2O4-2

文脈を考慮したアテンションメカニズムの計算量の削減

A Flexible Model for Reducing Computational Cost of Attention Mechanism

朱 中元　　中山英樹
Raphael Shu　Hideki Nakayama

東京大学 大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo

Recently, attention mechanism plays a key role to achieve high performance for Neural Machine Translation models. It applies a score function to the encoder states to obtain alignment weights. However, as this computation is done for all positions in each decoding step, the attention mechanism greatly increases the computational complexity. In this paper we propose a novel attention model which can reduce redundant attentional computation in a flexible manner. The proposed mechanism tracks the center of attention in each decoding step, and computes position-based penalties. In the test time, the computation of the score function for heavily penalized positions are omitted. In our experiments, we found that the computation in the attention model can be reduced over 50% in average with almost no loss of accuracy.

1. はじめに

近年，高性能なニューラル機械翻訳モデルを構築するために，アテンションメカニズムが不可欠だと考えられている [Bahdanau 14]．アテンションメカニズムはエンコーダー（リカレントニューラルネットワーク）の全ての隠れ状態に対して，アライメントスコアを計算する．このスコアによって，エンコーダーの隠れ状態の重み付き平均が計算される．しかし，このスコアの計算をデコーディング時の各ステップにおいて行う必要があるため，ニューラル機械翻訳モデルの総合的な計算コストが増大し，翻訳時間が長くなっている．また，機械翻訳の前処理として，単語レベルへの分割ではなく，部分文字列 [Sennrich 16]，もしくは文字レベルへの分割 [Chung 16] が注目されている．文をより小さな要素に分割することによって，アテンションメカニズムの計算コストが更に増加している．

本論文では，翻訳中の文脈を考慮することで，アテンションの対象となるエンコーダーの隠れ状態の部分系列（アテンション視野）を動的に決定するニューラルネットワークモデルを提案し，翻訳精度を維持しつつ，最大限削減できるアテンションモデルの計算量を検証する．本論文では，この提案モデルを Flexible Attention と名付ける．

2. 関連研究

アテンションの計算量の削減に関する研究は主に 2 つのアプローチに分けられる：(1) アテンションの視野のサイズを縮めるアプローチ (2) アテンションの視野のサイズによらず計算量が不変なモデリングアプローチ．前者では，2015 年に提案された Local Attention[Luong 15] が挙げられる．Local Attention は，デコーディング時の各ステップにおいて次のアテンションの焦点を予測し，予測した焦点の周辺の固定窓幅中のエンコーダーの隠れ状態だけを考慮するモデルである．Local Attention では，窓幅のサイズはハイパーパラメータであり，固定されている．Local Attention の潜在的な問題として，入力系列が極めて長い場合，最適な窓幅の選択が困難であり，正しく焦点を予測するのも難しくなると考えられる．

連絡先: 朱 中元，東京大学 大学院情報理工学系研究科，shu@nlab.ci.i.u-tokyo.ac.jp

一方，計算量がアテンションの視野のサイズの変化に左右されないアテンションメカニズムも提案されている [Brébisson 16]．このモデルは質問応答タスクにおいてアテンションベースモデルとアテンションを用いないモデルの中間的な精度を達成した．

3. アテンションメカニズム

アテンションメカニズム (Soft Attention) は [Bahdanau 14] によって提案された手法であり，勾配逆伝搬のパスを短縮することによって，翻訳モデルを有効に学習できる．エンコーダーの隠れ状態を $\bar{\boldsymbol{h}}_{\boldsymbol{1}}, ..., \bar{\boldsymbol{h}}_{\boldsymbol{S}}$ とし，t ステップ目のデコーダーの隠れ状態を $\boldsymbol{h_t}$ とする．アテンションメカニズムは各隠れ状態に対してアライメントの重みを計算する：

$$a_t(s) = \frac{\exp(\text{score}(\boldsymbol{h_{t-1}}, \bar{\boldsymbol{h}}_{\boldsymbol{s}}))}{\sum_{s'} \exp(\text{score}(\boldsymbol{h_{t-1}}, \bar{\boldsymbol{h}}_{\boldsymbol{s'}}))} . \tag{1}$$

score 関数の計算では数多くのバリエーションが存在するが，ここでは，下記の score 関数の定義を使う:

$$\text{score}(\boldsymbol{h_{t-1}}, \bar{\boldsymbol{h}}_{\boldsymbol{s}}) = \boldsymbol{v_a}^{\top} \tanh(\boldsymbol{W_a}[\boldsymbol{h_{t-1}}; \bar{\boldsymbol{h}}_{\boldsymbol{s}}]) . \tag{2}$$

