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], 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 (NAIST) 松本研究室

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 革新知能統合研究センターAIP

2018/3/14 NLP2018

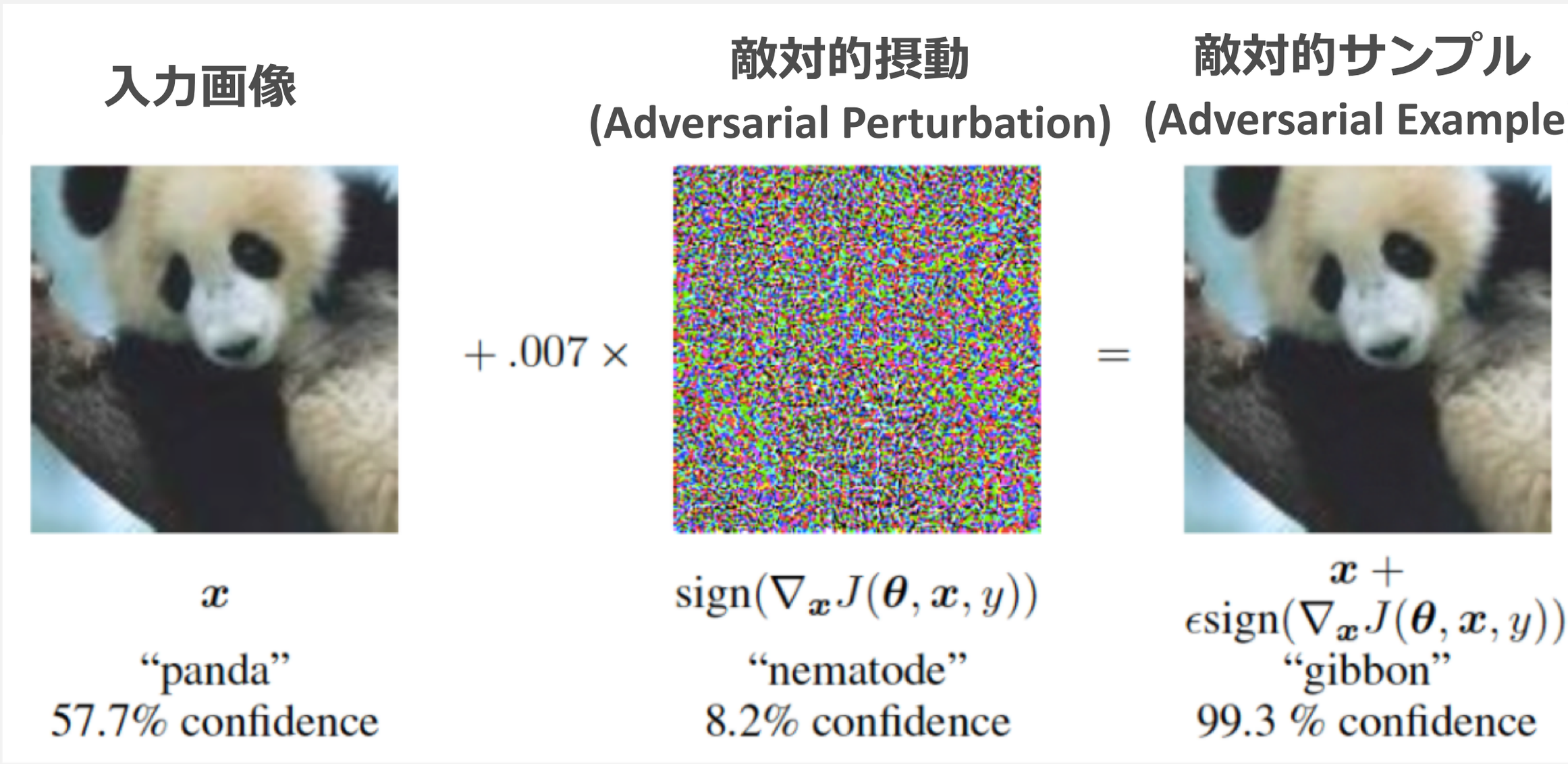
発表の概要

1. 背景・目的
   - 敵対的摂動・敵対的サンプル
   - 自然言語処理における敵対的摂動
2. 関連研究
3. 提案手法
4. 実験
5. 提案手法の分析
6. まとめと今後の課題

敵対的摂動

- **画像**に**摂動(ノイズ)**を加えると**分類器が間違える**ことが知られている

[Szegedy *et al.*, 2014, Goodfellow *et al* .,2015]



