自然言語処理における解釈可能な敵対的摂動の学習

佐藤 元紀 [1]　鈴木 潤 [2,3]　進藤 裕之 [1,3]　松本 裕治 [1,3]
[1] 奈良先端科学技術大学院大学 情報科学研究科
[2] NTT コミュニケーション科学基礎研究所
[3] 理化学研究所 革新知能総合研究センター
{sato.motoki.sa7, shindo, matsu}@is.naist.jp
suzuki.jun@lab.ntt.co.jp

1 はじめに

ニューラルネットワークは，画像，音声，自然言語処理の分野で広く実用化が進んでいる．ニューラルネットワークの研究題材の一つとして，入力に対して小さな摂動（ノイズベクトル）を与えることで，学習済み，あるいは学習中のニューラルネットを誤分類させる敵対的サンプル (Adversarial Example) を用いた方法論が近年多くの研究者に注目されている [15, 4]．例えば画像認識の分野では，正しく分類できているパンダの画像に対して摂動を加え，間違った分類をさせることが報告されている [4]．この摂動はニューラルネットワークの損失関数の勾配を用いて求めることができる．さらに，敵対的学習 (Adversarial Training) と呼ばれる摂動を加えた画像を正しく分類する損失関数を追加することで汎化性能が高まることも報告されている [4]．

自然言語処理の分野でも敵対的学習の適用は検討されており，汎化性能が向上することが報告されている [11]．しかし，自然言語処理と画像の分野の大きな違いとして，画像の分野では入力は連続的 (例:RGB 値) であるのに対して，自然言語処理の分野は入力は離散的 (例:単語) である点が挙げられる．この違いから，画像の場合は，入力画像に摂動を加えた画像は RGB の範囲 (0-255) で人間に解釈できる形で表現することができるが，自然言語処理の場合，入力の単語ベクトルに摂動を加えたベクトルがどのようなシンボルに対応するのか解釈することは難しい．これは，単語ベクトルが張る D 次元の単語ベクトル空間中で，各単語は 1 点に相当するため，たとえ数百万単語を用いたとしても，その空間中にまばらにしかベクトルが分布しない状態であることに起因する．

そこで本研究では，自然言語処理における敵対的学習の手法として，単語ベクトルに摂動を加える際に実際にシンボル（単語）が存在する方向に限定して摂動を加える手法を提案する (図 1)．提案手法には，少なくとも以下の 2 つの利点があると考えられる:

- 摂動の向きと大きさを単語の方向に基いて可視化することができる．
- 摂動を加えたことにより生成される擬似データを人間が解釈可能なシンボルに復元することができる．

つまり，提案法は，画像処理の分野で研究が盛んに進んでいる敵対的学習法を，入力が離散シンボルになるという自然言語処理分野の特性を鑑みて，より人間の解釈性が向上するように改良した取り組みである．本稿では，極性分類問題を用いて提案法の有効性を検証する．

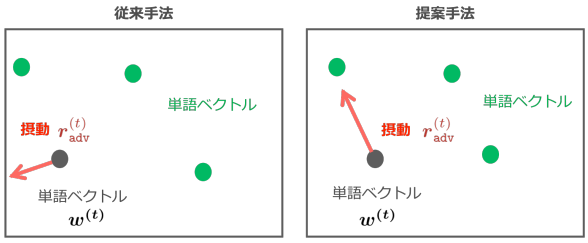

図 1: 従来手法 (左) と提案手法 (右) の違いを示す．従来手法では，単語ベクトルが存在しない方向に摂動を加えることができるが，提案手法は摂動が単語ベクトル方向へ限定することを目的とする．

2 関連研究

分類器の出力を誤らせる摂動は，文献 [15] によって問題提起された．この摂動は勾配を用いて求められる [15, 4]．敵対的サンプルを正しく分類するような損失関数を用いる敵対的学習 [4] も提案された．

自然言語処理における敵対的サンプル，及び敵対的学習の研究として，読解システム [7] や，ニューラル機械翻訳システム (NMT)[1, 6] の出力を変える敵対的な入力文を作る試みが行われている．敵対的な入力文を作ることで，モデルの性質を人間が理解することができるという利点がある．敵対的な入力文の作成は，人手で作る手法 [7] や，入力文の単語を同義語に置き換えた文を大量に作成し，予測が誤るか探索するもの [13] などがあるが，計算コストが高いという問題がある．