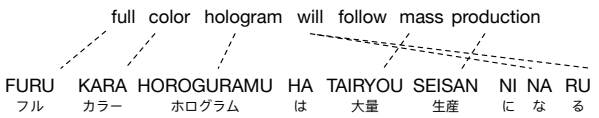
式 1 で計算したアラインメントの重みをもとに，次のようにしてコンテキストベクトルが計算される:

$$\boldsymbol{c}_t = \sum_s a_t(s) \bar{\boldsymbol{h}}_{\boldsymbol{s}} . \tag{3}$$

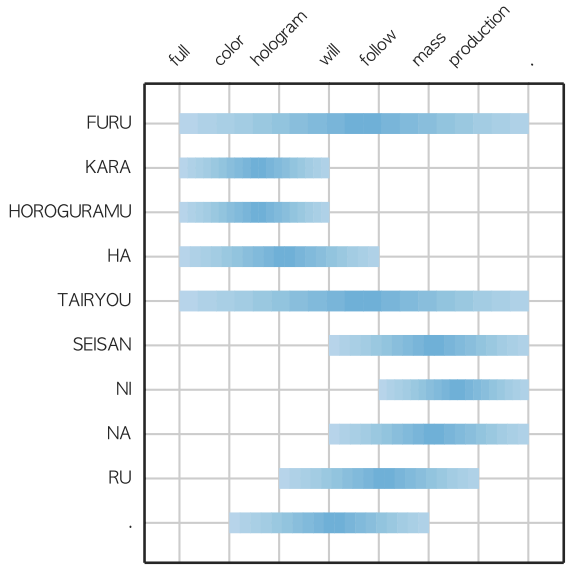
このコンテキストは毎デコーディングステップで計算され，入力としてデコーダーに与えられる．

4. 提案アテンションモデル

機械翻訳では，殆どの場合フレーズや文節のような局所的な意味単位の翻訳を行っている．英日翻訳のような文法的構造が異なる言語ペアを翻訳する際に，長距離の並び替えも必要だが，一文の翻訳ではあくまでも数回しか行う必要がないと思われている．例えば，図 1(a) に示されている翻訳例では，長距離の並び替えは一箇所だけ発生しているが，その他の単語に



(a)



(b)

図 1: (a) 長距離並び替えを持つ英日対訳ペアの例 (b) 提案モデルがデコーディングの各ステップにおいて予測したアテンションの窓幅

ついてほぼ並び替えなしで翻訳できる．これゆえに，我々はデコーダーが毎ステップで局所的にアテンションをかけても，必要な時だけ広範囲のエンコーダーの隠れ状態に対してアテンション計算を行えば，翻訳精度には影響がないと考える．本論文では，デコーダーの状態によって動的に最小限なアテンションの視野 (Vision Span) を予測するアテンションメカニズムを提案する．

提案モデルはまず各デコーディングステップで，アテンションモデルが計算した重みによって焦点 p_t を追跡する:

$$p_t = \sum_s a_t(s) \cdot s \,. \tag{4}$$

次に，下記の関数によって，前ステップの焦点と遠く離れるエンコーダーの隠れ状態に対してアテンションの重みにペナルティをかける:

$$\text{penalty}(s) = g(t) d(s, p_{t-1}) \,. \tag{5}$$

ここで，$d(\cdot)$ は距離関数であり，$g(t)$ は動的にペナルティの強さを制御するシグモイド関数である．Luong らによって提案した Local Attention[Luong 15] に従って，距離関数を下記のように定義する：

$$d(s, p_{t-1}) = \frac{1}{2\sigma^2}(s - p_{t-1})^2 \,. \tag{6}$$

ここで，σ は $g(t)$ が 1 を出力した時のペナルティの最大値を決めるハイパーパラメータである．$g(t)$ について，本論文では下記のように定義する：

$$g(t) = \text{sigmoid}(\boldsymbol{v_g}^\top \tanh(\boldsymbol{W_g}[\boldsymbol{h_{t-1}}; \boldsymbol{i_t}]) + b_g). \tag{7}$$