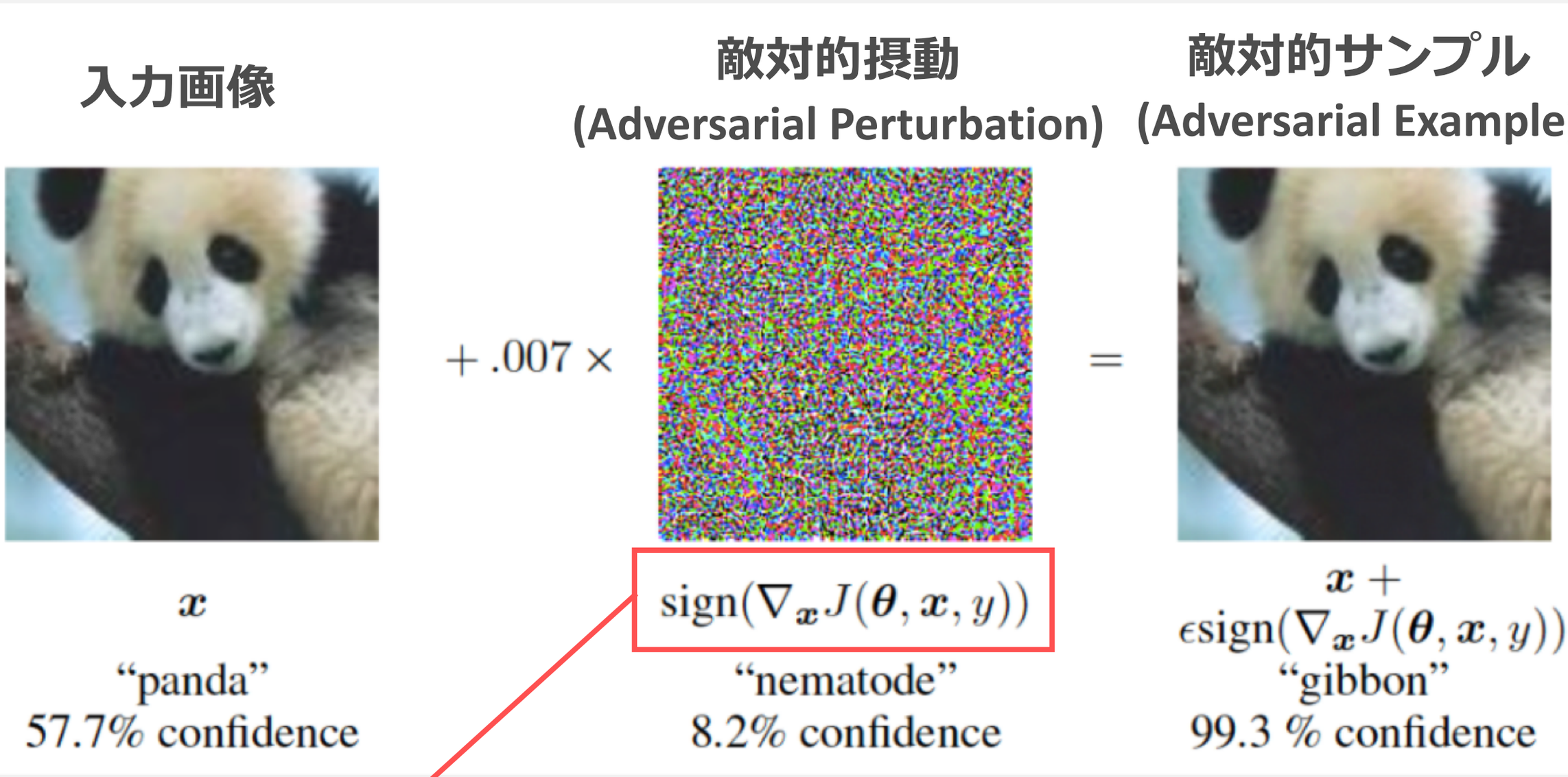
[Goodfellow *et al.*, 2015]から引用

- [Szegedy *et al.*, 2014] : “Intriguing properties of neural networks.”, ICLR 2014.
- [Goodfellow *et al.*,2015]:“Explaining and Harnessing Adversarial Examples”, ICLR 2015.

敵対的摂動

- **画像**に**摂動(ノイズ)**を加えると**分類器が間違える**ことが知られている

[Szegedy *et al.*, 2014, Goodfellow *et al* .,2015]



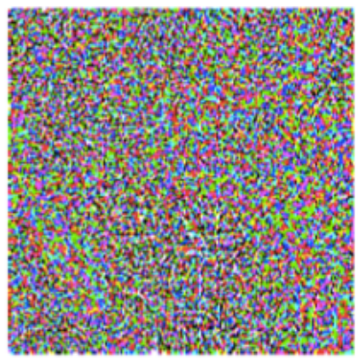
[Goodfellow *et al.*, 2015]から引用

目的関数の誤差が増大する**勾配**から求める

画像処理 と 自然言語処理

- **画像処理**
  - 入力は**連続的**（RGB値 0～255）
  - **画像　＋ 摂動ベクトル　→　画像**　として**解釈可能**

（例）

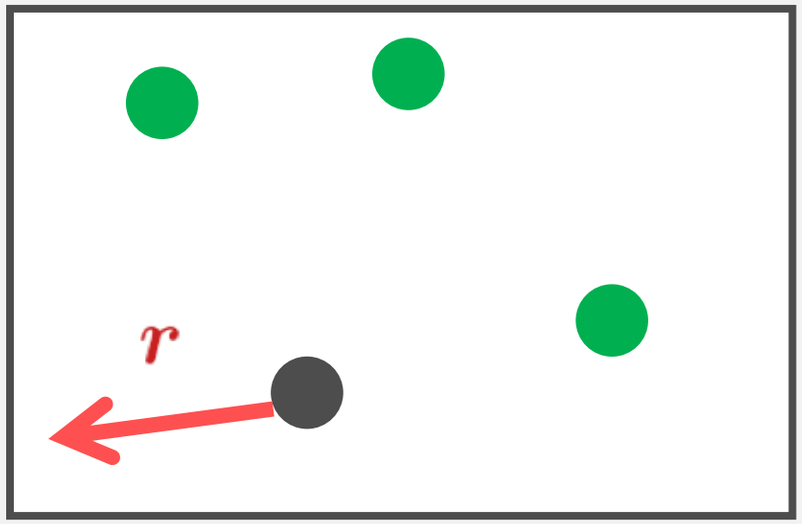




- **自然言語処理**
  - 入力は**離散的**（単語）
  - **離散シンボル⇔単語ベクトル**を変換するLookup Tableを用いる
  - **単語ベクトル ＋ 摂動ベクトル** → **？ ？ ？**