文献 [11] では，文書分類タスクにおいて，単語ベクトルに摂動を加える手法が提案されている．また，大きく汎化性能が向上することを示した．このように自然言語処理の分野でも徐々に敵対的学習の枠組みを用いることの有効性が示されるようになってきた．一方，実際に単語ベクトルに摂動が加わったことによってどういった効果が得られているのかを自然言語処理タスクで定性的に解釈することは，画像処理タスクほど容易ではない．画像処理タスクの場合は，摂動を加えたデータを画像として擬似的に復元することが比較的容易である．しかし，自然言語処理タスクでは，摂動を加えた擬似データを現実の単語や文章に対応させて復元することは困難である．これは摂動が加わったベクトルが現実のデータ点と一致することがほぼ起こらないことに起因する．

このように摂動をデータに加えて汎化性能が上がることを議論している研究は存在するが，その解釈性を議論した研究は，筆者らの知る限り現時点では報告されていない[1]．

3 敵対的摂動

本稿では，簡単な例として二クラス分類問題を対象に議論を行う．ここでは，入力を文章，出力を正 (Positive) または負 (Negative) の二クラスと仮定する．

入力文 X は T 個の単語系列で構成されるとする．また，入力文中に出現する単語集合 (語彙) を $\mathcal{V}$ とし，$x^{(t)} \in \mathcal{V}$ を t 番目の単語とする．この時，入力文 X は $X = (x^{(1)}, \ldots, x^{(T)})$ と表すことができる．簡単のため，系列 $(x^{(1)}, \ldots, x^{(T)})$ を短縮して $(x^{(t)})_{t=1}^{T}$ と記述する．次に，入力文 X に対する正解ラベルを y とする．$x^{(t)}$ に対応する単語ベクトルを $\boldsymbol{w}^{(t)} \in \mathbb{R}^D$ とする．以降，入力 X' はベクトルの系列と考えることとする．つまり，$X' = (\boldsymbol{w}^{(t)})_{t=1}^{T}$ である．

ここで，各単語ベクトルに摂動ベクトル $\boldsymbol{r}_{\mathtt{adv}}^{(t)}$ を加えることを考える．$\boldsymbol{r}_{\mathtt{adv}}^{(t)}$ は，単語ベクトルと同じ D 次元とする．

3.1 従来法

ここでは，提案法のベースとなる文献 [11] で述べられている手法の概略を述べる．文献 [11] では，各位置 t の単語ベクトル $\boldsymbol{w}^{(t)}$ に対応する摂動ベクトル $\boldsymbol{r}_{\mathtt{adv}}^{(t)}$ を以下の式を用いて計算する．

$$\boldsymbol{r}_{\mathtt{adv}}^{(t)} = \nabla_{\boldsymbol{w}^{(t)}} \mathcal{L} \quad (1)$$

$$\mathcal{L} = \log p\Big(y \mid (\boldsymbol{w}^{(t)})_{t=1}^{T}\Big) \quad (2)$$

$\boldsymbol{r}_{\mathtt{adv}}^{(t)}$ は，分類器の損失関数 $\mathcal{L}$ が増大する方向の勾配に相当する．つまり，求めた摂動ベクトル $\boldsymbol{r}_{\mathtt{adv}}^{(t)}$ を入力に加えたものは，分類器の損失関数を増大させ，誤分類を発生させやすくなると解釈できる．そこで学習の際に，摂動ベクトルを加えた入力を正しく分類するような損失関数を加えることで，分類器の汎化性能が向上することが期待できる．より具体的には，以下の損失関数を用いる．

$$\mathcal{J}_{\mathtt{adv}} = -\frac{1}{N}\sum_{n=1}^{N} \log p\Big(y \mid (\boldsymbol{w}^{(t)} + \epsilon \boldsymbol{r}_{\mathtt{adv}}^{(t)})_{t=1}^{T}\Big) \quad (3)$$

