論文読みまとめ
Effective Approaches to Attention-based Neural Machine Translation

井上智裕

2020 年 12 月 22 日

1 導入

1.1 扱う問題

英語-ドイツ語の双方向の翻訳タスク

1.2 問題意識

どこに問題意識を感じているのか？既存手法では何が足りないのか？

従来よりニューラル機械翻訳（NMT）という手法が使われてきた。NMT とは、原文 $x_1, ..., x_n$ を $y_1, ..., y_m$ に翻訳する際の条件付き確率を $p(y|x)$ 直接モデル化したニューラルネットワークのことである。NMT は各原文の表現 s を計算するエンコーダと、一度に 1 つの対象語を生成し、次のように条件付確率 $p(y|x)$ を分解するデコーダからなる。

$$\log p(y|x) = \sum_{j=1}^{m} \log p(y_j|y_{<j}, s) \tag{1.1}$$

すなわち、原文の表現 s とそれまでに生成された途中までの系列が得られたときの、ある系列が得られる条件付確率の積に分解する。この分解は原文の表現 s と生成された系列から確率を計算するタスクなので、RNN アーキテクチャが用いられる。デコーダには RNN、LSTM、GRU などが用いられ、エンコーダには CNN、LSTM、GRU などが用いられる。なお、従来手法の多くでは原文の表現 s はデコーダの隠れ状態の初期化のために一度だけ使用される。系列変換モデルでは系列情報を RNN ベースの手法で固定長のベクトルに変換するが、LSTM などを用いても短い系列に比べ長い系列の方がネットワーク内の情報伝播が難しくなるという問題がある。そこで、入力情報を出力時により直接的に利用する注意機構が用いられるようになった。

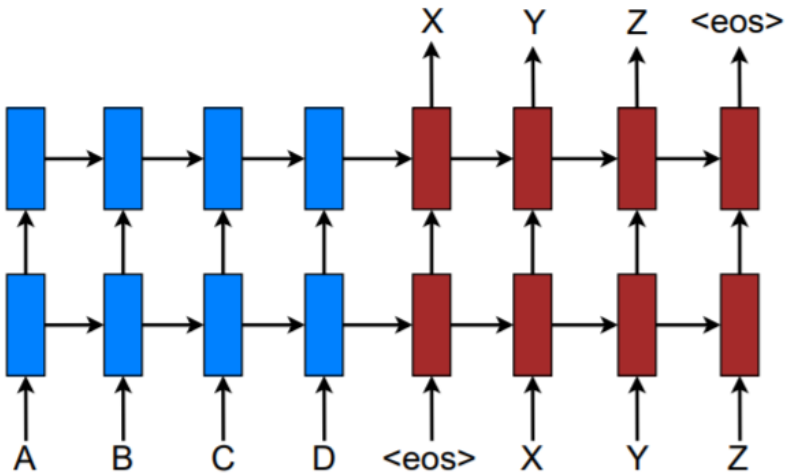

図 1.1　NMT

また、本研究以前の注意機構には、ソフト注意機構やハード注意機構があった。ソフト注意機構は、加重平均の対象が入力系列全体となるため、長い系列ほど計算コストが高くなるというデメリットがある。一方、ハード注意機構は微分できない過程を含むため、最適化が難しくなるという課題がある。本研究では、これらのデメリットに対処する方法として局所注意機構を提案している。

2　理論

どのようなアイデア・ロジック・仮定で問題を解決しようとしているか？なぜそれで問題が解決できるのか？

2.1　Attention-based models

アイデアをどのように定式に落とし込んでいるか？それぞれの式は何を意味しているのか？

本研究の Attention-based models は Global と Local の 2 種類に大別される。2 種類のモデルに共通しているのは、Attentoin layer においてデコーダの各ステップ t においてデコーダの LSTM の最上層の隠れ状態 h_t を入力としていることと、入力系列の情報を捉えるコンテクストベクトル c_t を出力することである。ただし、コンテクストベクトル c_t の計算方法が 2 つのモデルで異なっている。Attention layer で得られたコンテクストベクトル c_t と隠れ状態 h_t の情報は統合されて、以下のように隠れ状態 $\tilde{h}_t$ を生成する。

$$\tilde{h}_t = \tanh(W_c[c_t; h_t]) \tag{2.1}$$

この $\tilde{h}_t$ を用いて、予測される単語の確率分布は次で表される。

$$p(y_t|y_{<t}, x) = softmax(W_s\tilde{h}_t) \tag{2.2}$$

2.2 Global Attention

Global attentional model では、コンテクストベクトル c_t の計算時にエンコーダのすべての隠れ状態が考慮される。このモデルではエンコーダの時間ステップ数に等しいサイズの可変長アライメントベクトル a_t が、着目するデコーダのステップ t の隠れ状態 h_t とエンコーダの各ソースの隠れ状態 $\bar{h}_s$ を比較することによって導出される。

$$a_t(s) = align(h_t, \bar{h}_s) \tag{2.3}$$

$$= \frac{\exp(score(h_t, \bar{h}_s))}{\sum_{s'} \exp(score(h_t, \bar{h}_{s'}))} \tag{2.4}$$