（どの単語を表しているか**解釈できない**）

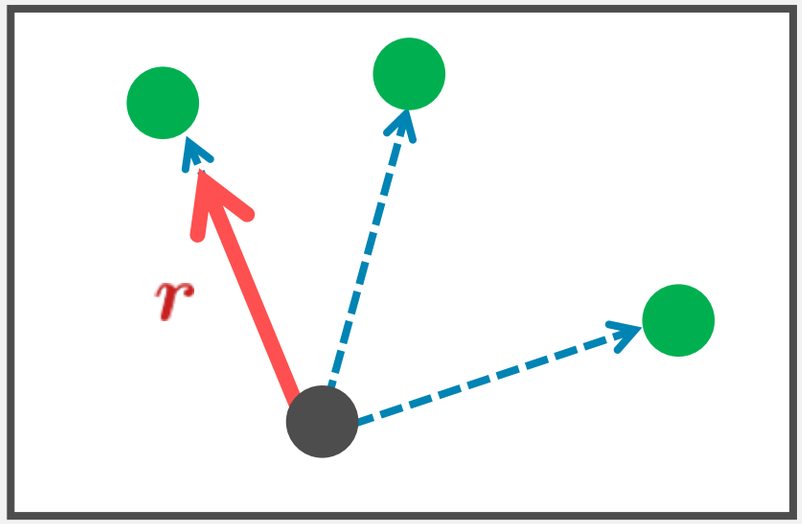
本研究の概要

既存手法

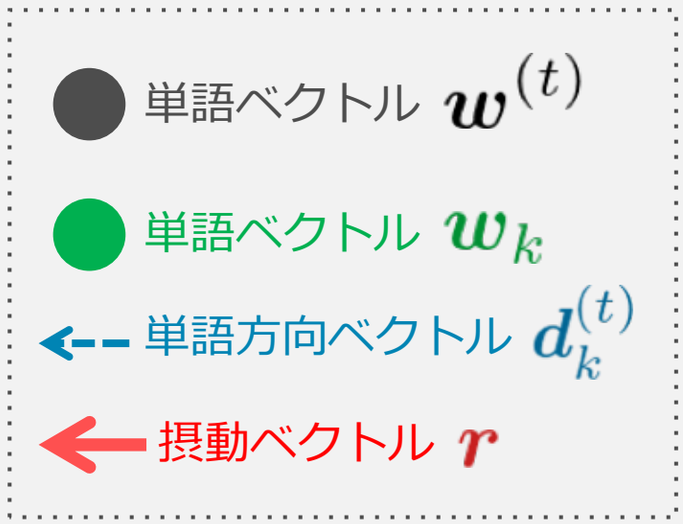


摂動ベクトルが
単語が存在する点を
向いていない

提案手法



摂動ベクトルが
単語が存在する点を
向いている



提案手法のメリット

1. **摂動を可視化し人間に解釈可能になる**
   (どのような単語の置き換えなのか)

2. **敵対的サンプルを勾配から生成することができる**

発表の概要

1. 背景・目的
   - 敵対的摂動・敵対的サンプル
   - 自然言語処理における敵対的摂動
2. **関連研究**
3. 提案手法
4. 実験
5. 提案手法の分析
6. まとめと今後の課題

関連研究

敵対的サンプル(Adversarial Example) for NLP

- 出力を変える入力文の作成
  - **クラウドソーシング**で**読解システム**を騙す入力文を作成する
    [Jia and Liang, 2017]
  - ランダムな**文字のスワップ**を考えて**NMT**を騙す入力文を**探索**する
    [Belinkov and Bisk, 2018]
  - **同義語**を置き換え大量の入力文を生成し,分類器を騙す出力文を**探索**する
    [Samanta and Mehta, 2017]
- **モデルの挙動**を知ることで**解釈性**が上がる.

関連研究

敵対的学習(Adversarial Training)

- **敵対的サンプル**を学習に加えて**汎化性能**を上げる[Goodfellow *et al* .,2015]

$$\tilde{J}(\boldsymbol{\theta}, \boldsymbol{x}, y) = \alpha J(\boldsymbol{\theta}, \boldsymbol{x}, y) + (1-\alpha) J(\boldsymbol{\theta}, \boldsymbol{x} + \epsilon \text{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)\right))$$

目的関数　　**敵対的サンプルを正しく分類する目的関数**

- **半教師あり学習**に敵対的学習を拡張 (Virtual Adversarial Training; VAT)

  [Miyato *et al.*, 2016]

Adversarial Training for Text

- **単語ベクトル**に摂動を加え，**敵対的学習**を行う **[Miyato *et al.*, 2017]**
  - テキスト分類において最高精度だが, 摂動に関する**解釈性**は議論していない