ϵ はハイパーパラメータであり，文献 [11] では固定の値 $\epsilon = 5.0$ を用いていた．

3.2 提案手法

前節で述べた摂動ベクトル $\boldsymbol{r}_{\mathtt{adv}}^{(t)}$ は，式 (1) からもわかるように，実際のデータ点とは無関係に計算される．よって，単語ベクトルに摂動を加えたデータ点 $(\boldsymbol{w}^{(t)} + \boldsymbol{r}_{\mathtt{adv}}^{(t)})$ が，実データ点でどのような単語に対応しているのか分からない．それに対して，提案法では，実際に存在する単語ベクトルの方向から，方向を選択し摂動ベクトルとする．そのために，ある単語 $x^{(t)}$ をどの単語に置換すると間違いが起こりやすいかを考慮しながら摂動を計算する方法を提案する．提案手法は 3 つのことを考慮する必要がある．(i) 語彙の選択 (ii) 摂動の方向 (iii) 摂動の大きさについて述べる．

(i) 語彙の選択

ここで，$|V|$ 個の全ての単語を考慮して摂動ベクトルを計算すると，最悪 $|V|^2$ の計算量が必要となり，計算コストが大きくなりすぎる問題がある．そこで，計算量を減らすために，部分単語集合 $\mathcal{V}'$ を考える ($|\mathcal{V}| > |\mathcal{V}'|$)．本研究では，単語 $x^{(t)}$ それぞれに対する周辺単語を，コサイン類似度に基いて計算し，上位 $|\mathcal{V}'|$ 個の部分単語集合 $\{w_k^{(t)} | k = 1 \ldots |\mathcal{V}'|\}$ を $\mathcal{V}$ の代わりに用いる (実験では $|\mathcal{V}'| = 10$ とした)

単語ベクトル $\boldsymbol{w}^{(t)}$ から単語ベクトル $\boldsymbol{w}_k^{(t)}$ への方向ベクトルは $\boldsymbol{d}_k = \boldsymbol{w}_k^{(t)} - \boldsymbol{w}^{(t)}$ である．

(ii) 摂動の方向の選択

まず，$|\mathcal{V}'|$ 次元のベクトル $\boldsymbol{\alpha}$ を定義する．まず，$\boldsymbol{\alpha}$ の k 番目の要素 α_k は，事前に選択された $|\mathcal{V}'|$ 個の各周辺単語中の k 番目の単語への方向ベクトルに対する重み付けスコアとする．このとき，提案手法における摂動ベクトル $\boldsymbol{r}_{\mathtt{adv}}^{(t)}$ を以下のように定義する．

$$\boldsymbol{r}_{\mathtt{adv}}^{(t)} = \sum_{k=1}^{|\mathcal{V}'|} f(\boldsymbol{\alpha}, k)\boldsymbol{d}_k \quad (4)$$

ここで，先行研究 [11] が摂動を勾配で計算したのと同様に $\boldsymbol{\alpha}$ は次の式を用いて計算する．

$$\boldsymbol{\alpha} = \nabla_{\boldsymbol{\alpha}} \log p\Big(y \mid (\boldsymbol{w}^{(t)} + \boldsymbol{r}_{\mathtt{adv-Rand}}^{(t)})_{t=1}^{T}\Big) \quad (5)$$

ここで，$\boldsymbol{\alpha}_k$ を導出する際 $\boldsymbol{r}_{\mathtt{adv-Rand}}^{(t)}$ は式 4 における $\boldsymbol{\alpha}_k$ をランダムに初期化したものとする．

次に，スコア関数 f について説明する．スコア関数 f の定義によって，性質の異なる摂動を考えることができる．例えば，本稿では以下の 2 種類の関数を用いる．

$$f^{\mathtt{Uniq}}(\boldsymbol{\alpha}, k) = \begin{cases} \frac{1}{\|\boldsymbol{\alpha}\|_2^2}\alpha_k & \text{if } \max(\boldsymbol{\alpha}) = \alpha_k \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

$$f^{\mathtt{Sum}}(\boldsymbol{\alpha}, k) = \frac{1}{|\mathcal{V}'|\|\boldsymbol{\alpha}\|_2^2}\alpha_k \quad (7)$$

