人工知能
IE229 – ARTIFICIAL INTELLIGENCE

第14回講義: 自然言語処理 (3)
Lecture 14: Natural Language Processing (3)

二宮 崇 (Takashi Ninomiya)
愛媛大学 (Ehime University)
ninomiya@cs.ehime-u.ac.jp

今回の講義内容

- **自然言語処理 (Natural Language Processing)**
  - ニューラル機械翻訳 (Neural Machine Translation)
    - アテンション付きエンコーダー・デコーダーモデル
    - Transformer

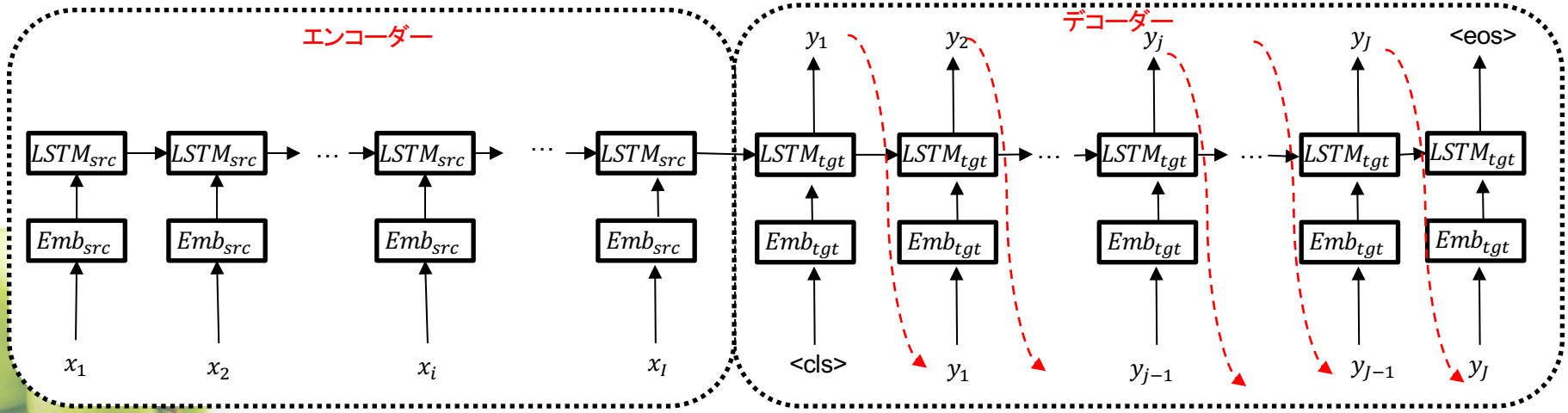
アテンション付きエンコーダー・デコーダーモデル (ENCODER-DECODER MODEL WITH ATTENTION)

LSTMに基づく エンコーダー・デコーダーモデル

- **LSTMに基づくエンコーダー・デコーダーモデル**
  - 入力文: $x_1, x_2, \cdots, x_i, \cdots, x_I$、 出力文: $y_1, y_2, \cdots, y_j, \cdots, y_J$
  - エンコーダー: $\boldsymbol{h}_i = LSTM_{src}(Emb_{src}(x_i), \boldsymbol{h}_{i-1})$
  - デコーダー: $\boldsymbol{h}_j = LSTM_{tgt}(Emb_{tgt}(y_{j-1}), \boldsymbol{h}_{j-1})$

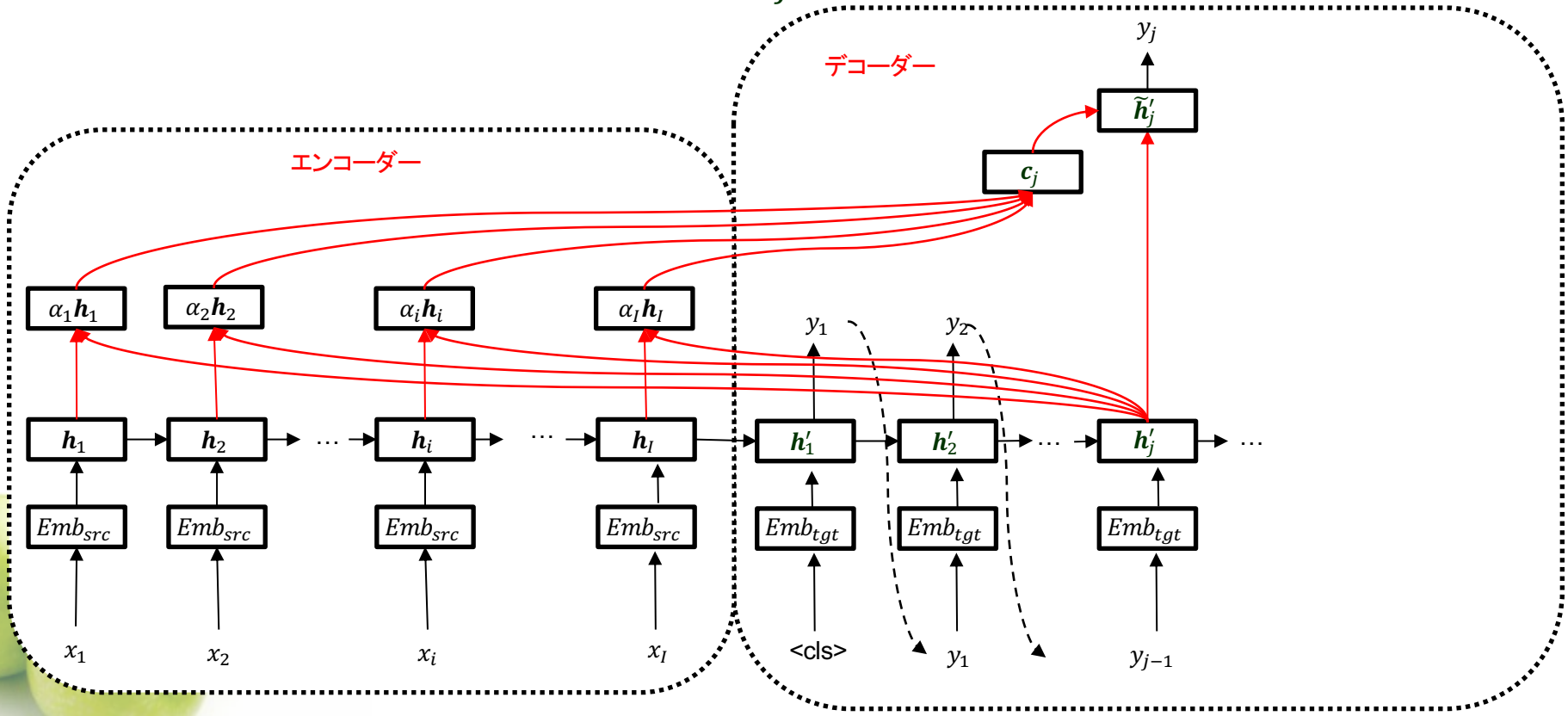
$$y_j = softmax(W_o \boldsymbol{h}_j + \boldsymbol{b}_o)$$

