敵対的サンプルを用いた対照学習の性能向上への取り組み

折口希実 [1]　小林一郎 [1]
[1] お茶の水女子大学
{g1620512,koba}@is.ocha.ac.jp

概要

アンカーとされる一つの画像から派生した画像群間の特徴量の類似度を最大化し，アンカーとは異なる画像から派生した画像群の特徴量とは類似度を最小化することにより自己教師あり学習を行う対照学習は，機械学習に新しい技術変革をもたらし優れた精度向上を実現した．本研究では，この対照学習を自然言語処理に導入した SimCSE [1] の枠組みに敵対的サンプルを導入することにより，さらに性能を向上させた対照学習を提案する．また，とくに，対照学習において明確な制約が存在しない負例の生成に着目し，様々な負例の導入方法を検討することにより，対照学習の性能向上を試みた．評価には意味的にテキストの類似性を測定できる SentEval の STS タスクを用いた．BERT ベースを用いた提案モデルでは先行研究モデルのスピアマンの相関 74.06%を 2.49%をも上回る 76.55%の相関を得られた．

1　はじめに

教師あり学習の精度を向上させるためには，大量のデータを必要とするが，データ全てに対してラベル付けを行うことは，多くの人的コストを要してしまう．一方，自己教師あり学習は，ラベルなしデータから自動で正解ラベル付きデータを生成することが可能であり，学習用データ構築のコストを大幅に下げることに有効な手段である．近年では，自己教師あり学習に対照学習を用いることで教師あり学習に匹敵する性能を実現している．対照学習とは，識別空間において識別対象と類似したデータは近くに，異なるデータは遠くに置かれるように学習する機械学習の手法の一つである．表現学習では対照学習を用いることで特徴量をうまく抽出できるモデルを得ることが可能であり，様々な下流タスクに利用できる．特に対照学習の枠組みを用いた教師なし学習は教師あり学習に匹敵することが示されている [2]．

本研究では，そのような対照学習をより高性能なものとするために，敵対的サンプル [3] を導入し，様々な観点から対照学習における負例の導入方法を検討する．

2　提案手法

本研究では，Gao ら [1] によって提案された埋め込み学習の枠組みである SimCSE モデルをベースにし，敵対的サンプルの導入や負例の生成方法を拡張したモデルを構築した．

2.1　SimCSE

対照学習では，アンカーとされるオリジナルのデータ x_i に対して，意味的に近いデータ x_i^+(正例と呼ぶ) のペアの集合 $D = \{(x_i, x_i^+)\}_{i=1}^m$ が与えられると，事前学習済みの汎用言語モデルによって入力をエンコードし，埋め込みベクトルを獲得する．一方，SimCSE モデルでは $x_i = x_i^+$ であり，データをドロップアウトマスクを用いて 2 度エンコードすることで僅かに異なる埋め込みベクトルの正例ペアを取得する．以下，汎用言語モデルを BERT として話を進める．通常 BERT への入力の際，バッチ単位で複数の文を同時にそれぞれベクトル化しているが，そのときのミニバッチ内のアンカーとする対象文以外の他の文については負例として扱う．損失関数である NT-Xent loss を用いて正例のペアのソフトマックスを 1 に近づけるように，そうでないものを 0 に近づけるように学習することで対照学習を行う．

SimCSE モデルの損失関数は式（1）で示される．

$$\ell_i = -\log \frac{e^{sim(\mathbf{h}_i^{z_i}, \mathbf{h}_i^{z_i'})/\tau}}{\Sigma_{j=1}^N e^{sim(\mathbf{h}_i^{z_i}, \mathbf{h}_j^{z_j'})/\tau}} \quad (1)$$

ここで，N はミニバッチ数であり，z はドロップアウトマスク，τ はソフトマックスの温度パラメータである．$\mathbf{h}$ はデータ x，x^+ それぞれの特徴表現を示す．また，$sim(\cdot)$ はコサイン類似度の関数であり，

文同士の類似度を計算している．正のペアは 1 に，負のペアは 0 に近づくように学習を行う．

対照学習におけるデータの拡張では，単語の削除や入れ替え等の既存手法が主流であったが [4]，ドロップアウトマスクのみというシンプルな仕組みで，SimCSE モデルは既存手法の精度を更新する十分な精度を実現している．