スコア関数に $f^{\mathtt{Uniq}}$ を用いた場合，摂動ベクトルは 1 つの単語を選択し，その方向ベクトルを利用するという解釈になり，$f^{\mathtt{Sum}}$ を用いた場合，全ての単語方向の平均的な方向の摂動を使うと解釈できる．それぞれの提案手法を `Uniq`，`Sum` と呼ぶ．

(iii) 摂動のノルム

従来法（式 3）では，摂動を加える際に固定のハイパーパラメータ ϵ を用いていた．提案手法は，摂動を最も近い単語ベクトルに近づけるように加える．摂動は実データ方向を向いているため，摂動のノルムを自動決定することができる．摂動 $\boldsymbol{r}_{\mathtt{adv}}^{(t)}$ と，コサイン類似度が

[1] 自然言語処理における敵対的サンプルの生成に関する研究 [3] があるが，One-hot なベクトルの勾配に注目し，敵対的サンプルの生成は幅優先探索を行っているため，計算量が高い．また，我々の摂動は敵対的な摂動を連続値のまま扱っているが，文献 [3] は離散的なシンボルに変換して学習に加えている点も異なる．

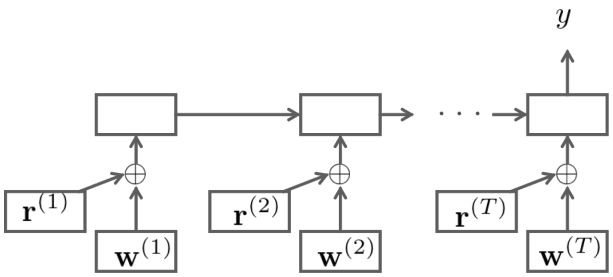

図 2: 敵対学習を LSTM に適応した図.

Method	Test error rate
Baseline	7.05 (%)
Adversarial [11]	6.02 (%)
(Ours) Adversarial-Real (Uniq)	6.34 (%)
(Ours) Adversarial-Real (Sum)	**5.89 (%)**
Virtual Adversarial [11]	5.69 (%)
(Ours) Virtual Adversarial-Real (Uniq)	6.15 (%)
(Ours) Virtual Adversarial-Real (Sum)	**5.66 (%)**
Full+Unlabeled+BoW [10]	11.11 (%)
Paragraph Vectors [9]	7.42 (%)
SA-LSTM [2]	7.24 (%)
One-hot bi-LSTM [8]	5.94 (%)

表 1: 実験結果 (IMDB における極性分類タスク).

最も高い方向ベクトルを $\boldsymbol{d}_{\mathtt{most}}$ とする.

$$\epsilon = \frac{\boldsymbol{d}_{\mathtt{most}} \cdot \boldsymbol{r}^{(t)}_{\mathtt{advREAL}}}{\boldsymbol{d}_{\mathtt{most}} \cdot \boldsymbol{d}_{\mathtt{most}}} \tag{8}$$

以上のように (i)〜(iii) を考えることで，解釈可能な摂動を計算することができる．従来法 (式 3) と同様に，損失関数に提案法の摂動を加え，敵対学習に用いる．

4 実験

4.1 実験設定

文書分類において提案手法の有効性を検証した．データセットとして，英語の 2 クラス極性分類 (Positive/Negative) のデータセット IMDB[10] を用いた．訓練データ 21,246 文，開発データ 3,754 文，テストデータ 25,000 文，ラベルなしデータ 50,000 文である．先行研究 [11] に従い，前処理として小文字化はせず，語彙の作成は訓練データとラベルなしデータから構築し，語彙数は 86,935 となった．

同様に先行研究 [11] に従い，分類器として 1 層の単方向 LSTM[5] を用いた（図 2）．LSTM は事前にラベルなしデータで言語モデルを訓練した．開発データを用いて Early Stopping を行い，開発データで最も高い正解率を得たモデルをテストデータで評価した．

