注意機構 (attention mechanism)

概要

- 入力情報の中で特に注目すべき箇所を指定するための機能
  - 画像処理、文字列処理 etc で使用されている
- メリット：
  - 入力の系列情報の冒頭部分を伝播でき、モデルの性能がよくなる
  - (特に画像処理で？) 特定の箇所に着目するので計算コストを抑えれる

種類

- ソフト注意機構
  - 入力情報の重み付き平均を用いる方法
- ハード注意機構
  - 入力情報のどれか一つを確率的に選択して用いる方法
- 自己注意機構（self attention）
  - Transformer 等で使用されている機構

ソフト注意機構

- 系列変換モデルを考える
- 入力系列 $\{x_1, \ldots, x_I\}$、符号化されたベクトル $\{h_1^{(s)}, ..h_I^{(s)}\}$ として、各時刻の符号化層の隠れ状態ベクトルは

$$h_i^{(s)} = \Psi^{(s)}\left(x_i, h_{i-1}^{(s)}\right)$$

- 入力初期の情報（ex. 文頭の情報）は復号化器に伝播するには $\Psi^{(s)}$ が I 回適用されて尚有用な情報として残っておく必要がある
  - もう少し直接的に復号化器に伝播する方法はないか？

- a_{ij} による重み付き平均を考える
  - なぜ j を添えているか？ Issue #15

$$\bar{h}_j = \sum_{i=1}^{I} a_{ij} h_i^{(s)}$$

- 復号化器がj番目の単語の予測を行う際に$\bar{h}_j$を利用する
  - （旧版では連結がめちゃめちゃなので注意 Issue #5）

$$\hat{h}_j^{(s)} = \tanh\left(W^{(a)} \begin{bmatrix} \bar{h} \\ h_j^{(t)} \end{bmatrix}\right)$$

- 最初に与えた $\{a_1, \ldots, a_I\}$
  - ニューラルネットで計算する
- 関数 Ω で $h_i^{(s)}$ と $h_j^{(t)}$ の重みを計算する
  - $e_i = \Omega\left(h_i^{(s)}, h_j^{(t)}\right)$
  - softmax で 1 に規格化して確率化する
- 関数 Ω は複雑な関数、というわけでもなく（定義は自由）
  - ...

ソフト注意機構 (一般化した定義)

- 復号化器の隠れ状態を $h_j^{(t)}$ とし、
  参照したい符号化器の隠れ状態を $Y = \{y_1, \ldots, y_N\}$とする
- $h_j^{(t)}$ に対してどのYが重要化を $\{a_1, \ldots, a_N\}$で表す
  - この重要度を計算するための情報を c_i とする

$$a_i = \frac{\exp(\Omega(c_i))}{\sum_{k=1}^{N} \exp(\Omega(c_k))}$$

- 復号化器からの出力 $h_j^{(t)}$ と $\hat{y} = \sum_{i=1}^{N} a_i y_i$ を用いて最終的な情報を決定する

ハード注意機構

- N個の参照したい情報Yの重み（=確率）をa_iとして、この確率に従ってYの値をただひとつに決定する
- 目的関数 $f(\hat{y})$ の最小化
  - 直接これを最小化できないので、期待値を最小化する

$$\nabla E[f(\hat{y})] = \nabla^s um_{x=1}^N f(y_x)a_x$$

- いずれの項もxの取りうる全ての範囲$1, \ldots, N$に対して計算が必要

§4.2 記憶ネットワーク

- LSTMを始めとするRNNで、文の状態を記憶することができていた
  - しかし記憶の内容（隠れ状態ベクトル）は固定長で限定的だった
- より直接的に記憶の仕組みをモデル化する研究が行われている→記憶ネットワーク

モデル

- 入力情報変換
- 一般化
- 出力情報変換
- 応答

教師あり記憶ネットワーク

end-to-end 記憶ネットワーク

動的記憶ネットワーク

- Dynamic memory networks (DMN)
  - 入力
  - 意味記憶
  - 質問
  - エピソード記憶
  - 回答

§4.3 出力層の高速化