2.2 敵対的サンプルの対照学習への導入

敵対的学習 [3] において，モデルの出力を変えてしまうようなノイズベクトルのことを「敵対的摂動」と呼び，意図的に作成される．作成方法については，主に 2 つの設定 (White-Box, Black-Box) により異なり，モデル知識の有無で分かれる．モデル知識がない場合が Black-Box であり，ある場合の設定が White-Box である．本研究での取り組みは，White-Box に該当するため，以下それについてのみ言及する．敵対的サンプルを作成する手法の代表的なものの 1 つに Fast Gradient Sign Method (FGSM) がある [3]．摂動は，式 2 で示すように与えられる．

$$x = x + \epsilon sign(\nabla_x Loss(x, y)) \tag{2}$$

ここで，x は入力データ，y はそのラベルである．また，$sign$ 関数は正の値を +1，負の値を −1 にする関数で，ϵ は摂動に対する調整係数である．損失を最大化する方向へ入力 x を ϵ ずつ更新していく．

モデルを誤認識させてしまう敵対的サンプルを混ぜて学習を行うことでより堅牢なモデルを得られることが報告されており [5, 6]，本研究においても対照学習における正例に対して，敵対的サンプルを生成させ，モデルの学習に用いた．図 1 に概要を示す．正例から生成した敵対的サンプルを負例のように扱い，損失関数は式 (1) を用いた．

2.3 対照学習の性能向上に向けて

敵対的サンプルの導入により，対照学習における正例の拡張を行なったのと同様に，負例に関しても拡張を考える．対照学習における負例の生成においては，制約が緩く，工夫の余地がある．以下に，本研究で導入した負例の生成について記す．

1. **損失関数のコサイン類似度を用いた負例の拡張**
   Contrastive Loss を計算過程では文の特徴表現の類似度としてコサイン類似度にて算出する．負例は基本的に正例と全く似つかないものであるため負例ペアの類似度はかなり低いものとなってしまう．そのため正例ペアのものと負例ペアのものとの間には類似度に一定の乖離が存在している．その乖離部分は正例ペアに近い境界部分の負例のペアの類似度を表していると推測し，その部分の拡張を行うことは境界部分のデータを拡張していることと同義で，潜在空間をより補完できるのでははないかと考えた．損失関数は式 (3) のように表せる．

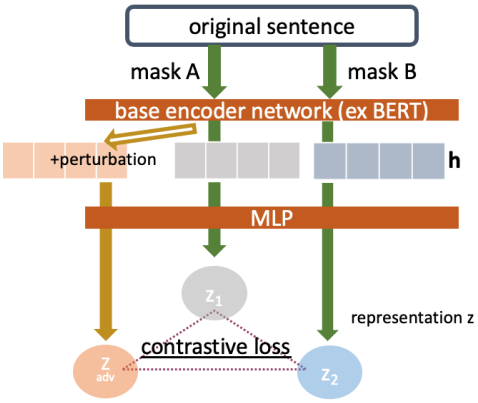

図 1 提案モデル

$$\ell_i = -\log \frac{e^{sim(\mathbf{h}_i^{z_i}, \mathbf{h}_i^{z_i'})/\tau}}{\Sigma_{j=1}^{N} e^{sim(\mathbf{h}_i^{z_i}, \mathbf{h}_j^{z_j'})/\tau} + \lambda \cdot \Sigma_{k=1}^{M} e^{simA/\tau}} \tag{3}$$

$simA$ は正例ペアと負例ペアの間のコサイン類似度の間の範囲から生成される擬似コサイン類似度であり，M は生成される数を示し，λ は調整係数を示すハイパーパラメータである．

2. **margin を考慮した負例の生成**
   深層距離学習 [7] において，アンカー，正例，負例の 3 つ組の埋め込み空間の状態を捉えた損失関数 triplet loss の設定やデータ間の距離を捉える margin といった考え方がある．アンカーから遠く離れた負例は学習にとってさほど重要な意味を持たず，アンカーに近ければ近いほど効果的な学習ができるため，その効果的な学習を可能とする範囲が margin である．学習にとって理想的な負例は，正例に極めて近いが正例ではない境界の近くに存在するものと考えることができる．ここでは，学習にとって有益な負例となるデータが存在する範囲のことを margin とし，文のベクトルにノイズを加え，正例を負例に近づけるように拡張させた敵対的サンプル（正例の限界点とみなす）との距離を考

表 1　実験結果 (STS タスク)