- **しかし、LSTMを使っても、エンコーダーとデコーダーの間でかなりの距離があるので情報がうまく伝播されない**



アテンション付き エンコーダー・デコーダーモデル

- デコーダーの出力の際にエンコーダーの内部状態を直接参照することで長距離の問題を解決する
- デコーダーの内部状態($\boldsymbol{h}'_j$)に対するエンコーダーの内部状態($\boldsymbol{h}_i$)の重み (α_i)を求め、 $\boldsymbol{h}_i$の加重平均($\boldsymbol{c}_j$; コンテキストベクトル)を求める



アテンション付き エンコーダー・デコーダーモデル

- **アテンション付きエンコーダー・デコーダーモデルの形式化**
  - 入力文: $x_1, \cdots, x_i, \cdots, x_I$
  - 出力文: $y_1, \cdots, y_j, \cdots, y_J$
  - エンコーダー: $\boldsymbol{h}_i = LSTM_{src}(Emb_{src}(x_i), \boldsymbol{h}_{i-1})$ $(\boldsymbol{h}_i \in \mathbb{R}^d)$
  - デコーダー: $\boldsymbol{h}'_j = LSTM_{tgt}(Emb_{tgt}(y_{j-1}), \boldsymbol{h}'_{j-1})$ $(\boldsymbol{h}'_j \in \mathbb{R}^d)$

アテンションの重み $\alpha_i = \dfrac{\exp(\boldsymbol{h}_i^T \boldsymbol{h}'_j)}{\sum_{i'=1}^{I} \exp(\boldsymbol{h}_{i'}^T \boldsymbol{h}'_j)}$ $(\alpha_i \in \mathbb{R})$

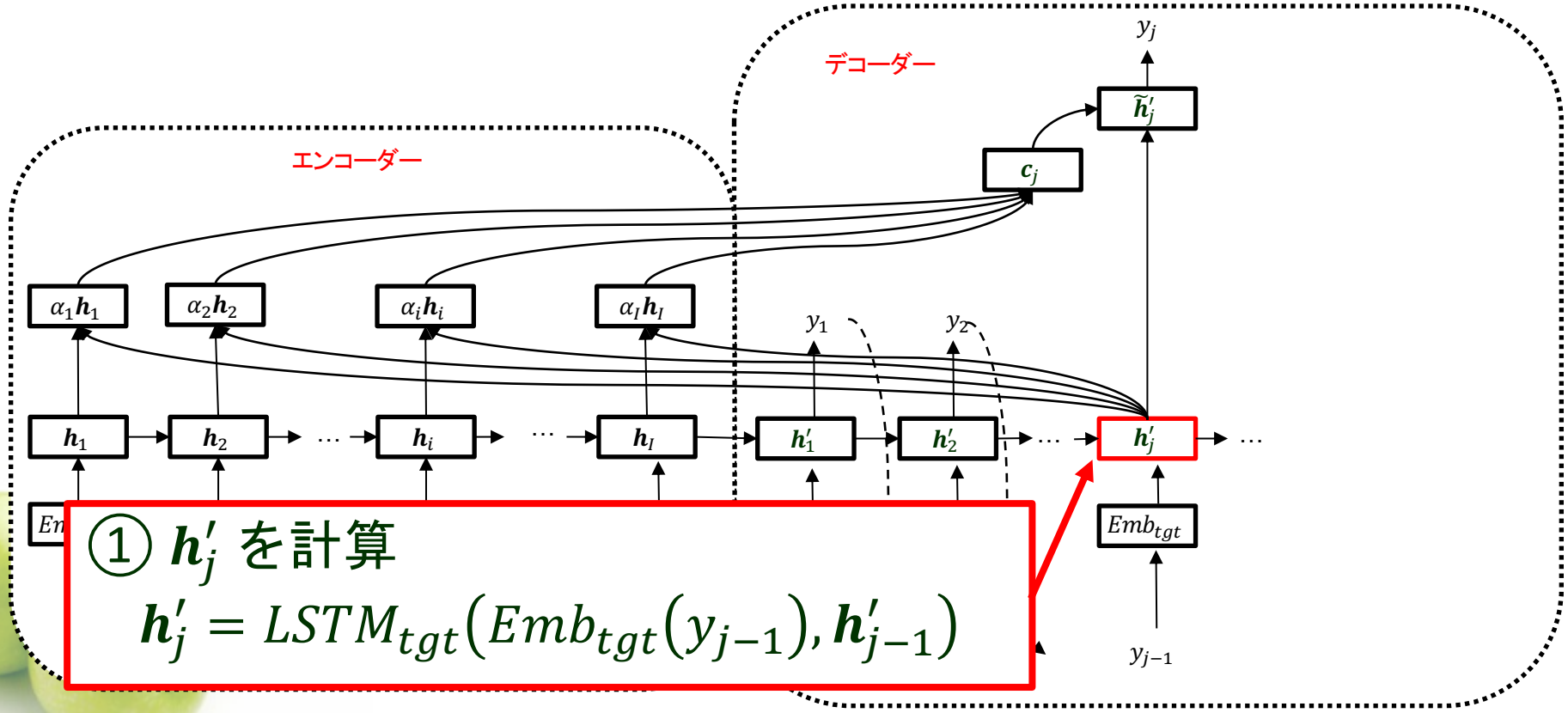
コンテキストベクトル $\boldsymbol{c}_j = \sum_{i=1}^{I} \alpha_i \boldsymbol{h}_i$ $(\boldsymbol{c}_i \in \mathbb{R}^d)$

