_sim _sim

進捗報告

1 今週やったこと

- 実験内容のまとめ

2 要素技術

2.1 RNN と GRU

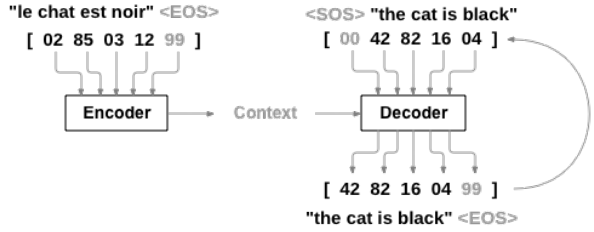
再帰型ニューラルネットワーク (RNN) とは [7], 回帰構造を持つニューラルネットワークである.

通常のニューラルネットワークでは, あるレイヤの出力は, 次のレイヤの入力として利用されるが,RNN では, 同じニューラルに対して, 当時刻の入力だけでなく, 前時刻の出力も入力し, その出力も次の時刻の入力として扱い, 再帰的構造を持ったニューラルネットワークである.

しかし,RNN は勾配消失と重み衝突の問題が存在するので, 長期的な特徴を学習できない, その問題を解決するため, ゲート付き回帰型ユニット (GRU) を導入する.GRU は [5]RNN と同じ, 当時刻の入力と前時刻の出力をモデルの入力とし, 当の出力と次の時刻の入力を出力する構造であるが,GRU の内部構造はリセットを制御するリセットゲートと更新を制御する更新ゲートを導入し, 前時刻の出力をどれくらい記憶し、利用するがを制御する構造である.GRU により,RNN は勾配消失と重み衝突の問題を解決する同時に, 高速に学習することができるめるセルである. 本実験は GRU を導入した Encoder-Decoder 構造を使用する.

2.2 seq2seq

Sequence To Sequence(seq2seq) とは [8],2014 年に Google が発表した, 言語モデルである. 従来の DNN(ディープニューラルネットワーク）が扱いにくい時系列データ問題を解決するため,seq2seq は RNN を用いた Encoder-Decoder モデルを導入した.Encoder は入力する時系列データをベクトルに圧縮し, そのベクトルを Decoder に渡し出力系列を生成する.RNN を利用したため,Decoder の出力は自動的に調整する. 始めと終わりのシンボルとして, ＜ SOS ＞と＜ EOS ＞のトークンを導入した, この中に, ＜ SOS ＞は始めの信号になり, ＜ EOS ＞は終わりの信号になる. 図 1 に本実験用の seq2seq モデルを示す. figure[H]



実験用 seq2seq の構造

3 データセット

3.1 ManyThings Bilingual Sentence Pairs

ManyThings データセットは [1]Tatoeba プロジェクトで収集され, 英語からフランス語や中国語などの 81 国の言語に対応するペアで集まるデータセットである. 収集されたデータは英語-他言語のペアを, 単語の少ない方から多いの方までソートされる.

本実験に使われるのは m,ManyThings データセットの英語-中国語データセットである. 英語-中国語データセットは 24,360 の英語-中国語文章ペアがあり, ペアの後ろに Tatoeba プロジェクトで収集されるに関する情報があるので, すべての文章に対し, 特殊符号と無関係情報を除去した.

3.2 ASPEC-JC コーパス

本実験の日本語 - 中国語対訳実験は,Asian Scientific Paper Excerpt Corpus(ASPEC) の日中学術論文抜粋コーパス (ASPEC-JC) を使用した [6].

表 1 は ASPEC-JC コーパスはデータセットごとに文章ペア数を示す.

表 1: ASPEC-JC

DataSet	Number of pairs
Train	672,315
Dev	2,090
DevTest	2,148
Test	2107

4　jieba と janome の紹介

英語とは違う, 日本語と中国語は単語を区別する空白がないので, 実験前に分かち書きの処理しないと, 文章はこのまま seq2seq モデルに入り, 実験はうまくいかないので,jieba と janome でデータを処理する.

jieba は [2]2013 年にリリースされ, 中国語文章の分かち書きに専用するライブラリである. 本実験は jieba の cut メソッドで中国語の文章を単語に分ける.

janome は [4]Python で記述された, 辞書内包の形態素解析ライブラリである. 本実験は janome の tokenizer メソッドで日本語の文章を単語に分ける.

5　実験の流れ

本実験は pytorch のチュートリアル [3] のプログラムで実行する.

5.1　データ処理

まずはデータセットの処理. データセットの処理は,jieba（中国語データ）,janome（日本語データ）で行う.

処理したデータは,word2index（単語からインデックス）,index2word（インデックスから単語）,word2count（単語から単語が出現した回数）というディクショナリの形式で保存する. ボキャブラリー数は n_words に保存する. 表 2 に各データセットの文章ペアを示す. 表 3 表 4 に各データセットの言語ごとのボキャブラリー数を示す.