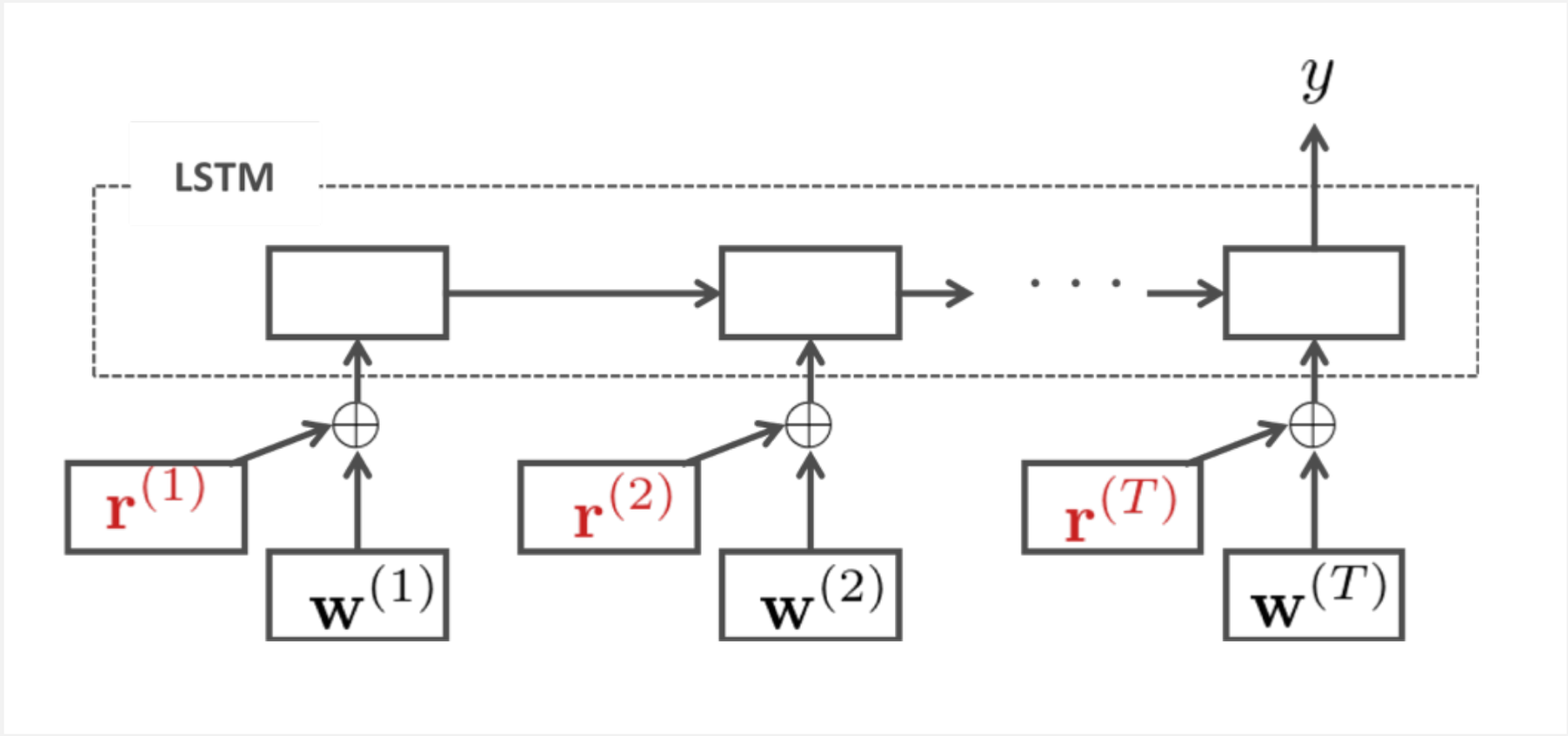
既存手法： [Miyato *et al.*, 2017]　について詳しく述べる

既存手法：[Miyato *et al.*, 2017]

- **Takeru Miyato**, Andrew M Dai, and Ian Goodfellow, **ICLR 2017**

  “Adversarial training methods for semi-supervised text classification.”

- **単層LSTM** + **Pre-Training** (Language Model) + **Adversarial Training**



敵対的摂動ベクトル：$\boldsymbol{r}^{(t)}$

単語ベクトル： $\boldsymbol{w}^{(t)}$

既存手法：[Miyato *et al.*, 2017]

敵対的摂動ベクトル：$\boldsymbol{r}$

単語ベクトル：$\boldsymbol{w}^{(t)}$

摂動の定義

ϵ：ハイパーパラメータ　(例: 1.0)

$$\tilde{X}_{+\boldsymbol{r}} = \left(\boldsymbol{w}^{(t)} + \boldsymbol{r}^{(t)}\right)_{t=1}^{T}$$

：**単語ベクトル**に**摂動**を加えた入力

$$\boldsymbol{r}_{\mathtt{AdvT}} = \underset{\boldsymbol{r},||\boldsymbol{r}||\leq\epsilon}{\operatorname{argmax}} \left\{\ell(\tilde{X}_{+\boldsymbol{r}}, \tilde{Y}, \mathcal{W})\right\}$$

：**損失関数 ℓ を増大させる$\boldsymbol{r}$を求める**

摂動の求め方

$$\boldsymbol{r}_{\mathtt{AdvT}}^{(t)} = \frac{\epsilon \boldsymbol{g}^{(t)}}{||\boldsymbol{g}||_2}, \quad \boldsymbol{g}^{(t)} = \nabla_{\boldsymbol{w}^{(t)}} \ell(\tilde{X}, \tilde{Y}, \mathcal{W})$$

：**勾配を求め, L2正規化**

敵対的学習

$$\mathcal{J}_{\mathtt{AdvT}}(\mathcal{D}, \mathcal{W}) = \frac{1}{|\mathcal{D}|} \sum_{(\tilde{X},\tilde{Y})\in\mathcal{D}} \ell(\tilde{X}_{+\boldsymbol{r}_{\mathtt{AdvT}}}, \tilde{Y}, \mathcal{W})$$

：**摂動を加えた入力を**
正しく分類する目的関数

$$\hat{\mathcal{W}} = \underset{\mathcal{W}}{\operatorname{argmin}} \left\{\mathcal{J}(\mathcal{D}, \mathcal{W}) + \lambda \mathcal{J}_{\mathtt{AdvT}}(\mathcal{D}, \mathcal{W})\right\}$$