$\widetilde{\boldsymbol{h}}'_j = \tanh(W_c[\boldsymbol{c}_j; \boldsymbol{h}'_j] + \boldsymbol{b}_c)$ $(W_c \in \mathbb{R}^{d \times 2d})$

$y_j = softmax(W_o \widetilde{\boldsymbol{h}}'_j + \boldsymbol{b}_o)$ $(W_o \in \mathbb{R}^{V \times d})$

;は連結、dは内部状態の次元数、Vは語彙数

アテンション付き
エンコーダー・デコーダーモデル

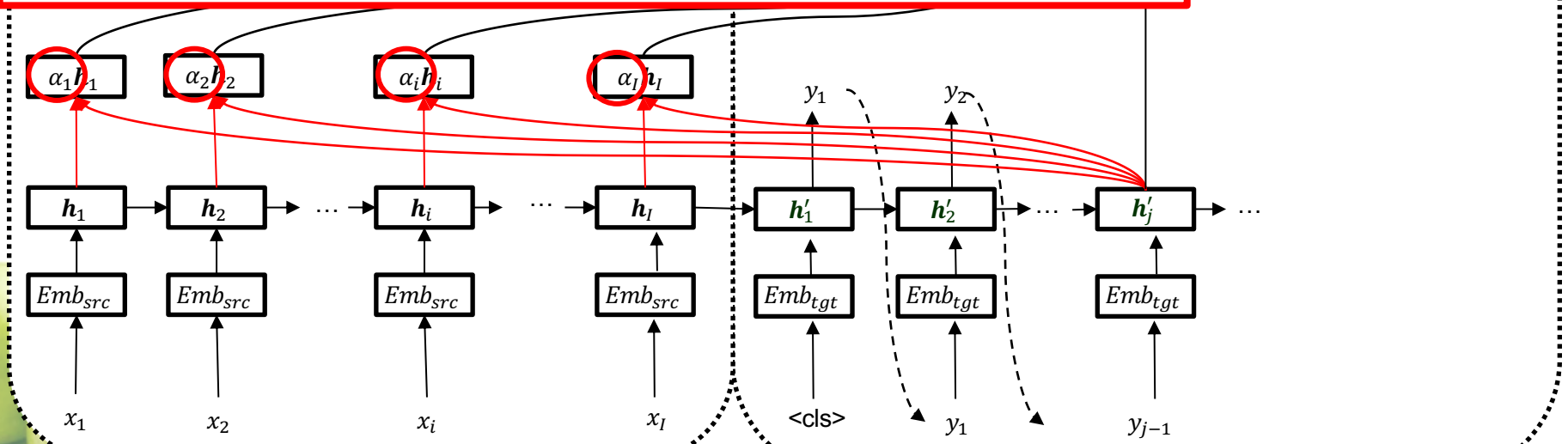


アテンション付き エンコーダー・デコーダーモデル

② $\boldsymbol{h}_i$と$\boldsymbol{h}'_j$からα_iを計算$(1 \leq i \leq I)$

$$\alpha_i = \frac{\exp\left(\boldsymbol{h}_i^T \boldsymbol{h}'_j\right)}{\sum_{i'=1}^{I} \exp\left(\boldsymbol{h}_{i'}^T \boldsymbol{h}'_j\right)}$$

- α_iは$\boldsymbol{h}_i$と$\boldsymbol{h}'_j$の内積にsoftmaxをかけたもの
- $0 \leq \alpha_i \leq 1, \sum_{i=1}^{I} \alpha_i = 1$となることに注意