教師あり学習と半教師あり学習の 2 つの設定を試した．教師あり学習では Adversarial Training(Adv)[11] と，提案手法 (Adversarial-Real) を比較し，半教師あり学習では Virtual Adversarial Training(VAT)[12, 11] と，提案手法 (Virtual Adversarial-Real) を比較した．

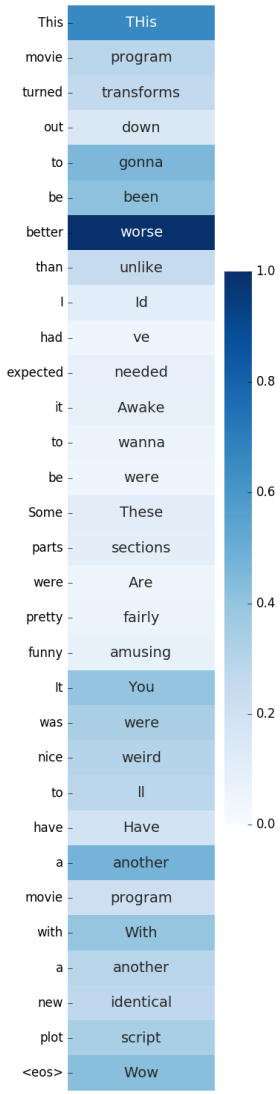

図 3: Positive → Negative

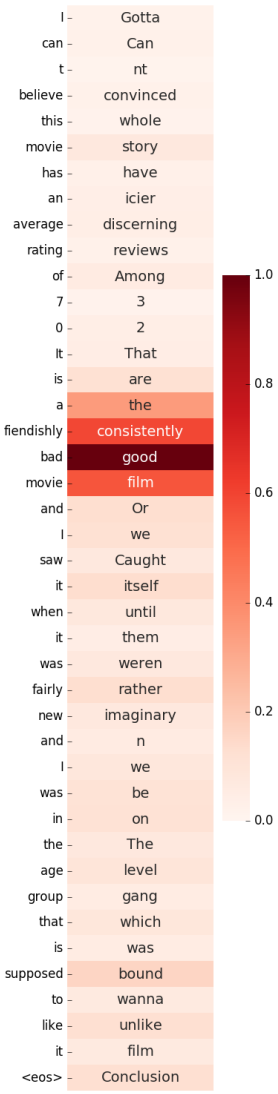

図 4: Negative → Positive

4.2 実験結果

表 1 に提案手法と既存手法の IMDB におけるエラー率 (%) を示す．2 つの提案手法 (Uniq, Sum) で Baseline を超える汎化性能を得ることができることができた．またデータ点を考慮しない既存手法よりも Adversarial では高い汎化性能を得ることができ，Virtual Adversarial-Real では同程度の性能を得ることができた．Adversarial 及び Virtual Adversarial の両方で Sum が高い汎化性能を得ることがわかった．より複雑なネットワーク構造を用いた先行研究の他の手法 [8] よりも良い結果を得ることができた．比較手法の Adversarial 及び Virtual Adversarial[11] の実験結果は，再実装し提案法と同じ条件で学習を行った結果である．本稿の実験結果は，文献中の報告結果より良いスコアを得られている．(Adv:6.21%, Vat:5.91%)

	文	分類器の予測
実際のテストデータ文	This movie turned out to be better than I had expected it to be Some parts were pretty funny It was nice to have a movie with a new plot <eos>	Positive
敵対的サンプル	This movie turned out to be worse than I had expected it to be Some parts were pretty funny It was nice to have a movie with a new plot <eos>	Negative
実際のテストデータ文	There is really but one thing to say about this sorry movie It should never have been made The first one one of my favourites An American Werewolf in London is a great movie with a good plot good actors and good FX But this one It stinks to heaven with a cry of helplessness <eos>	Negative
敵対的サンプル	There is really but one thing to say about that sorry movie It should never have been made The first one one of my favourites An American Werewolf in London is a great movie with a good plot good actors and good FX But this one It stinks to heaven with a cry of helplessness <eos>	Positive

表 2:　学習済み分類器の予測を誤らせる入力文 (敵対的サンプル) の生成例．上部の例は better→worse への置き換え，this→that への置き換えによって予測結果が反転している．