また、score 関数は 3 種類提案されている。

$$score(h_t, \bar{h}_s) = \begin{cases} h_t^T \bar{h}_s & dot \\ h_t^T W_a \bar{h}_s & general \\ v_a^T \tanh(W_a[h_t; \bar{h}_s]) & concat \end{cases} \tag{2.5}$$

こうして得られたアライメントベクトル a_t を重みとして、入力系列全体の加重平均としてコンテクストベクトル c_t が与えられる。

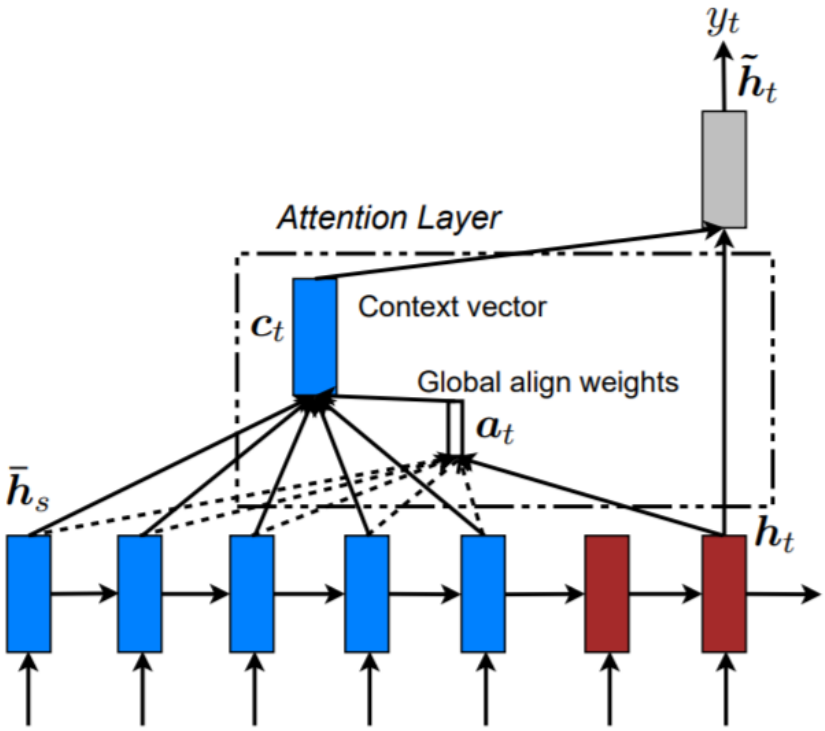

図 2.1 Global attentional model

2.3 Local Attention

Global attentional model（ソフト注意機構）には、各ターゲット単語に対して、入力系列全体を見る必要があるため、コストがかかり、段落や文書といった長い系列に対して不利であるというデメリットがある。このデメリットを解決する従来手法としてハード注意機構がある。ソフト注意機構がアライメントベクトル a_t を重みとして、入力系列全体の加重平均を計算していたのに対して、ハード注意機構ではアライメントベクトルの最も大きい成分に対応する入力系列のステップの隠れ状態をそのままコンテクストベクトル c_t として利用する。ハード注意機構は、推論時間が短くなる反面、微分できず、学習のためにより複雑な手法を必要とするデメリットがある。

これらの手法の改善策として、Local attentional model（局所注意機構）では、各ターゲット単語ごとに入力系列の一部に焦点を当てる。これにより、ソフト注意機構に比べ計算コストが低く、微分可能であるため、ハード注意機構に比べ学習が容易である。このモデルではまずデコーダのステップ t において、アライメント位置 p_t を生成する。コンテクストベクトル c_t は入力系列に対して窓枠 $[p_t - D, p_t + D]$ 内の隠れ状態の集合に対する加重平均として導出される。この p_t の生成には、2 つのバリエーションがある。1 つ目は入力系列とターゲット系列がほぼ単調にアラインメントされていると仮定して $p_t = t$ とする方法である。2 つ目は以下の式を用いて、位置を予測する方法である。

$$p_t = S \cdot sigmoid(v_p^T \tanh(W_p h_t)) \tag{2.6}$$

また、a_t を以下の式で計算する。

$$a_t(s) = align(h_t, \bar{h}_s) \exp(-\frac{(s - p_t)^2}{2\sigma^2}) \tag{2.7}$$