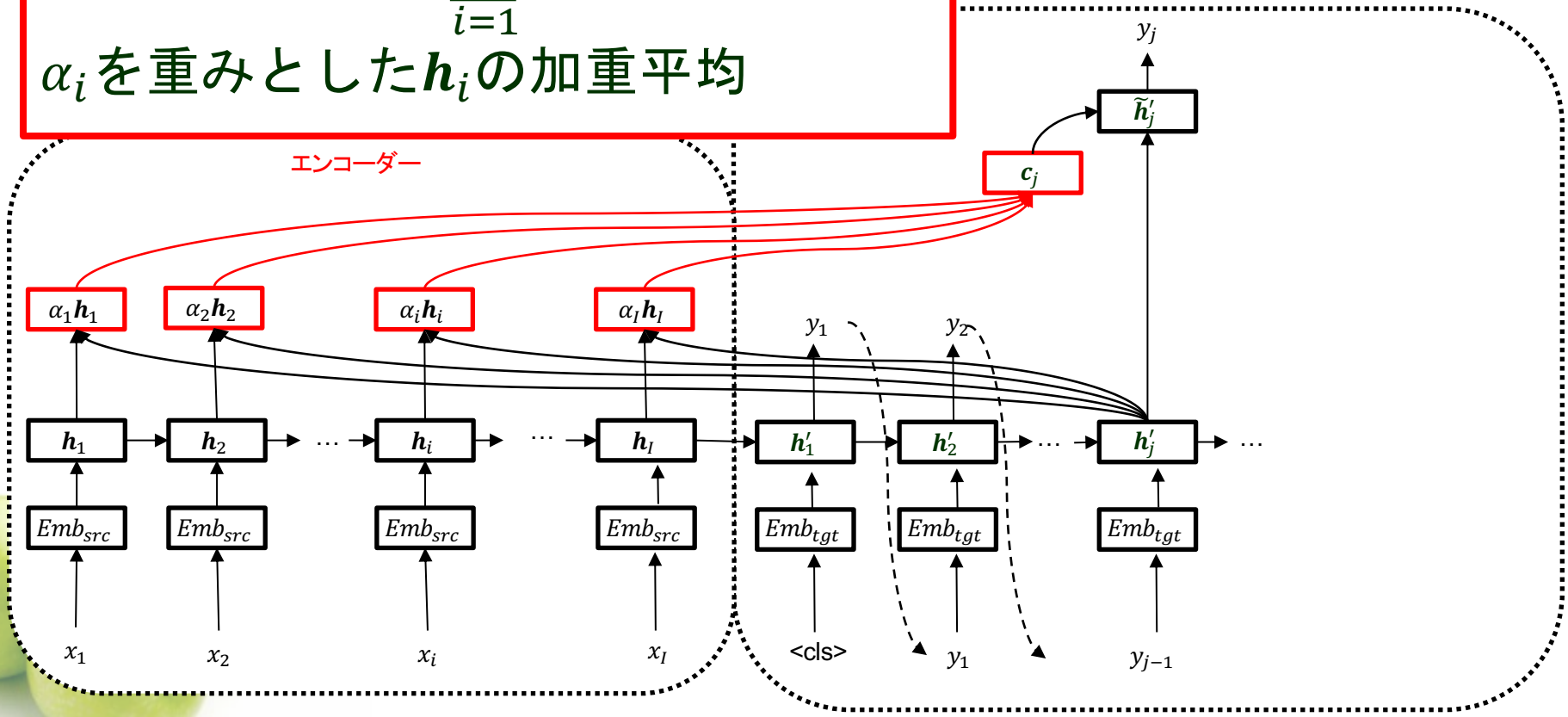


アテンション付き エンコーダー・デコーダーモデル

③ コンテキストベクトルc_jを計算

$$c_j = \sum_{i=1}^{I} \alpha_i h_i$$

α_iを重みとしたh_iの加重平均

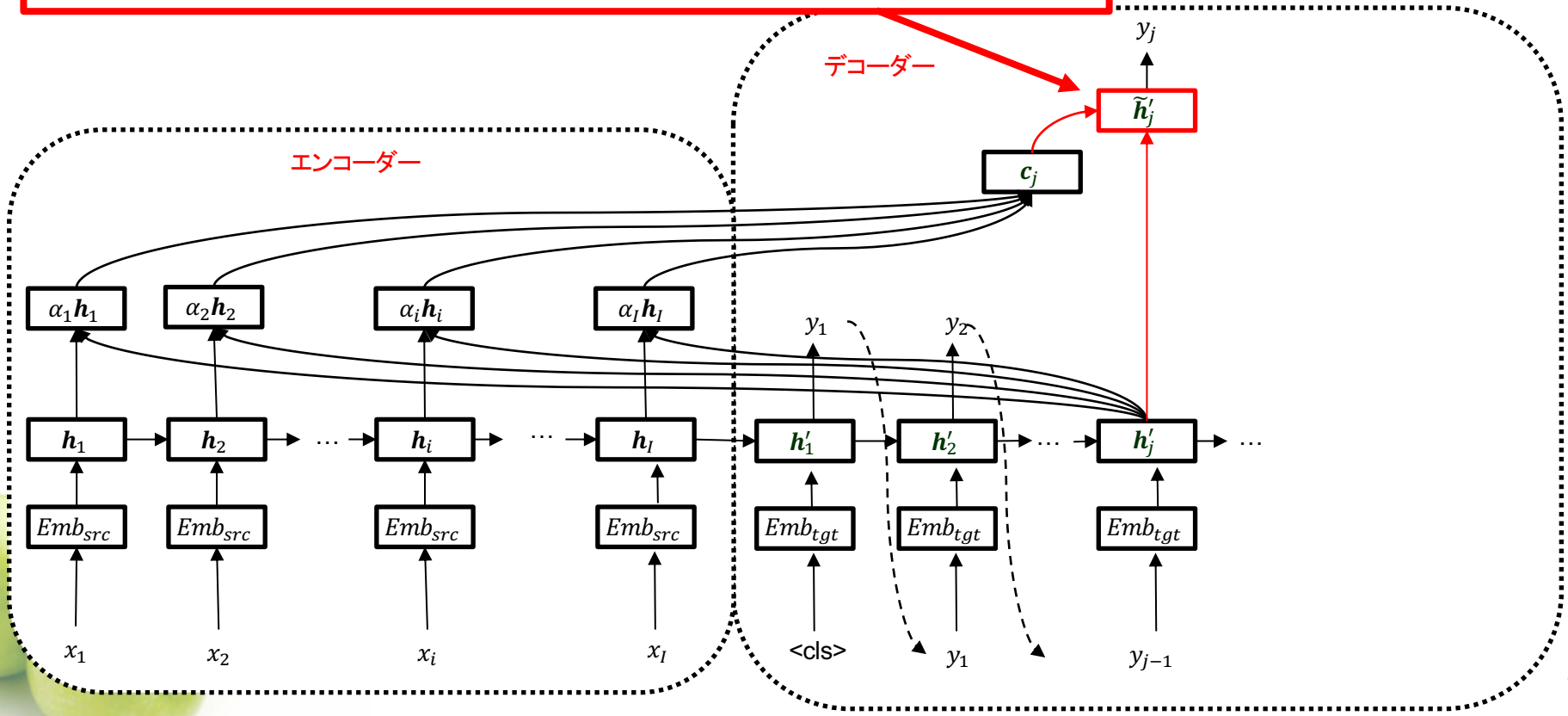


アテンション付き エンコーダー・デコーダーモデル

④新しい内部状態$\tilde{\boldsymbol{h}}'_j$を計算

$$\tilde{\boldsymbol{h}}'_j = \tanh\left(W_c\left[\boldsymbol{c}_j; \boldsymbol{h}'_j\right] + \boldsymbol{b}_c\right)$$

