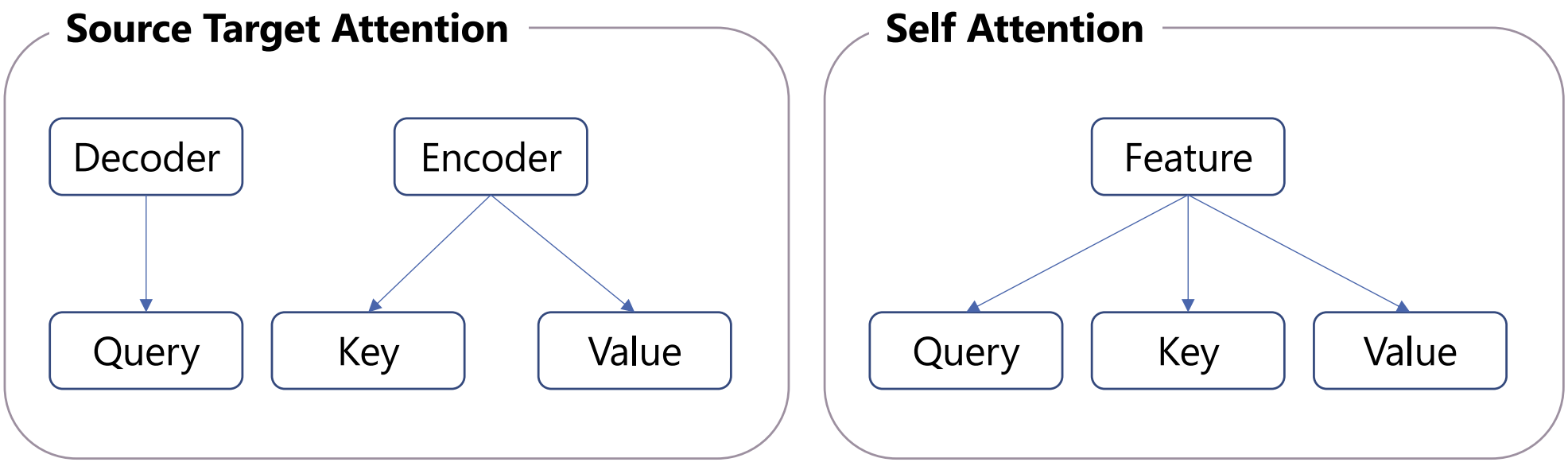
AttentionはDecoderの隠れ層であるQueryによって，Encoderの隠れ層であるmemory（KeyとValueのペア）から重要な情報を取り出す機構とみなすことができ，次のように表せる．

$$Attention(Query, Key, Value) = SOFTMAX(Query, Key^T) \cdot Value$$

QueryとKeyの関連度をsoftmaxで正規化してAttention weightを計算し，Keyの位置に対応したValueを加重和として取り出す．

Self Attention

Source Target Attentionでは，QueryがDecoderの隠れ層でKeyとValueのペアがEncoderの隠れ層を表していたが，Self Attentionでは，QueryもKeyもValueも同じ特徴量から生成される．



Multi-head Attention

各単語に対し１組のQuery, Key, Valueを割り当てるのではなく，

複数のheadという単位に分けてQuery, Key, Valueの組を用意.

各headで潜在表現を計算し，最終的にheadの潜在表現を並列に結合することで，

様々な側面から見た各単語の潜在表現を得る.