ここで，$\boldsymbol{v_g}$ と $\boldsymbol{W_g}$ はモデルのパラメータで，$\boldsymbol{i_t}$ は前のステップで出力した単語の単語ベクトルである．この式で，$\boldsymbol{i_t}$ を考慮する理由は，あるフレーズの翻訳が完了されているかどうかを判断するには直前の出力単語の情報が重要だからである．最終的に，従来のアテンションメカニズムのスコア関数に加えて，提案モデルは下記のようにアライメント重みを計算する：

$$a_t(s) = \frac{\exp(\text{score}(\boldsymbol{h_{t-1}}, \bar{\boldsymbol{h}}_{\boldsymbol{s}}) - \text{penalty}(s))}{\sum_{s'} \exp(\text{score}(\boldsymbol{h_{t-1}}, \bar{\boldsymbol{h}}_{\boldsymbol{s'}}) - \text{penalty}(s'))}. \tag{8}$$

提案モデルによって，アテンションモデルの視野をステップごとに細かく制御することができる．

4.1 計算量の削減

式 8 では，ある位置 s に対して，もしペナルティ項 $\text{penalty}(s)$ が十分大きいければ，スコア関数の出力によらずに，アライメント重み $a_t(s)$ が極めて小さい値になることが分かる．したがって，冗長な計算を避けるために，閾値 τ を設定し，$\text{penalty}(s) < \tau$ を満たす位置 s のエンコーダーの隠れ状態のみに対して，アテンションの重みを計算することができる．この不等式によって，毎ステップで，下記の範囲中のエンコーダーの隠れ状態だけが考慮される：

$$s \in \left(p_{t-1} - \sigma\sqrt{2\tau/g(t)},\ p_{t-1} + \sigma\sqrt{2\tau/g(t)} \right). \tag{9}$$

各デコーディングステップでは $g(t)$ の計算は一回しか行う必要がないので，このようにアテンションの視野を予測することによって，スコア関数の計算回数を大幅に削減することができる．

4.2 ファインチューニング

式 5 によって，$g(t)$ の値が大きいほど前ステップの焦点と離れる位置に対して大きいペナルティが与えられる．即ち，$g(t)$ が大きい値を出力すると，アテンションの視野が狭く制限されて，必要なスコア関数の計算回数が下がることが分かる．しかし，これまでの式の中で，明示的に $g(t)$ が大きい値を出力するように促進する仕組みがないため，アテンションの視野が最小限まで制限されているとは限らない．そこで，我々は下記の損失関数を使って，前述のアテンションモデルをファインチューニングすることを提案する:

$$J = \sum_{i=1}^{D} -\log p(y^{(i)}|x^{(i)}) - \beta \frac{1}{T} \sum_{t=1}^{T} g(t)^{(i)}. \tag{10}$$

この式の中で，従来のクロスエントロピーに加えて，$g(t)$ の平均値に基づいた損失項が追加されている．ここで，β はハイパーパラメータである．我々の実験では，この損失関数によって通常の学習の後に一エポックファインチューニングすることで，翻訳精度に影響を及ばずにスコア関数の計算回数を削減することが確認された．

5. 実験

本章では，提案アテンションモデルによって翻訳精度に影響を与えずに削減できるスコア関数の計算量を検証する．

表 1: 英日タスク及び独英タスクでの評価結果

	英日タスク			独英タスク	
	平均窓幅（単語数）	BLEU(%)	RIBES	平均窓幅（単語数）	BLEU(%)
Global Attention ベースライン	24.4	34.87	0.810	20.7	20.62
Local Attention ベースライン	18.4	34.52	0.809	15.7	21.09
Flexible Attention (τ=∞)	24.4	35.01	0.814	20.7	21.31
Flexible Attention (τ=1.2)	16.4	34.90	0.812	7.8	21.11
+ fine-tuning (τ=1.2)	**10.7**	34.78	0.807	**7.4**	20.79

5.1 実験設定

我々は英日翻訳及び独英翻訳タスクにおいて，提案モデルの性能を評価する．英日タスクは，ASPEC 英日翻訳データ [Nakazawa 16] の先頭の 150 万文ペアを用いる．独英タスクでは，WMT'15 で提供している対訳データ (450 万文ペア) を用い，テストデータは newstest2015 を利用する．テストデータでは，一文当たりの平均単語数はそれぞれ 24.4 単語と 20.7 単語ある．