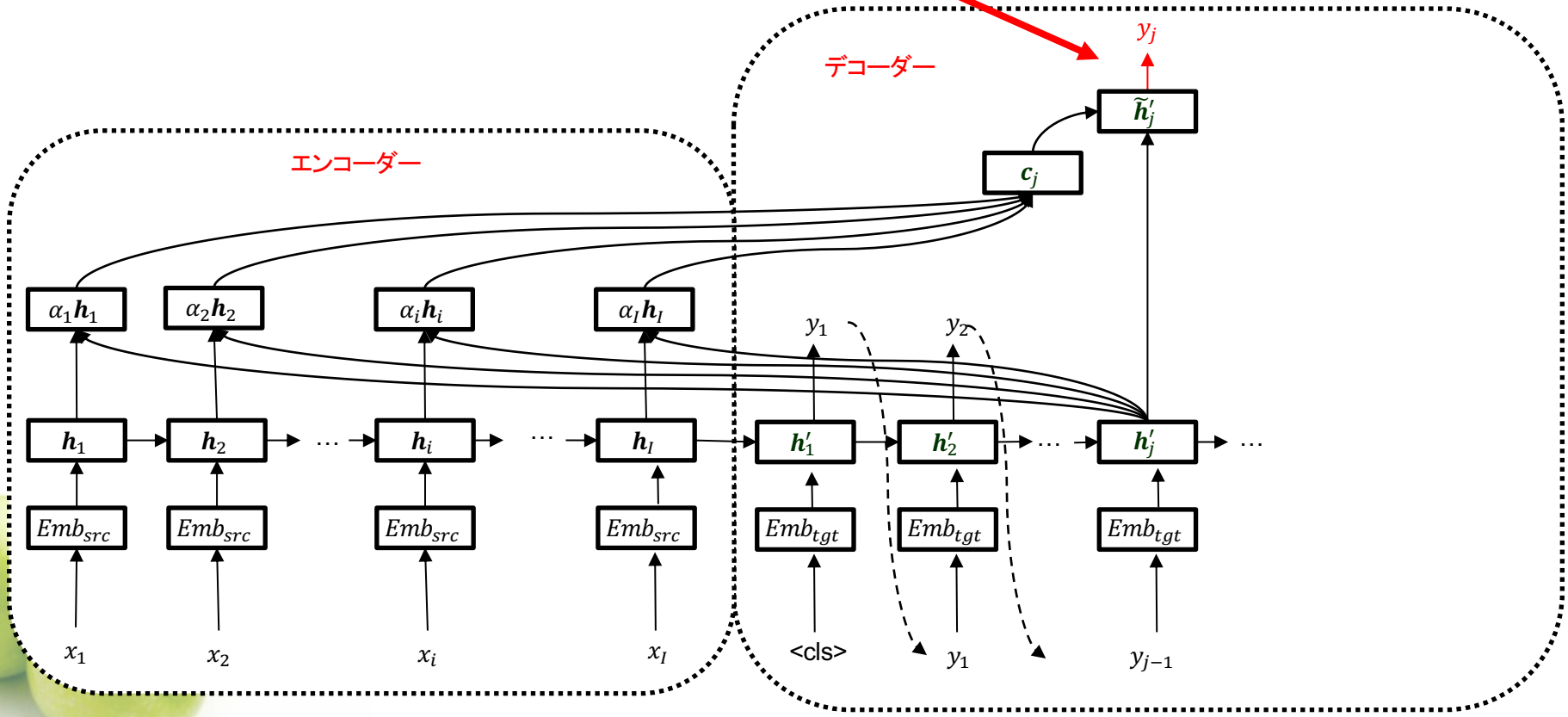
$\left[\boldsymbol{c}_j; \boldsymbol{h}'_j\right]$は$\boldsymbol{c}_j$と$\boldsymbol{h}'_j$を連結したベクトル



アテンション付き エンコーダー・デコーダーモデル

⑤出力を計算

$$y_j = softmax(W_o \tilde{\boldsymbol{h}}'_j + \boldsymbol{b}_o)$$



アテンション付き エンコーダー・デコーダーモデル

- **まとめ**
  - エンコーダーの内部状態に対する重み付け

$$\alpha_i = \frac{\exp\left(\boldsymbol{h}_i^T \boldsymbol{h}'_j\right)}{\sum_{i'=1}^{I} \exp\left(\boldsymbol{h}_{i'}^T \boldsymbol{h}'_j\right)}$$

    - 基本的に内積で計算されていることに注意
    - $0 \leq \alpha_i \leq 1, \sum_{i=1}^{I} \alpha_i = 1$となることに注意
  - コンテキストベクトルは、エンコーダーの内部状態の加重平均

$$\boldsymbol{c}_j = \sum_{i=1}^{I} \alpha_i \boldsymbol{h}_i$$

  - 新しい内部状態はコンテキストベクトルと内部状態を連結したベクトルに対する線形変換で得られる

$$\tilde{\boldsymbol{h}}'_j = \tanh\left(W_c\left[\boldsymbol{c}_j; \boldsymbol{h}'_j\right] + \boldsymbol{b}_c\right)$$

アテンション付き エンコーダー・デコーダーモデル

- **重み付けのバリエーション**

$$\alpha_i = \frac{\exp\left(score\left(\boldsymbol{h}_i, \boldsymbol{h}'_j\right)\right)}{\sum_{i'=1}^{I} \exp\left(score\left(\boldsymbol{h}_{i'}, \boldsymbol{h}'_j\right)\right)}$$

  - Global Attention (dot)

$$score\left(\boldsymbol{h}_i, \boldsymbol{h}'_j\right) = \boldsymbol{h}_i^T \boldsymbol{h}'_j$$

    - 先程説明したものと同じ通常のアテンション

ベクトル間の関係を内積で表現

  - Global Attention (general)

$$score\left(\boldsymbol{h}_i, \boldsymbol{h}'_j\right) = \boldsymbol{h}_i^T W_a \boldsymbol{h}'_j$$

ベクトル間の関係を重み行列W_aで学習

  - Global Attention (concat)

$$score\left(\boldsymbol{h}_i, \boldsymbol{h}'_j\right) = \boldsymbol{v}^T \tanh\left(W_a\left[\boldsymbol{h}_i; \boldsymbol{h}'_j\right]\right)$$