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
SimCSE	66.07	79.99	71.14	78.91	77.74	75.16	69.39	74.06
+GS-InfoNCE(GS)	67.24	81.41	73.05	79.21	78.38	75.30	69.59	74.88
ours								
+adversarial examples(ae)	68.88	**82.15**	73.71	80.97	79.13	**77.98**	69.71	76.08
+GS+ae	66.78	79.49	72.00	79.46	78.80	76.02	70.28	74.69
+cos-sim	**69.61**	81.67	73.65	81.16	79.46	77.74	72.39	76.53
+cos-sim+ae	69.47	81.65	73.41	**81.41**	**79.60**	77.85	**72.43**	**76.55**
+margin+ae	68.28	81.61	**73.77**	79.86	78.84	76.57	70.40	75.62

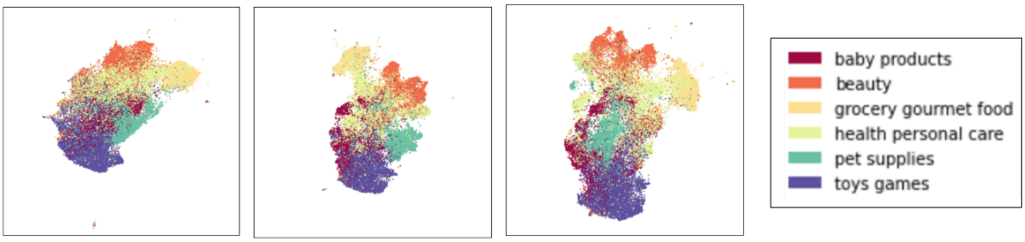

図 2　埋め込みの視覚化：(左) BERT (中央) BERT を用いた SimCSE モデル (右) 敵対的サンプルを用いた提案モデル

慮することにより，有効な負例を見つけ学習に用いる．具体的には，アンカーからもっとも遠いとみなされる正例の限界点を敵対的サンプルとし，負例との間の範囲を margin としている．敵対的サンプルと新たな負例との間のベクトルを生成することで margin の範囲にある負例とした．

3. **GS-InfoNCE**

Wu ら [8] は，SimCSE をベースとして，その損失関数の分母に，Gauss 分布で表現されるノイズを追加した手法である Gaussian Smoothing InfoNCE (GS-InfoNCE) を提案している．まず，ガウス分布から文のベクトルと同じ次元で M 個のガウスノイズベクトル $\mathbf{g}$ をランダムにサンプリングする．これらのベクトルはサンプルと高い信頼性を持つ負のペアとなる．そのような負のペアを生成して負例を拡張することによって，表現空間をより充実させ，滑らかなものとする．次のような式にて損失関数を示す．

$$\ell_i = -\log \frac{e^{sim(\mathbf{h}_i,\mathbf{h}_i^+)/\tau}}{\Sigma_{j=1}^{N} e^{sim(\mathbf{h}_j,\mathbf{h}_i)\tau} / + \lambda \cdot \Sigma_{k=1}^{M} e^{sim(\mathbf{g}_k,\mathbf{h}_i)\tau}} \tag{4}$$

この手法では，SimCSE モデルを上回る結果となったことが示されており，本研究での提案手法との比較のための追実験を行った．

表 2　実験設定

バッチサイズ	64
学習率	3e-5
温度パラメータ	0.05
metric	Spearman correlations
hugging-face	bert-base-uncased，roberta-large

3　実験

提案モデルを含め，2.2，2.3 項で述べた方法について実験を行い，対照学習の性能を向上させる方法を検証する．

3.1　実験設定

表 2 に実験設定を示す．パラメータ等は [1] を元にして行った．

評価方法　評価方法については，先行研究と同様に SentEval: evaluation toolkit for sentence embeddings [1)] を用いる．SentEval を使用し，STS (Semantic Textual Similarity) を用いてタスク遂行に際して、精度を向上させるような文の埋め込み表現が獲得できたかを評価する．STS は，文章の等価性を評価するタスクであり，2 つの文の関係について類似度を対象として評価する．本研究では STS の 7 つのタスク [2)] を用いて評価を行った．対照学習は「似た表現を近くに，異なる表現を遠くへ」と学習させるものである

1)　https://github.com/facebookresearch/SentEval
2)　STS12，STS13，STS14，STS15，STS16，STS Benchmark，SICK Relations

表 3　学習済み言語モデルの比較 (Avg.)