前処理について，英日タスクでは「tokenizer.perl」と Kytea[Neubig 11] を使い，独英タスクでは [Li 14] と同じ前処理手法に従う．それぞれのタスクの語彙数 (softmax 層のサイズ) について，英日では 8 万と 4 万，独英では両方 5 万単語を選択し，未知語は全部「UNK」に書き換えた．50 単語以上の文を学習データから省いた上，ミニバッチサイズは 64 とした．

ネットワーク構造に関して，[Bahdanau 14] のモデルに従って，全ての LSTM の隠れ状態のサイズを 1000 にした．提案モデルのハイパーパラメータである σ をチューニングデータで経験的に 1.5 に設定した [*1]．ここで，σ はただ $g(t)$ が 1 を出力した時のペナルティを決定するもので，固定した視野の窓幅を決めるものではないことに注意されたい．

NMT モデルの学習には，我々は Adam[Kingma 14] を用い，初期学習率を 0.0001 に設定し，6 エポックを学習した．4 エポック目から，各エポックごとに学習率を半減させる．

翻訳結果の評価に関して，英日タスクでは BLEU と RIBES[Isozaki 10][*2]，独英タスクでは tokenized BLEU の評価スコアを使う．

5.2 提案アテンションモデルの評価

本章では，提案アテンションモデル (Flexible Attention) の閾値 τ を調整することによって，翻訳精度をほぼ失わず (チューニングデータで 0.5 BLEU 以内) に最大限削減できるスコア関数の計算量、即ちステップごとの平均窓幅を評価する．評価結果は表 1 に示す．比較するために，同じ条件で学習した通常のアテンションモデル (Global Attention) と Local Attention でのスコア関数の計算量も報告する．

通常のアテンションモデルでは，各デコーディングステップで全てのエンコーダーの隠れ状態を考慮する必要があるので，スコア関数の平均計算回数はテストコーパスの一文当たりの平均単語数になる．Local Attention は固定した窓幅を 2 1 単語に設定した時に英日と独英タスクでは一番よい性能を得た．この場合，平均計算回数について，Local Attention は英日タスクと独英タスクでそれぞれ 18.4 と 15.7 を達成した．

提案モデル (Flexible Attention) について，$\tau = \infty$ の翻訳精度との差を 0.5 BLEU 以内収めるように，ペナルティの閾値を $\tau = 1.2$ に設定した．評価結果に示されている通り，提案モデルによって，Global Attention と比べると平均窓幅を 50%以上削減することができる．英日タスクで，閾値 τ を変化させた場合の，翻訳精度とアテンションの平均窓幅の関係を図 2 で示す．平均窓幅を 10 単語以下にさげようとすると，翻訳精度が激的に悪化する傾向が見られる．

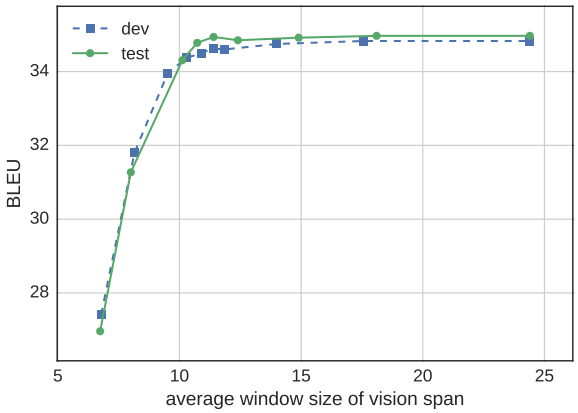


図 2: 英日翻訳タスクのチューニングデータとテストデータにおける翻訳精度とアテンションの平均窓幅の関係

5.3 文字ベース翻訳モデルでの評価

入力系列が極めて長い場合について，提案モデルの性能を検証するために，英日タスクの学習データを文字単位に分割して実験を行った．文字ベース翻訳モデルでの評価結果を表 2 に示す．学習データを文字単位に分割すると，全体的に翻訳精度が単語ベースモデルより劣るが，本実験は提案モデルによる計算量の削減を検証することを目的とする．評価結果で示されているように，入力が極めて長い場合でも，提案モデルは平均窓幅を約 46 %狭められることが分かる．よって，提案モデルが自動的に学習データの性質に適応して，ペナルティの強さを調整できることが分かる．

5.4 定性的分析

式 7 によって予測した窓幅の妥当性を検証するために，長距離並び替えが生じている翻訳サンプルを用いて各ステップにおいて予測した窓幅の可視化を行った．図 1 (a) に示している翻訳例では，4 番目の英単語が日本語の末尾に翻訳されている．図 1 (b) では，提案アテンションモデルが，各日本語単語を出力する際に予測した窓幅がで示す．