5　モデルの分析

5.1　摂動の可視化

図 3，4 に摂動ベクトルを可視化した．この可視化には，提案手法の Uniq を用いている．二列の単語は，左列がテストデータの実際の文であり，右列の単語は提案手法 (Uniq) の最大 $\boldsymbol{\alpha}_k$ となる周辺単語である．色（青または赤）の濃淡は提案法により計算された摂動ベクトルの大きさであり，それぞれの単語へ置き換えるた際に，分類器が間違え易いかどうかの度合いを示していると解釈できる．ただし，濃淡は文全体で正規化した値である．図 3 は，正解が Positive クラスの文であり，文全体で better→worse に置き換えると分類器が間違えやすいことを捉えている．同様に図 4 は Negative クラスの文であり，最も大きい摂動ベクトルの方向は bad→good の置き換え方向であることが分かる．

5.2　敵対的な入力文の作成

提案法では，前説で示したように，どの単語を別のどの単語に置き換えると分類器が誤認識しやすいか，というのを計算することができる．これを応用し，人間が解釈しやすい実際の単語へ置き換えた擬似データを作成することも可能である．

従来法では，摂動方向と単語ベクトル方向は無関係に計算されているため，単語ベクトルに摂動を加えたデータ点が，どのような単語に対応しているのかを解釈することができない．一方で，提案法は，実在する単語ベクトルの方向に摂動を限定することで，摂動方向と単語の置き換えを対応付けることができるようになった．

その性質を利用して，テスト文の単語を，最も強い摂動の単語へ置き換えることで，元文の予測結果を間違える敵対的な入力文を作成することを考える．表 2 に実際に作成した敵対的サンプルの例を示す．ここで用いたテストデータは，学習済み分類器 (Baseline) が正しく分類できた文の中からサンプリングした．

今回の事例から，分類器が間違える入力には大きく分けて 2 種類あることが分かる．

1. 文の意味が変わり，予測が反転する (表 2 上部)．
2. 文の意味は変わらず，予測が反転する (表 2 下部)．

このように分類器が間違える敵対的サンプルを生成することで，分類器の性質を知ることができる．

また本稿の実験では単語ベクトルを入力としているが，文字ベクトルを用いるモデルに適応することで，文字レベルの敵対的な入力（タイポ）を生成することも可能になる．この応用先としてエラー訂正タスクで間違えやすいデータの生成や，起こりやすいタイポ分析への応用に用いることができる．

6　おわりに

本研究では，自然言語処理における分類器を騙す摂動を加える際に，摂動を実データ方向に限定する手法を提案した．提案手法を用いることで，人間に解釈可能な摂動を計算することができることを示した．さらに，敵対的な文を生成することで，分類器の性質を人間に解釈可能となる．本稿では，最初の取り組みとして，簡単な二クラス分類タスクで提案法の有効性を検証したが，エンコーダデコーダモデル [14] への拡張も，生成側の入力を間違えるクロスエントロピーロスを最大化するような摂動を考えることで，任意の出力をするような入力文の生成が行える可能性もある．今後は，様々なタスク，モデルへの適用を考えていきたい．

参考文献

[1] Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. *CoRR*, abs/1711.02173, 2017.

[2] Andrew M. Dai and Quoc V. Le. Semi-supervised sequence learning. In *NIPS*, 2015.

[3] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for nlp. *arXiv preprint arXiv:1712.06751*, 2017.

[4] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014.

[5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[6] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. Deceiving google's perspective api built for detecting toxic comments. *CoRR*, abs/1702.08138, 2017.

[7] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *EMNLP*, 2017.

[8] Rie Johnson and Tong Zhang. Supervised and semi-supervised text categorization using lstm for region embeddings. In *ICML*, 2016.

[9] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML*, 2014.

[10] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *ACL*, 2011.

[11] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification. *ICLR*, 2016.

[12] Takeru Miyato, Shin ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. Distributional smoothing with virtual adversarial training. 2015.

[13] Suranjana Samanta and Sameep Mehta. Towards crafting text adversarial samples. *arXiv preprint arXiv:1707.02812*, 2017.

[14] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[15] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013.