	SimCSE	Ours
BERT-base-uncased	74.06	76.08
RoBERTa-large	**77.73**	**78.16**

ので，類似を評価する STS タスクを用いることで文の埋め込み表現のパフォーマンスについて評価が可能である．また，学習された潜在空間のクラスタ化がなされているかを検証するために UMAP[3] を用いて，データ数約 40 万の amazon の製品レビューでカテゴリのラベルのついたデータの潜在空間の視覚化を行う．UMAP とは t-SNE よりも高速・高性能に次元削減・可視化する手法であり，似たカテゴリ同士は近くに，似ていないカテゴリ同士は遠くに配置される．

3.2　実験結果

STS の 7 つのタスクに対し，2.2，2.3 項で述べた提案手法を用いて評価を行った結果を先行研究らの追実験の結果と比較し表 1 に示す．実験で用いた提案モデルは 5 つであり，手法を組み合わせたモデルも用いた．表 1 で示すモデルの事前学習済み汎用言語モデルは全て BERT を用いて行ったものとなっている．また学習済み汎用言語モデルとして BERT と RoBERTa それぞれを採用した際の STS のタスクの結果の比較を表 3 に示す．この結果は STS の 7 つのタスクのスピアマン相関の平均値であり，先行研究モデルの SimCSE と敵対的サンプルを用いた提案手法で比較を行った．

また，敵対的サンプルを用いた提案手法と先行研究モデルとで可視化し比較したものを図 2 に示す．左から BERT，BERT ベースのモデルを用いた SimCSE モデル，敵対的サンプルを用いた提案モデルの順になっている．

3.3　考察

表 1 に示すように，5 つの提案モデルは先行研究モデルである SimCSE と Wu らによって提案されたモデル [8] のスコアのほとんど全てを上回ったことがわかる．これにより，2.2，2.3 項で示した手法は全て効果的に作用したといえる．しかし，5 つの提案モデルの結果はタスクによってベストスコアを出すモデルは異なっているためどの手法が優れているかについては一概にはいうことはできない．だが，基本的には敵対的サンプルを単独や組み合わせて用いたモデルが，7 つのタスク中 6 つのタスクで最も高いスピアマンの相関を得ることができており，かなり効果的な手法の一つと考える．また，**ours**(+GS+ae) モデルのように GS-InfoNCE モデルと **ours**(+adversarial examples) モデルを組み合わせることでスコアを落としてしまうこともあり，安易に手法を組み合わせることで逆効果になり得ることがわかる．

表 3 からは 7 つのタスクのスピアマン相関の平均スコアが学習済みの汎用言語モデルとして BERT を用いるより RoBERTa を用いる方が SimCSE モデル，提案モデルのベストスコアでそれぞれ 3.67%，2.08%上回った．このことから RoBERTa をベースとしたモデルの方が優れていることを示している．

図 2 では文の埋め込み表現のパフォーマンスについて BERT，SimCSE モデル，提案手法の一つである敵対的サンプルを用いた提案モデルの 3 つで比較を行っているが，敵対的サンプルを用いた提案モデルが最もクラスタが分かれており，効果的な対照学習はより解きほぐされた表現空間を生成することに繋がっていることがわかる．

4　おわりに

本研究では，自然言語処理において効果的な対照学習を行うことができるように敵対的サンプルを用いたデータの拡張方法の模索や負例の生成方法について検討をし，実験を行った．実験を通じて，敵対的サンプルを単独，または組み合わせたモデルが STS の 7 つのタスクのほとんどのベストスコアを達成しており，有効な手法であることを示した．損失関数のコサイン類似度を用いて負例を拡張したモデルについてもかなり高いスピアマン相関を達成しており，さらに改善の余地のあるパラメータの調整等を行うことでより良いスコアを得られるのではないかと考える．今後の課題として、それらを行った実験を行い性能向上に努め，また，具体的な潜在空間を可視化して負例のデータ分布を margin の中で一様に補完するような負例の生成をしたいと考えている．提案手法は，自然言語処理において優れた表現の獲得を可能とし，また潜在空間をより明確に整理が行えるため，分類系統のタスクにおいて特に役立つと期待している．

3）　https://umap-learn.readthedocs.io/en/latest/index.html

参考文献

[1] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In **Empirical Methods in Natural Language Processing (EMNLP)**, 2021.

[2] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. **arXiv preprint arXiv:2006.10029**, 2020.

[3] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.

[4] Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. Clear: Contrastive learning for sentence representation, 2020.

[5] Chih-Hui Ho and Nuno Vasconcelos. Contrastive learning with adversarial examples. **ArXiv**, Vol. abs/2010.12050, , 2020.

[6] Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, **Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual**, 2020.

[7] Kaya and Bilge. Deep metric learning: A survey. **Symmetry**, Vol. 11, No. 9, p. 1066, August 2019.

[8] Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. Smoothed contrastive learning for unsupervised sentence embedding, 2021.