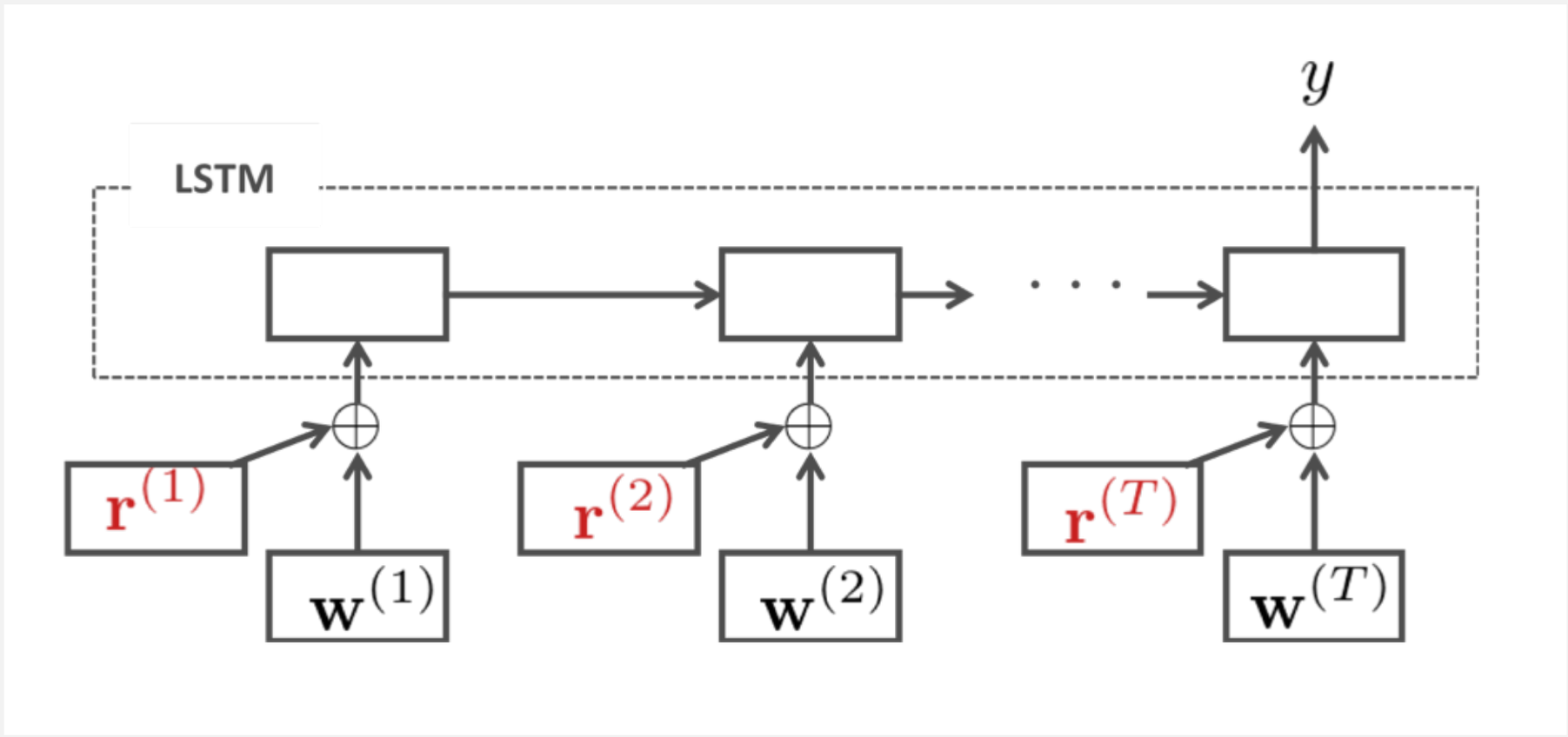
：**目的関数に追加**

既存手法：[Miyato *et al.*, 2017]

- **Takeru Miyato**, Andrew M Dai, and Ian Goodfellow, **ICLR 2017**

  “Adversarial training methods for semi-supervised text classification.”

- **単層LSTM** + **Pre-Training** (Language Model) + **Adversarial Training**



敵対的摂動ベクトル：$r^{(t)}$

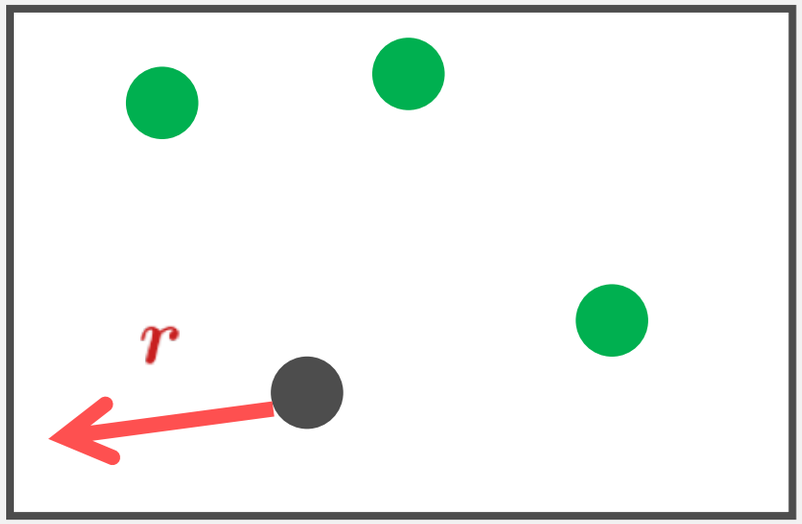
単語ベクトル： $w^{(t)}$

発表の概要

1. 背景・目的
   - 敵対的摂動・敵対的サンプル
   - 自然言語処理における敵対的摂動
2. 関連研究
3. **提案手法**
4. 実験
5. 提案手法の分析
6. まとめと今後の課題

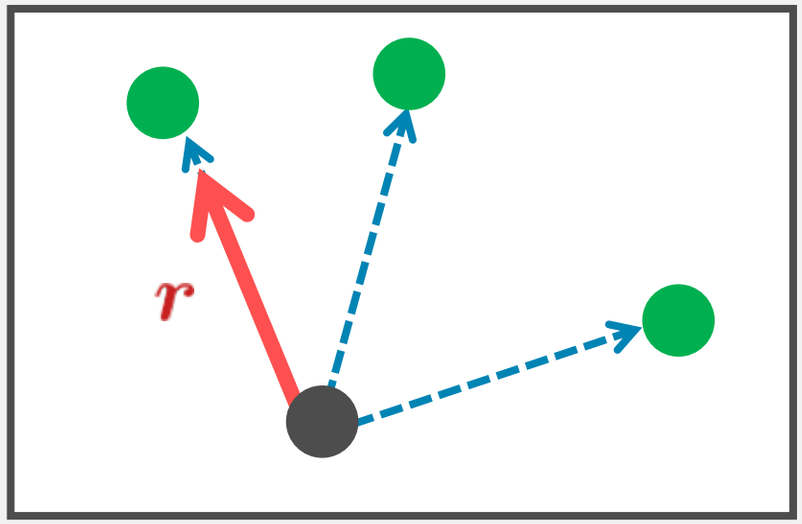
提案手法

既存手法:

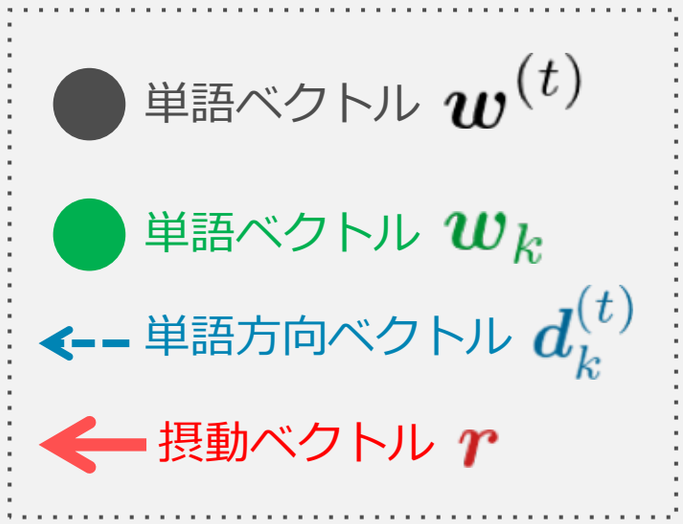


摂動ベクトルが
単語が存在する点を
向いていない

提案手法:



摂動ベクトルが
単語が存在する点を
向いている



単語方向ベクトル $\boldsymbol{d}_k^{(t)}$ を考慮し, **勾配**から**摂動**を求める

$$\boldsymbol{d}_k^{(t)} = \boldsymbol{w}_k - \boldsymbol{w}^{(t)}$$

提案手法

単語ベクトル： $\boldsymbol{w}^{(t)}$

摂動の定義

$$\boldsymbol{d}_k^{(t)} = \frac{\tilde{\boldsymbol{d}}_k^{(t)}}{||\tilde{\boldsymbol{d}}_k^{(t)}||_2}, \quad \text{where} \quad \tilde{\boldsymbol{d}}_k^{(t)} = \boldsymbol{w}_k - \boldsymbol{w}^{(t)}$$ ：単語の**方向ベクトル**

$$\boldsymbol{r}(\boldsymbol{\alpha}^{(t)}) = \sum_{k=1}^{|\mathcal{V}|} \alpha_k^{(t)} \boldsymbol{d}_k^{(t)}$$ ：**重みスコアαとの総和を摂動とする**

$$\tilde{X}_{+\boldsymbol{r}(\boldsymbol{\alpha})} = \left(\boldsymbol{w}^{(t)} + \boldsymbol{r}(\boldsymbol{\alpha}^{(t)})\right)_{t=1}^{T}$$ ：**単語ベクトル**に**摂動**を加えた入力

$$\boldsymbol{\alpha}_{\texttt{iAdvT}} = \underset{\boldsymbol{\alpha}, ||\boldsymbol{\alpha}|| \leq \epsilon}{\operatorname{argmax}} \left\{ \ell(\tilde{X}_{+\boldsymbol{r}(\boldsymbol{\alpha})}, \tilde{Y}, \mathcal{W}) \right\}$$ ：**損失関数 ℓ を増大させるaを求める**

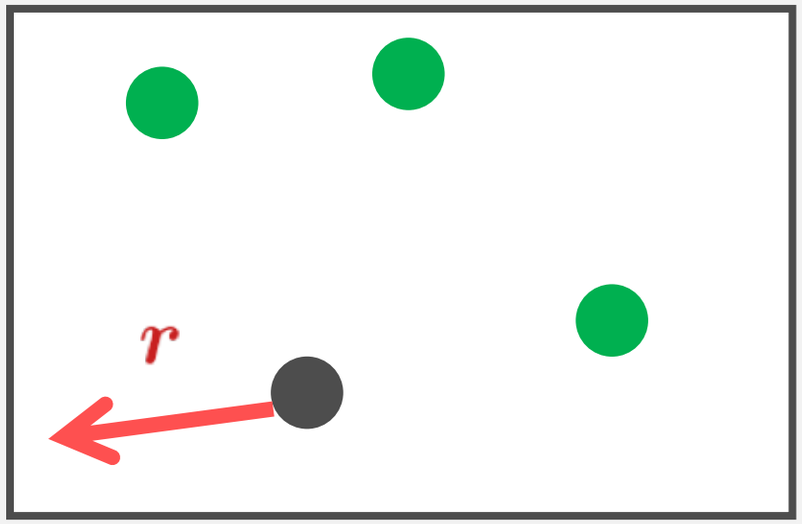
摂動の求め方

$$\boldsymbol{\alpha}_{\texttt{iAdvT}}^{(t)} = \frac{\epsilon \boldsymbol{g}^{(t)}}{||\boldsymbol{g}||_2}, \quad \boldsymbol{g}^{(t)} = \nabla_{\boldsymbol{\alpha}^{(t)}} \ell(\tilde{X}_{+\boldsymbol{r}(\boldsymbol{\alpha})}, \tilde{Y}, \mathcal{W})$$ ：**勾配**から**α**を求める