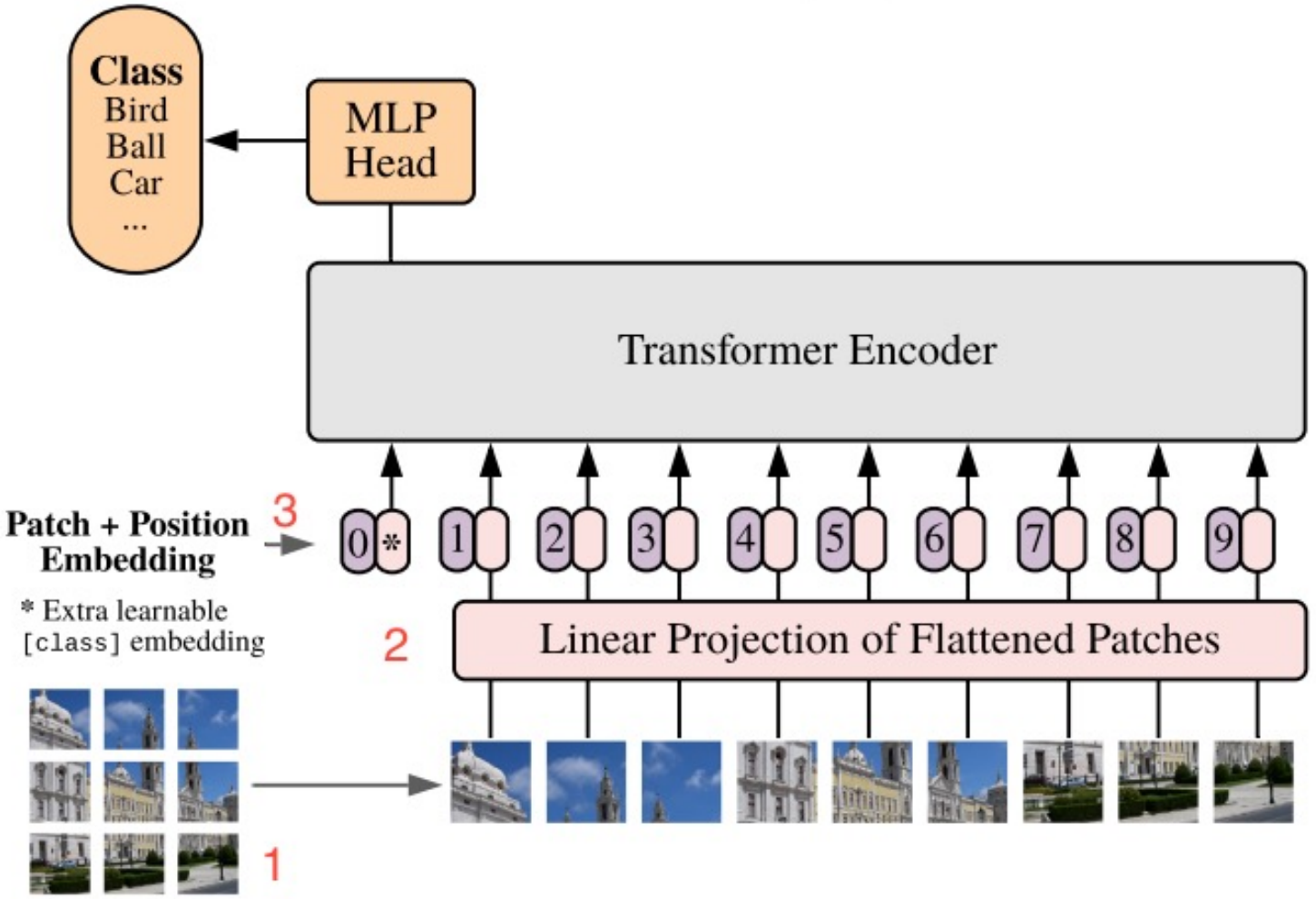
学習時において各ヘッドの**統合には全結合層**が適用されるのが一般的だが，

Attention Mapの**可視化の際にはヘッド間のAttention Scoreの平均**を使う

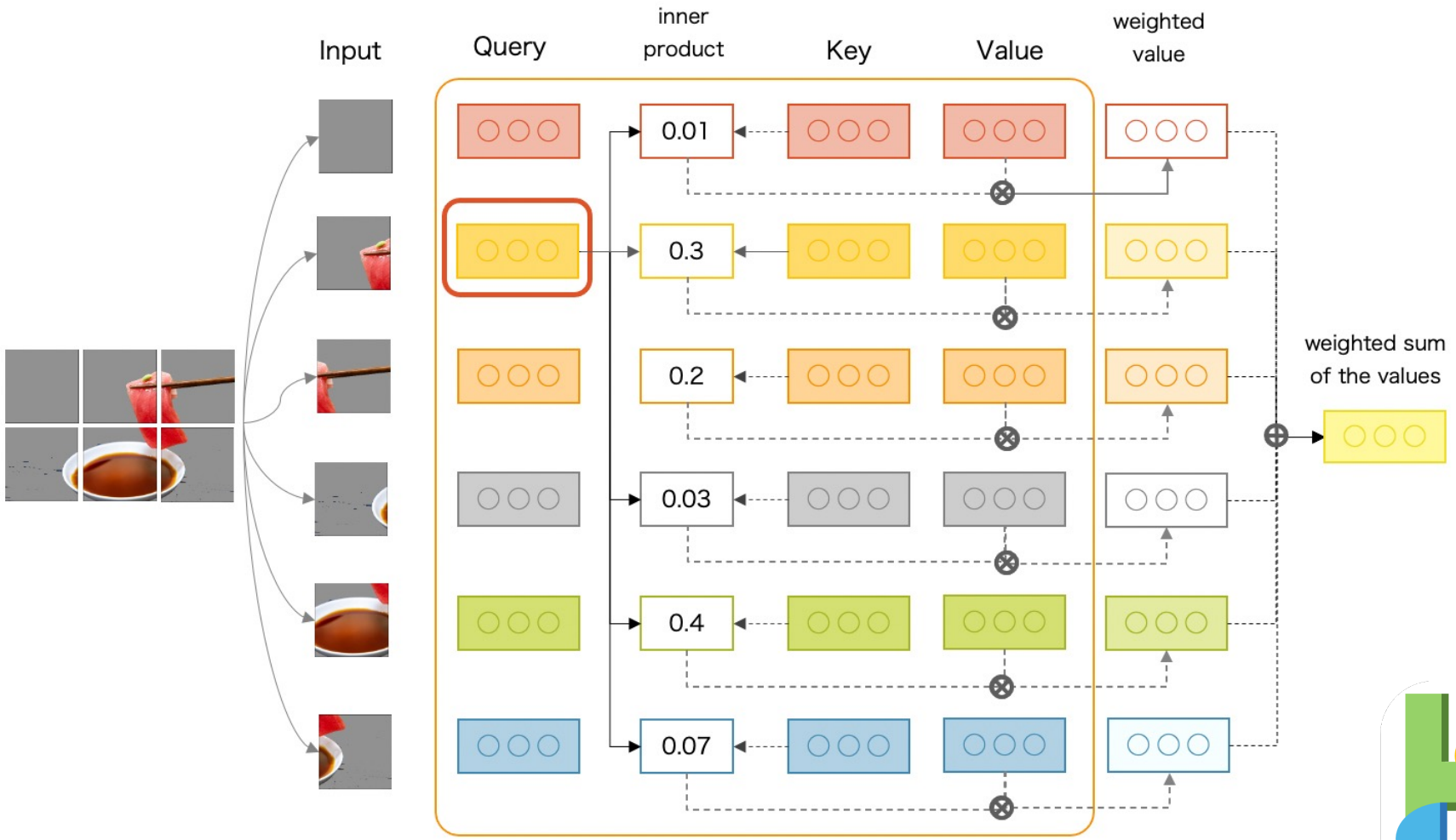
ViTとは？

- **畳み込みと完全にさようならしたSoTAを達成したモデル**
- **ViT(Vision Transformer)の大きな特徴**
  - 画像パッチを単語のように扱う
  - アーキテクチャはTransformerのエンコーダー部分
- **SoTAを上回る性能を約1/15の計算コストで実現**

ViTとは？

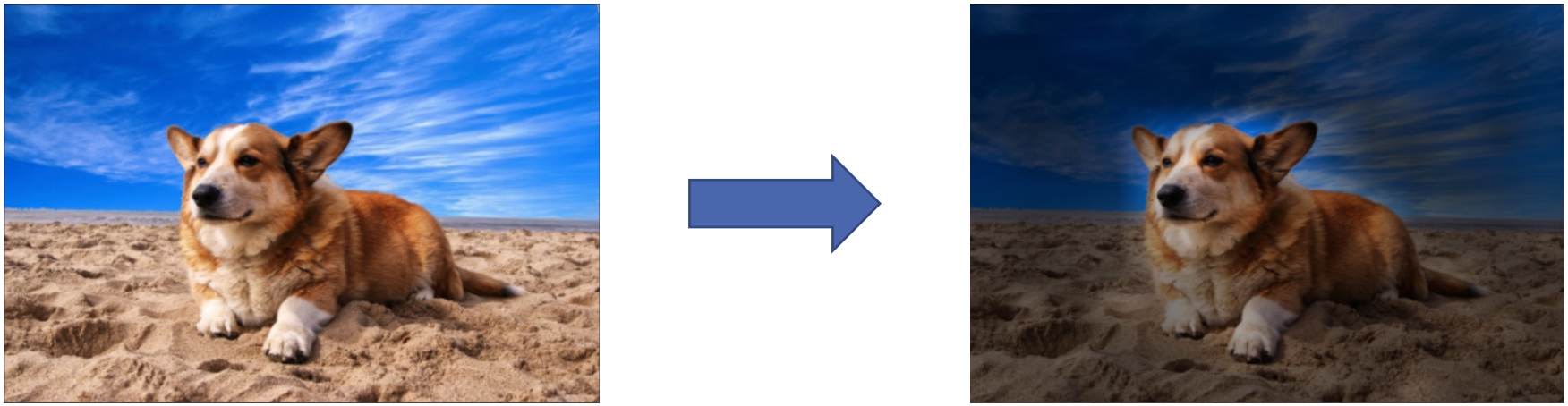

Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
3
* Extra learnable [class] embedding
2
Linear Projection of Flattened Patches
1
0
*
1
2
3
4
5
6
7
8
9

ViTにおけるAttention



CVにおける注目領域の可視化

QueryとKeyの類似度の高さを参考に元画像にヒートマップを重ねてみると...



実際のPyTorchとTensorflowでの実装を見てみよう

ごちゃごちゃ説明しましたが,

結局，コード見て理解するのが一番手っ取り早いので

アテンション機構の実装のレポジトリを用意しました.

https://github.com/tech-tatsuma/AttentionMechanism

まとめ

- Attention機構は入力データの注目すべきものに重みをつけて学習を行なっていく手法
- Transformerに使われているAttention手法以外にもあらゆるAttentionが提案されてきた
- Attention Scoreを可視化することで推論根拠の可視化もできるようになる

参考文献

・[深層学習]図で理解するAttention機構

・transformer-GitHub

・画像認識の大革命．AI界で話題爆発中の「Vision Transformer」を解説