*1 σ の選択に関して，チューニングデータで $\sigma \in (1.5, 5, 7.5, 10)$ を試した上決定した．

*2 後処理について，WAT2016 の後処理手順に従い，Kytea を使った場合のスコアを報告する

表 2: 英日タスクにおける文字ベース翻訳モデルの評価結果

	平均窓幅	BLEU	RIBES
Global Attention ベースライン	144.9	26.18	0.767
Flexible Attention (τ=∞)	144.9	26.68	0.763
Flexible Attention (τ=1.0)	80.4	26.18	0.757
+ fine-tuning (τ=1.0)	**77.4**	26.23	0.757

図 1 の可視化によって，$g(t)$ 関数は翻訳中の文脈によって異なる値を出力していることが確認できる．例えば，4 番目の単語「は」を出力した後，次に翻訳する単語は離れた位置に在るので，$g(t)$ は小さい値を出力し，全域で次に翻訳する対象になる単語を探索する．この定性分析によって，アテンションモデルは毎ステップで全ての位置を考慮せずに，翻訳中の文脈によって，次の出力に必要な窓幅を予測することが可能であり，翻訳性能を維持しつつアテンションモデルの計算量を抑えられることが示唆される．

6. おわりに

本論文では，翻訳中の文脈に応じて動的にアテンションの窓幅を調整できるアテンションモデルを提案した． 英日タスクと独英タスクの実験では，アテンションモデルのスコア関数の平均計算量をで 50%以上安全に減らすことができた．文字単位の学習データにおいても，提案アテンションモデルは 46 %の削減率を達成することができた．

定性的な分析では，提案モデルは，特に長距離並び替えが必要な場合に，動的に大きい窓幅を予測できることが確認されている．本論文の実験によって，従来のアテンションメカニズムは冗長な計算を行っていることを確認できた．動的に不要な計算を減らすことで，NMT モデルは効率的に長いシーケンスを翻訳できること，または計算コストの高いスコア関数 [Socher 13, Yang 16] を組み込めることが期待される．

参考文献

[Bahdanau 14] Bahdanau, D., Cho, K., and Bengio, Y.: Neural machine translation by jointly learning to align and translate, *arXiv preprint arXiv:1409.0473* (2014)

[Brébisson 16] Brébisson, de A. and Vincent, P.: A Cheap Linear Attention Mechanism with Fast Lookups and Fixed-Size Representations, *arXiv preprint arXiv:1609.05866* (2016)

[Chung 16] Chung, J., Cho, K., and Bengio, Y.: A Character-level Decoder without Explicit Segmentation for Neural Machine Translation, in *ACL*, pp. 1693–1703 (2016)

[Isozaki 10] Isozaki, H., Hirao, T., Duh, K., Sudoh, K., and Tsukada, H.: Automatic evaluation of translation quality for distant language pairs, in *EMNLP*, pp. 944–952 (2010)

[Kingma 14] Kingma, D. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014)

[Li 14] Li, L., Wu, X., Vaıllo, S. C., Xie, J., Way, A., and Liu, Q.: The dcu-ictcas mt system at wmt 2014 on german-english translation task, in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 136–141 (2014)

[Luong 15] Luong, T., Pham, H., and Manning, D. C.: Effective Approaches to Attention-based Neural Machine Translation, in *EMNLP*, pp. 1412–1421 (2015)

[Nakazawa 16] Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S., and Isahara, H.: ASPEC: Asian Scientific Paper Excerpt Corpus, in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 2204–2208 (2016)

[Neubig 11] Neubig, G., Nakata, Y., and Mori, S.: Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis, in *ACL*, pp. 529–533 (2011)

[Sennrich 16] Sennrich, R., Haddow, B., and Birch, A.: Neural Machine Translation of Rare Words with Subword Units, in *ACL*, pp. 1715–1725 (2016)

[Socher 13] Socher, R., Chen, D., Manning, C. D., and Ng, A.: Reasoning With Neural Tensor Networks for Knowledge Base Completion, in *NIPS*, pp. 926–934 (2013)

[Yang 16] Yang, Z., Hu, Z., Deng, Y., Dyer, C., and Smola, A.: Neural Machine Translation with Recurrent Attention Modeling, *arXiv preprint arXiv:1607.05108* (2016)