表 2: Pairs

DataSet	Eng-Cmn	ASPEC-JP
Pairs	24360	2107

表 3: ManyThings_Vocabs

ManyThings	ENG	CMN
Vocab	7484	14716

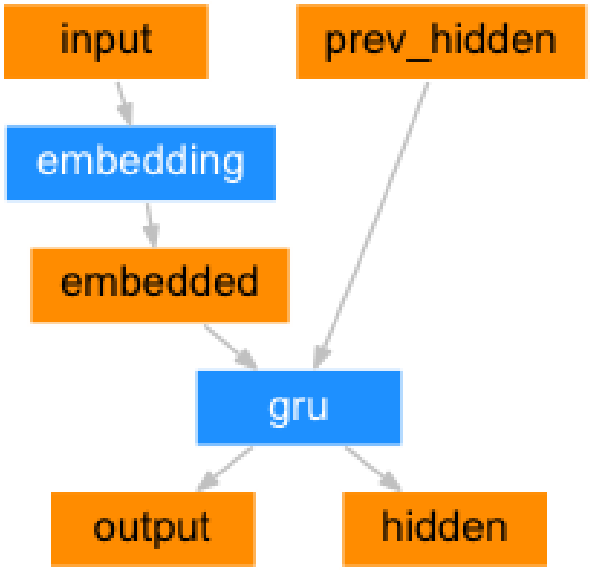
表 4: ASPEC-JP_Vocabs

ASPEC-JP	JPN	CMN
Vocab	6640	6844

5.2　モデルの実装

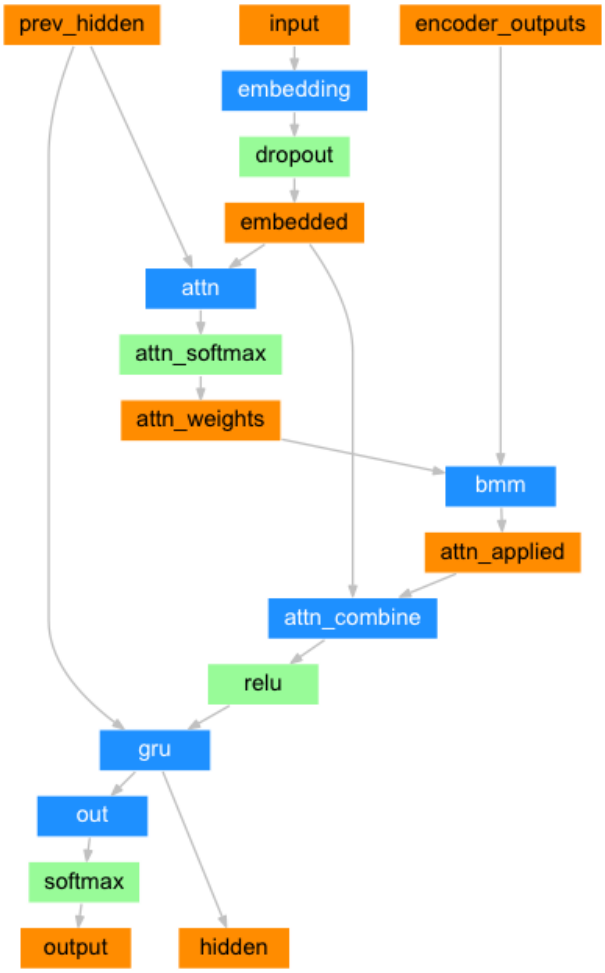
実験に使う seq2seq モデルの Encoder は,RNN である. 入力と前時刻の hidden 状態を入力した時,Encoder は現時刻の出力と hidden 状態を出力する, 出力した現時刻の hidden 状態は, 次の RNN の入力として次の入力と一緒に入力する. 最後の慎吾＜ EOS ＞を入力すると,Encoder の運作は停止し,Decoder の運作が始める. 図 2 に Encoder の構造を示す.

figure[H]



実験用 Encoder の構造

実験に使う Decoder は RNN の上に,Attention Mechanism を導入したモデルである.Decoder は入力, 先回のヒドゥン状態, エンコーダーの出力を入力とし, 出力, 現時刻の hidden 状態を出力とする. 図 3 に Decoder の構造を示す. figure[H]



実験用 Decoder の構造

5.3 トレーニング

トレーニングする前に, データペアを入力 tensor と目標 tensor の形式に変換し, そして入力 tesnsor を seq2seq モデルに入力し, 出力のデータも tensor の形式である.

モデルをトレーニングする時,Encoder 毎回の出力と最後の hidden 状態を追跡する. そして< SOS >トークンを Decoder の最初の入力として入力し,Encoder 最後の hidden 状態を Decoder 最初の hidden 状態にする. 実験誤差は,seq2seq モデルの出力 tensor と, データセットの目標 tensor により,NLLLoss で計算できる.

表 5 に実験に用いたパラメータを示す.

6 来週目標

- BiGRU を実装する

表 5: parameters

parameter	Value
input_size	7484
output_size	14716
hidden_size	256
n_layer	1
batch_size	1
dropout	0.1
attn_activation	Softmax
attn_combine$_act$	ReLU

- MeCab で日本語データを処理する

参考文献

[1] Bilingual Sentence Pairs Selected Sentences from the Tatoeba Corpus. `http://www.manythings.org/bilingual/`.

[2] "Jieba" (Chinese for "to stutter") Chinese text segmentation: built to be the best Python Chinese word segmentation module. `https://github.com/fxsjy/jieba`.

[3] NLP FROM SCRATCH: TRANSLATION WITH A SEQUENCE TO SEQUENCE NETWORK AND ATTENTION. `https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html`.

[4] Welcome to janome 's documentation! `https://mocobeta.github.io/janome/`.

[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014.

[6] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. Aspec: Asian scientific paper excerpt corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings*

of the Ninth International Conference on Language Resources and Evaluation (LREC 2016), pp. 2204–2208, Portorož, Slovenia, may 2016. European Language Resources Association (ELRA).

[7] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, Vol. 404, p. 132306, Mar 2020.

[8] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks, 2014.