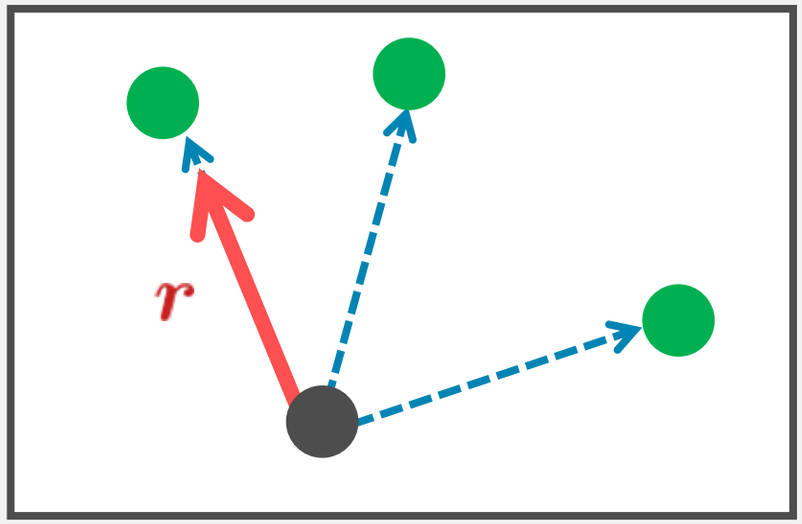
$$\boldsymbol{r}(\boldsymbol{\alpha}^{(t)}) = \sum_{k=1}^{|\mathcal{V}|} \alpha_k^{(t)} \boldsymbol{d}_k^{(t)}$$ ：**重みα**と**単語方向ベクトルd**から**摂動**を求める

提案手法

既存手法:



提案手法:



単語ベクトル $\boldsymbol{w}^{(t)}$

単語ベクトル $\boldsymbol{w}_k$

単語方向ベクトル $\boldsymbol{d}_k^{(t)}$

摂動ベクトル $\boldsymbol{r}$

$$\boldsymbol{r}(\boldsymbol{\alpha}^{(t)}) = \sum_{k=1}^{|\mathcal{V}|} \alpha_k^{(t)} \boldsymbol{d}_k^{(t)}$$

- **可視化　　：最大のa 以外は0にフィルターすることで唯一の方向を向くようになる**
- **敵対的学習：上記の式を使う.**

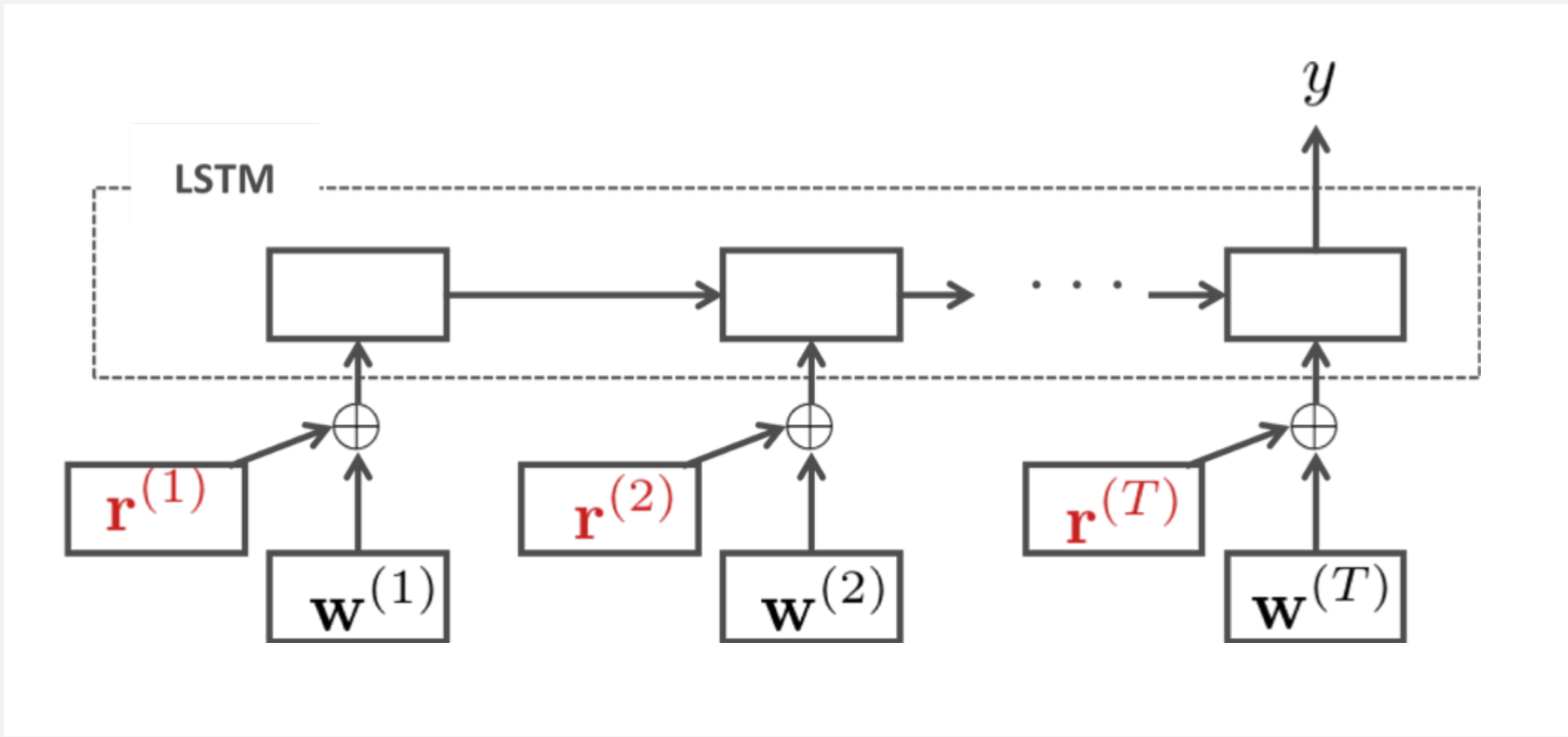
発表の概要

1. 背景・目的
   - 敵対的摂動・敵対的サンプル
   - 自然言語処理における敵対的摂動
2. 関連研究
3. 提案手法
4. **実験**
5. 提案手法の分析
6. まとめと今後の課題