2つのベクトル間の重み付けを重み行列W_aと重みベクトル$\boldsymbol{v}$で学習

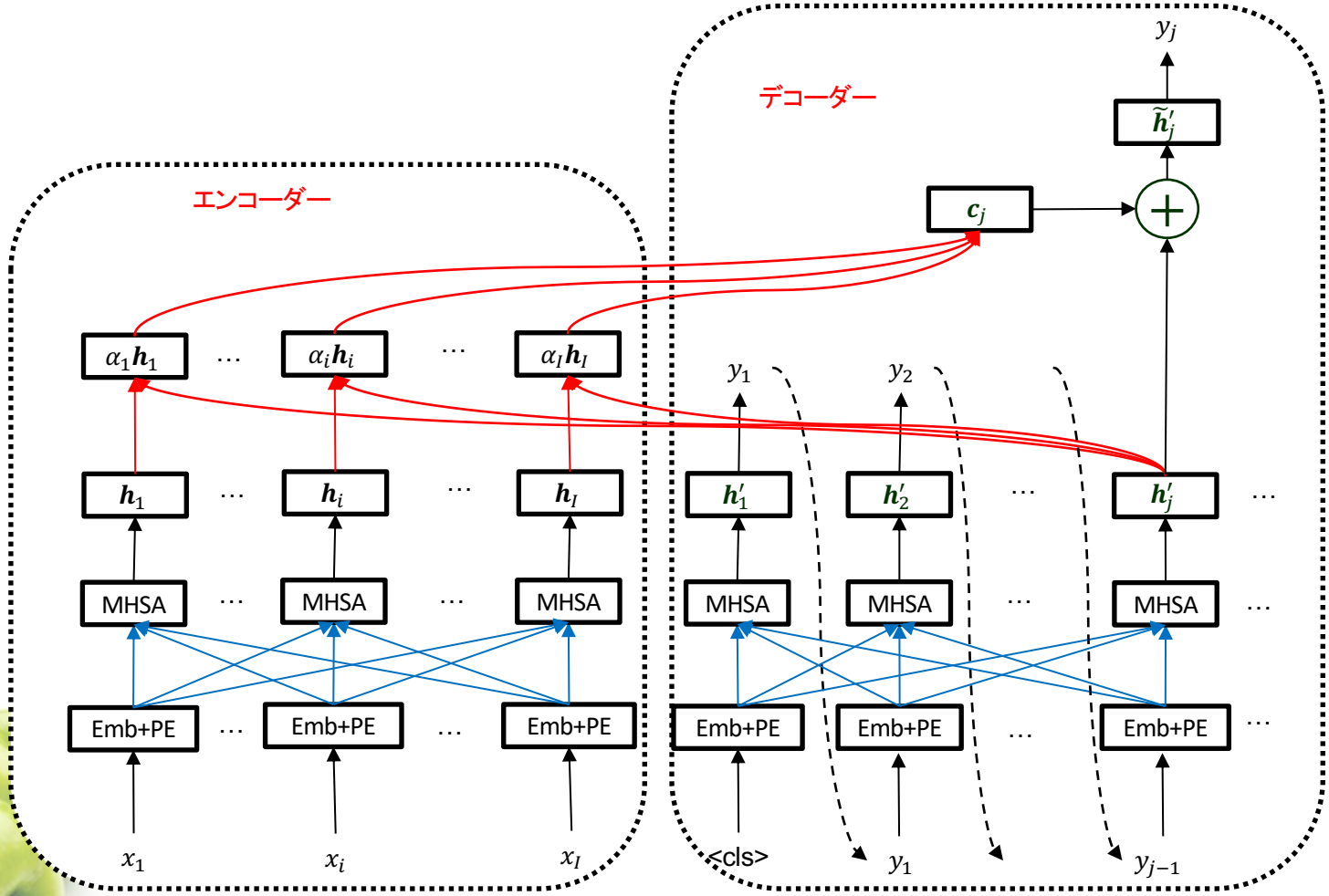
TRANSFORMER

Transformer

- **Transformerモデル**
  - 現在のニューラル機械翻訳のベースラインモデル
  - RNNでもCNNでもない新しいタイプのニューラル機械翻訳モデル
  - アテンションの計算だけで翻訳を行うことが特徴
  - キーテクノロジー
    - 自己アテンション (Self Attention)
    - 位置エンコーディング (Positional Encoding)
    - マルチヘッドアテンション (Multi-head Attention)