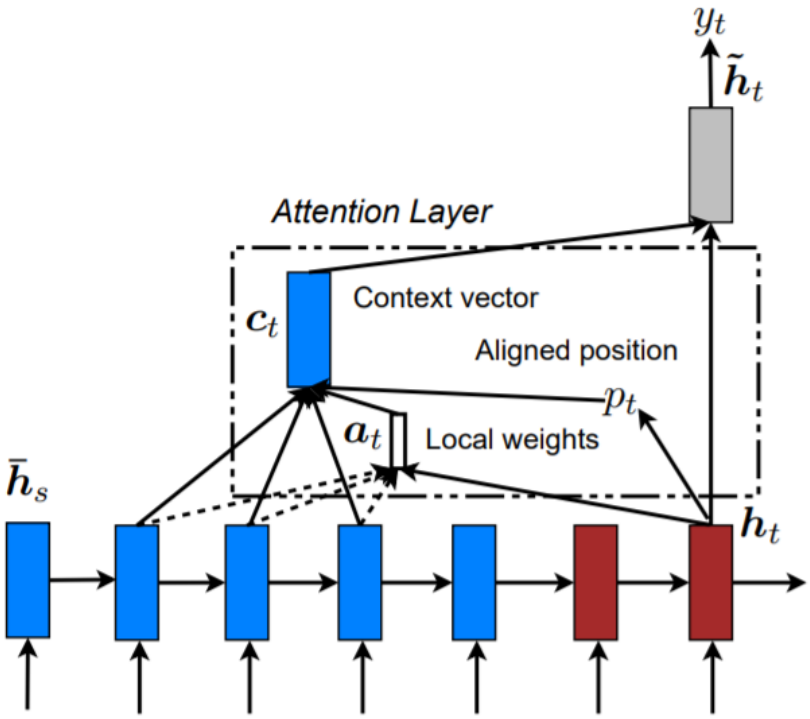

図 2.2　Local attentional model

3　実験

どんな実験設定か？どんな点で他の手法より良くなったか？

4.5M の文対（116M の英単語、110M のドイツ語の単語）からなる WMT'14 の train データを学習に使用した。語彙は両言語でもっとも頻繁に使用される単語の上位 50K で、これに含まれない単語はユニバーサルトークンに置き換えられる。我々の NMT システムを訓練するとき、長さが 50 単語を超える文のペアをフィルタリングし、我々が進むにつれてミニバッチをシャッフルする。使用した LSTM モデルは、4 つの層を持ち、それぞれが 1000 個のセルを持ち、1000 次元の埋め込みを持つ。パラメータは一様に $[-0.1, 0.1]$ で初期化し、最適化には SGD を使用して 10 エポック学習させた。学習率は 1 で開始し、5 エポック以降はエポックごとに学習率を半分にした。ミニバッチサイズは 128 で、正規化された勾配は、そのノルムが 5 を超えるたびにスケーリングし直した。さらに、LSTM に確率 0.2 のドロップアウトを使用した。ドロップアウトモデルについては、12 エポック学習させ、8 エポック後に学習率を半減させた。局所注意機構については、経験的に窓の大きさを $D = 10$ とした。様々な設定の 8 つの異なるモデルによるアンサンブル学習に未知語の置換を組み合わせたところ、英独翻訳タスクにおいて WMT ' 14、WMT ' 15 のいずれでも既存手法より高い性能を示した。独英翻訳タスクでは SOTA には及ばなかったものの注意機構の導入により改善が確認された。

4　結論・展望

他に改善するとしたらどこか？本論文では、NMT のための 2 つのシンプルで効果的な注意機構が提案された。すなわち、常に原文全体に注目するグローバルアプローチと、一度に原文のサブセットのみに注目するローカルアプローチである。英語とドイツ語の両方向の WMT 翻訳タスクで我々のモデルの有効性をテストしています。局所的注意機構は、ドロップアウトのような既知の技術をすでに組み込んでいる非注意モデルと比較して、最大 5.0BLEU の大きな改善をもたらした。英語からドイツ語への翻訳については、本研究のアンサンブルモデルは、WMT'14 と WMT'15 の両方で、既存のベストシステムを 1.0 BLEU 以上も上回る、新たな最先端の結果を確立した。一方でドイツ語から英語への翻訳では有効性は見られたが、SOTA には及ばなかった。