修士論文 要旨

敵対的正則化を用いた文章埋め込み表現の学習

数理情報学専攻 48-196226　福地成彦
指導教員　田中久美子 教授

1　背景・関連研究

1.1　文章埋め込み

深層学習が自然言語処理の分野でも圧倒的な性能を見せている中、自然言語のベクトル表現は連続値の情報を入力する深層学習モデルにおいて必要不可欠な技術になっている。ベクトル表現は単語や文章など様々な単位を表現する。近年、文章埋め込みは、深層学習において高い性能を示している。文章埋め込みを獲得するためには、何らかのタスクで事前学習をすることが必要である。文章埋め込みは、文章の特徴抽出器でもあり、転移学習に応用することができる。本研究では、転移学習での文章埋め込みの性能を向上させるために、事前学習に対する敵対的正則化を探索した。転移学習では、事前学習で学習していない文章やフレーズ、単語をベクトルに埋め込むことになる。そこで、転移学習で文章埋め込みの性能を向上をさせるために、未知語や未知のフレーズを埋め込むことができるように正則化の方法を検討した。

本研究では、提案手法の検証のために、自然言語推論を事前学習に用いる文章埋め込みの既存手法 InferSent[1] を用いた。自然言語推論は、与えられた文章のペアに対して論理的含意、矛盾、自然のいずれかの論理的関係性のラベルを予測するタスクである。InferSent では、2 つの文章をそれぞれ単語ベクトルの系列に変換し、その系列を文章埋め込みを行う深層学習モデルが文章ベクトルのペアに変換する。事前学習用の補助の深層学習モデルが文章ベクトルのペアから関係性ラベルを予測する。

1.2　敵対的正則化

敵対的正則化は、機械学習モデルを誤識別させるような入力である敵対的サンプルを、学習データとして用いる正則化である [2]。敵対的サンプルは元のサンプルに微小な敵対的摂動を加えて生成される。敵対的摂動は、機械学習モデルの損失関数を最大化させるような摂動で、画像の場合、敵対的摂動は 1 反復の勾配法を用いて計算できる。敵対的正則化には、敵対的サンプルに対する誤識別率の低減や機械学習モデルの汎化性能向上に効果があると報告されている。

自然言語処理においても、文書分類や機械学習などのタスクで敵対的正則化の研究が報告されている [3, 4]。テキストを構成する単語は離散的な記号である。そのため、単語に対して直接、勾配法を用いて敵対的摂動を計算する事はできない。そこで、先行手法では単語ベクトルに対して勾配法を用いて敵対的摂動を計算し、敵対的正則化を行っている。

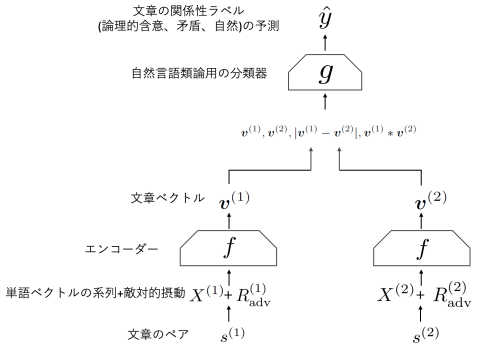

図 1. 提案手法の深層学習モデルの概要。InferSent のネットワーク構造を利用する。敵対的正則化として、単語ベクトルの系列 $X^{(1)}$, $X^{(1)}$ に対してそれぞれ敵対的摂動 $R^{(1)}_{\mathrm{adv}}$, $R^{(2)}_{\mathrm{adv}}$ を加算する。

1.3　研究の目的

自然言語処理においても様々なタスクで敵対的正則化の研究が報告されているが、文章埋め込みの事前学習に対しては敵対的正則化の研究はほとんど報告されていない。また、文章埋め込みでは、未知語の埋め込みが課題である。そこで、本研究の目的は、敵対的正則化が文章埋め込みの事前学習に効果があるかどうかの検証と、未知語の埋め込み性能向上を目指したとした敵対的摂動の検討である。

2　提案手法

本研究では、3 つの文章埋め込みの事前学習に対する敵対的正則化を提案・検証した。3 つの手法とも、InferSent に敵対的正則化を施している。また、単語には直接敵対的摂動を加えられないため、図 1 のように単語埋め込みに敵対的摂動を加えている。

2.1　近傍単語ベクトルを用いた摂動

未知語での埋め込み向上のため、文章の各単語を「言い換えた」文章でデータ拡張を行いたい。この手法では、[5] のように単語ベクトルの近傍探索によって各単

表 1. 自然言語推論での事前学習での正解率および SenEval[6] の応用タスクでの転移学習の転移スコア。事前学習の列は事前学習のテストデータに対する正解率 (%) を示している。左から 4 列目以降の各列は SentEval の各タスクに対応している。転移スコアは 0 から 100 までの値をとり、高いほど文章埋め込みの性能が高いことを表している。太字は各列で最も正解率が高いものを表している。