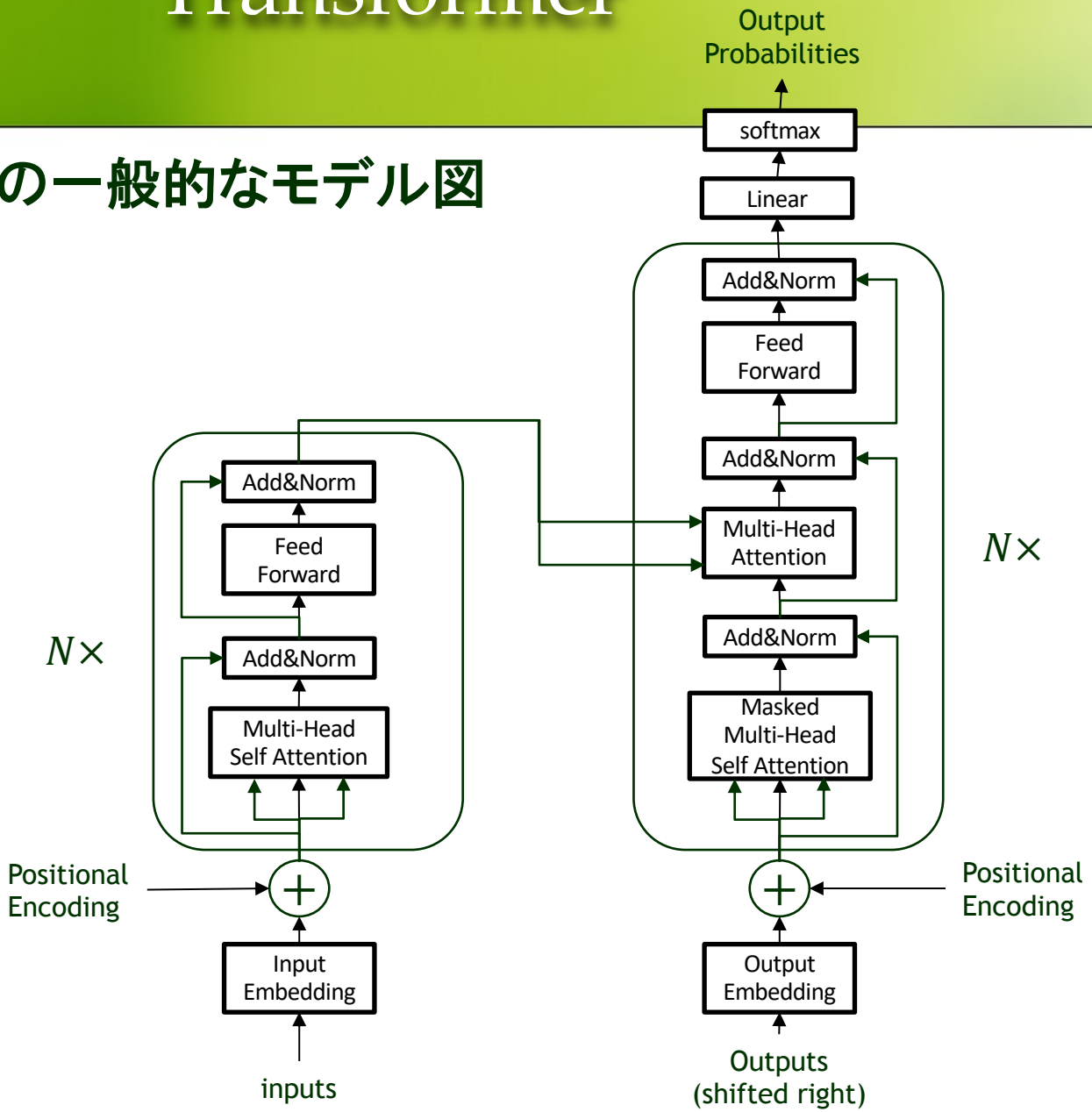
Transformer

- Transformerのモデル概要図



Transformer

- Transformerの一般的なモデル図



Transformer

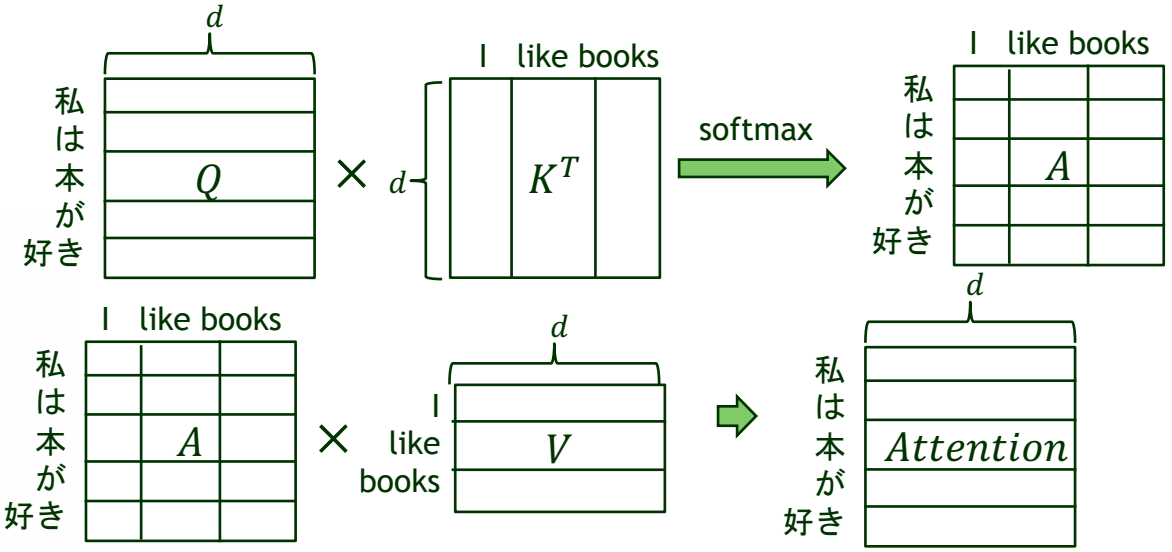
- **アテンション**
  - d: 内部状態の次元数
  - Query, Key, Valueの入力があると考えてアテンションを捉え直す
    - Query(Q): 計算対象の内部状態 (デコーダーの内部状態)
    - Key(K): アテンションの対象 (エンコーダーの内部状態)
    - Value(V): アテンション先の内部状態 (エンコーダーの内部状態)
  - Q、K、Vは(文長×d)の行列を表し、文全体のアテンションをまとめて計算
  - QとKから重み(A)を計算して、重みとVの加重平均を計算。Aは(文長×文長)の行列

$$A = softmax\left(\frac{QK^T}{\sqrt{d}}\right)$$

$$Attention(Q, K, V) = AV$$

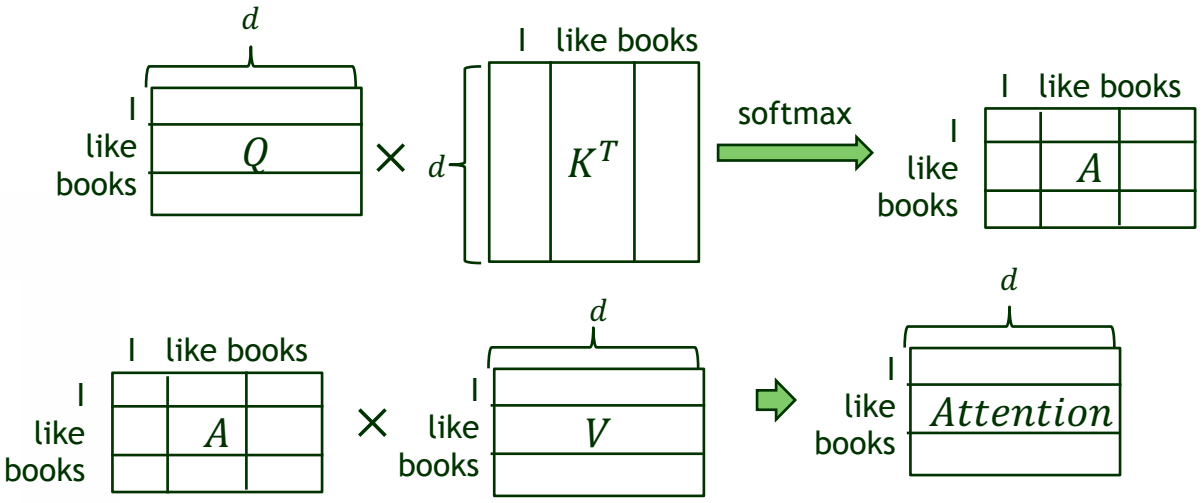
Transformer: アテンションの例

- **言語間アテンション(Encoder-Decoder Attention, Cross-Lingual Attention)**
  - エンコーダー(“I like books”)→デコーダー(“私 は 本 が 好き”)
    - Q: デコーダーの内部状態
    - K: エンコーダーの内部状態
    - V: エンコーダーの内部状態(KとVは同じ)
    - A: QK^Tを計算し、各行ごとにsoftmaxの計算



Transformer: 自己アテンション

- **自己アテンション(Self Attention)**
  - エンコーダーの場合(“I like books”)
    - Q: エンコーダーの内部状態
    - K: エンコーダーの内部状態
    - V: エンコーダーの内部状態(QとKとVは同じ)
    - A: QK^Tを計算し、各行ごとにsoftmaxの計算