正則化	ϵ	事前学習	MR	CR	SUBJ	MPQA	SST-2	SST-5	TREC	MRPC	SICK-E	SICK-R
近傍探索	0.1	83.97	74.81	80.53	88.72	**87.13**	78.2	41.36	80.4	**74.38**	84.33	**87.06**
近傍探索	5.0	83.97	74.4	78.97	88.5	86.59	76.66	41.04	**83.4**	71.83	82.20	85.57
言語モデル	0.1	83.62	74.96	79.28	**88.99**	87.92	80.18	**42.76**	76.8	71.48	83.88	87.04
言語モデル	5.0	82.93	75.32	80.11	88.21	87.93	79.35	41.09	79.0	72.58	84.23	87.23
標準基底	0.1	**84.53**	**75.61**	**80.66**	88.97	86.84	80.23	42.35	80.2	72.99	**84.49**	87.10
標準基底	5.0	82.56	74.70	79.89	88.48	86.51	77.81	40.86	78.8	72.00	83.01	86.41
正則化なし	-	83.80	75.45	80.13	88.62	87.53	**80.45**	41.90	78.8	73.28	83.17	87.18

語に意味が近い「類語語」を k 個だけ探索し、元のサンプルの単語ベクトルから k 個の近傍単語の単語ベクトルへ向かう方向を摂動の基底ベクトルとする。基底ベクトル上で、損失関数を最大にする敵対的摂動を計算し、その摂動を敵対的正則化に用いる。

2.2 言語モデルを用いた摂動

「類義語」を探索するのに、単語ベクトルの近傍探索を用いると文章の文脈や連語を考慮することができない。そこで、ある位置までの単語の系列を入力に次の単語の生成確率を出力する言語モデルを用いて「類義語」を探索する。言語モデルの生成確率が上位を k 個の単語のベクトルを摂動の基底ベクトルとして、敵対的摂動を計算する。

2.3 標準基底での摂動

この方法では、未知語の埋め込みでの性能向上については考慮せず、文章埋め込みを行う深層学習モデル自体を正則化するために、単純に単語ベクトルに単語ベクトルの次元と同じ次元の敵対的摂動を加算する。このとき敵対的摂動の基底は単語ベクトル空間の標準基底で、単語ベクトル空間の全ての方向に摂動を加えることができる。

3 実験・考察

3 つの提案手法の転移学習での性能を評価するために、事前学習と転移学習からなる実験を行った。事前学習では、摂動の大きさを表すパラメーター ϵ を 0.1, 5.0 として提案した 3 つの敵対的正則化をそれぞれ施した。転移学習では、SentEval[6] という文章埋め込みの評価ツールを用いた。SentEval には複数の転移学習用のタスクがある。文章埋め込みを各タスクで文章の特徴量抽出器として用いた転移学習を行い、その転移学習の精度などのスコアを文章埋め込みの性能として評価する。

実験の結果、表 1 の結果を得た。事前学習においては、標準基底で敵対的摂動の大きさ ϵ が 0.1 のときに正解率が最大になったが、$\epsilon = 5.0$ では正解率が悪化しており、摂動の大きさが大きすぎる場合には、敵対的正則化が学習に対してノイズになったと考えられる。単語ベクトルの近傍探索および言語モデルを用いた正則化での正解率の変化は小さかった。転移学習では、3 つの提案手法で敵対的正則化なしの場合から転移スコアが向上しているタスクも見られた。しかし、各転移タスクでの正則化なしの場合を基準にしたときの各転移スコアの比率の平均が 0 以上だったものは、敵対的摂動の大きさ $\epsilon = 0.1$ のときの標準基底を用いた摂動の正則化のみだった。

この結果から、未知語への埋め込み能力向上を目指して「類義語」の方向に敵対的摂動を加えるよりも、文章埋め込みをする深層学習モデル自体を正則化するほうが転移学習に効果的であることが考察される。この理由として、単語ベクトルや言語モデルから得られた「類義語」が適切な類義語になっていなかった可能性が考えられる。

参考文献

[1] A. Conneau et al., “Supervised learning of universal sentence representations from natural language inference data,” EMNLP 2017

[2] I. J. Goodfellow et al., “Explaining and harnessing adversarial examples,” ICLR 2015

[3] Miyato et al, “Adversarial training methods for semi-supervised text classification,” ICLR 2017.

[4] M. Sato et al., “Effective adversarial regularization for neural machine translation,” ACL2020.

[5] M. Sato et al., “Interpretable adversarial perturbation in input embedding space for text,”IJCAI 2018.

[6] A. Conneau et al.,“SentEval: An evaluation toolkit for universal sentence representations,”LREC 2018.