Transformer

- **マルチヘッドアテンション (Multi-head Attention)**
  - 内部状態を表すベクトル($\boldsymbol{h}$)をヘッド数(H)に分割して、それぞれの分割されたベクトルに対してアテンションの計算を行う
  - d: 内部状態の次元数
  - H: ヘッド数
    - 例: $d = 512, H = 8$なら、各ヘッドの次元数は512/8=64次元)
  - 線形変換による分割($W_h^Q, W_h^K, W_h^V \in \mathbb{R}^{d\times(d/H)}$)

$$Q_h = QW_h^Q, \qquad K_h = KW_h^K, \qquad V_h = VW_h^V$$

  - 各ヘッドのアテンションの計算

$$Head_h = softmax\left(\frac{Q_h K_h^T}{\sqrt{d/H}}\right)V_h$$

  - アテンションの計算($W^o \in \mathbb{R}^{d\times d}$)

$$Attention(Q, K, V) = [Head_1; \cdots; Head_H]W^o$$

Transformer

- **Feed Forward層**
  - 各単語ごとの全結合層(Fully Connected Layer)2層+ReLU

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

Transformer

- **レイヤー正規化 (Layer Normalization)**
  - ベクトル内の全ての次元の値の平均、分散を求めて正規化すること
  - $\boldsymbol{h}$: 内部状態 ($\boldsymbol{h} = (h_1, \cdots, h_d)^T$)
  - d: 内部状態の次元数
  - $\boldsymbol{\gamma}, \boldsymbol{\beta}$: パラメータ

$$\mu = \frac{1}{d}\sum_{i=1}^{d} h_i$$

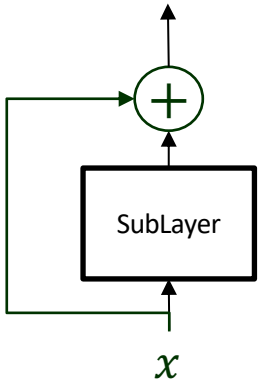
$$\sigma = \sqrt{\frac{1}{d}\sum_{i=1}^{d}(h_i - \mu)^2}$$

$$LayerNorm(\boldsymbol{h}) = \boldsymbol{\gamma} \odot \frac{\boldsymbol{h} - \mu}{\sigma} + \boldsymbol{\beta}$$

Transformer

- **残差接続(Residual Connection)**
  - ニューラルネットワークのある部分ネットワークSubLayerがあったとき、次の計算を行うことを残差接続(Residual Connection)という

$$Res(\boldsymbol{x}) = \boldsymbol{x} + SubLayer(\boldsymbol{x})$$



- **Add&Norm層**

$$AddNorm(\boldsymbol{x}) = LayerNorm\big(\boldsymbol{x} + SubLayer(\boldsymbol{x})\big)$$

Transformer: 位置エンコーディング

- **位置エンコーディング (Positional Encoding)**
  - 自己アテンションで同じ文内の任意の単語間の関係が計算されているが、単語間の距離が考慮されていない(隣の単語も遠くの単語も同じ様にアテンションの計算が行われる)
  - 単語埋め込みの後に、位置情報を埋め込むことでこの問題を解決する

  - 位置エンコーディングの式
    - pos: 文中の位置
    - $2i, 2i+1$: 位置埋め込みベクトルの次元
    - d: 埋め込みベクトルの次元数

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right)$$

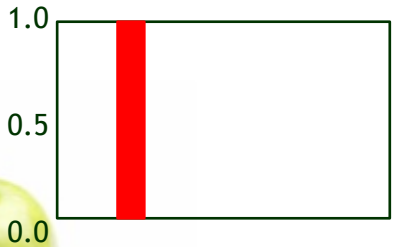
Transformer
Dropoutとラベルスムージング

- **各層の後にDropoutを適用**
- **出力にラベルスムージング(Label Smoothing)を適用**
  - 通常の交差エントロピー損失

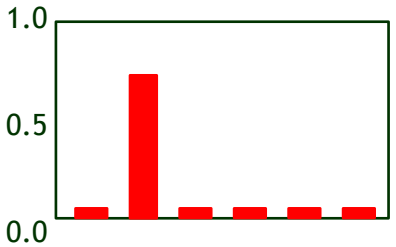
$$L = -\sum_{j=1}^{J}\sum_{k=1}^{K} t_{jk} \log P\left(y_{jk} \middle| x, y_{<j}\right)$$

  - ラベルスムージング

$$L = -\sum_{j=1}^{J}\sum_{k=1}^{K} \left[(1-\epsilon)t_{jk} + \epsilon\frac{1}{K}\right] \log P\left(y_{jk} \middle| x, y_{<j}\right)$$



通常の交差エントロピー



ラベルスムージング

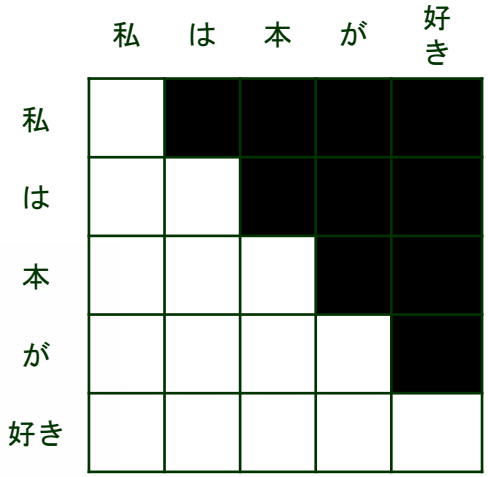
$t_{jk} \in \{0,1\}$はj番目の出力の
k番目のラベルの正解

ϵはスムージングの度合いを
調整するハイパーパラメータ

Transformer: デコーダー

- **マスク付きマルチヘッドアテンション**
  - デコーダーの自己アテンションの計算時にマスクをかける
    - 推論時には逐次的に単語が生成されるため、学習時にもマスクをかけて後ろの単語にアテンションがかからないようにする

デコーダーのアテンション行列A



黒い箇所を自己アテンションの計算対象にしない
(アテンションスコアを0にする)

ニューラル機械翻訳のまとめ

- **アテンション付きエンコーダー・デコーダーモデル**
- **Transformer**
  - 自己アテンション
  - マルチヘッドアテンション
  - レイヤー正規化
  - 残差接続
  - 位置エンコーディング
  - マスク付き自己アテンション