実験

- **敵対的学習の実験** (汎化性能が上がるかを調べる)
  - 比較手法：[Miyato *et al.*, 2017]
  - データセット: **IMDB (極性分類タスク)**

    (Train : 21,246　Dev: 3,754　Test: 25,000　Unlabeled: 50,000)



敵対的摂動ベクトル：$\boldsymbol{r}^{(t)}$

単語ベクトル：　$\boldsymbol{w}^{(t)}$

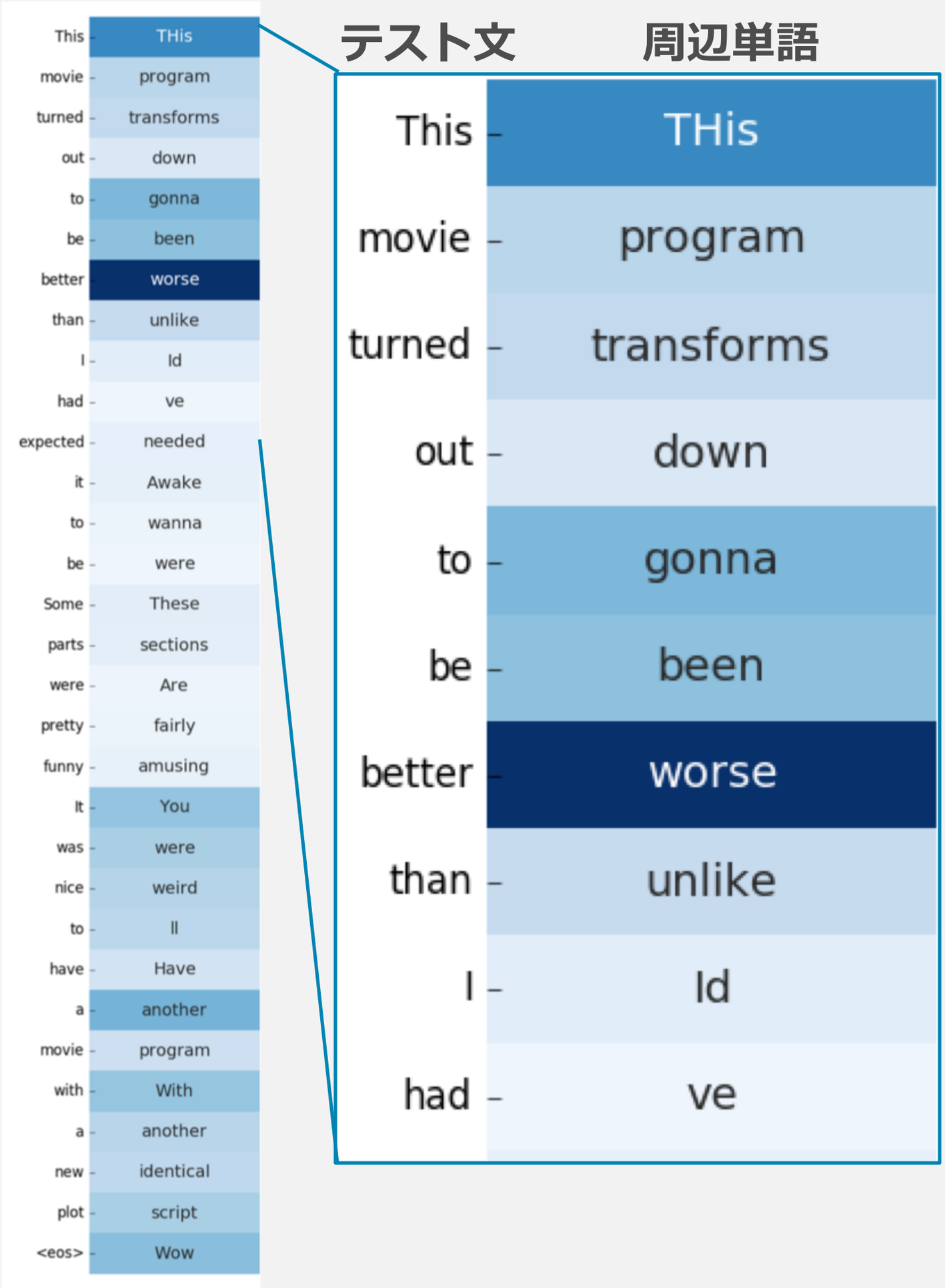
実験結果

Method (Semi-supervised : †)	Test error rate
Baseline	7.05 (%)
Random Perturbation (Labeled)	6.74 (%)
AdvT-Text [Miyato *et al.*, 2017]	6.12 (%)
iAdvT-Text (Ours)	**6.08 (%)**
Random Perturbation (Labeled + Unlabeled)†	6.44 (%)
VAT-Text [Miyato *et al.*, 2017]†	5.69 (%)
iVAT-Text (Ours)†	**5.66 (%)**
Full+Unlabeled+BoW [Maas *et al.*, 2011]	11.11 (%)
Paragraph Vectors [Le and Mikolov, 2014]	7.42 (%)
SA-LSTM [Dai and Le, 2015]†	7.24 (%)
One-hot bi-LSTM [Johnson and Zhang, 2016]†	5.94 (%)

- 既存手法と同等 or 少し良い性能を得ることができた.
- ランダムな摂動ベクトルよりも高い性能を得ることができた

発表の概要

1. 背景・目的
   - 敵対的摂動・敵対的サンプル
   - 自然言語処理における敵対的摂動
2. 関連研究
3. 提案手法
4. 実験
5. **提案手法の分析**
6. まとめと今後の課題



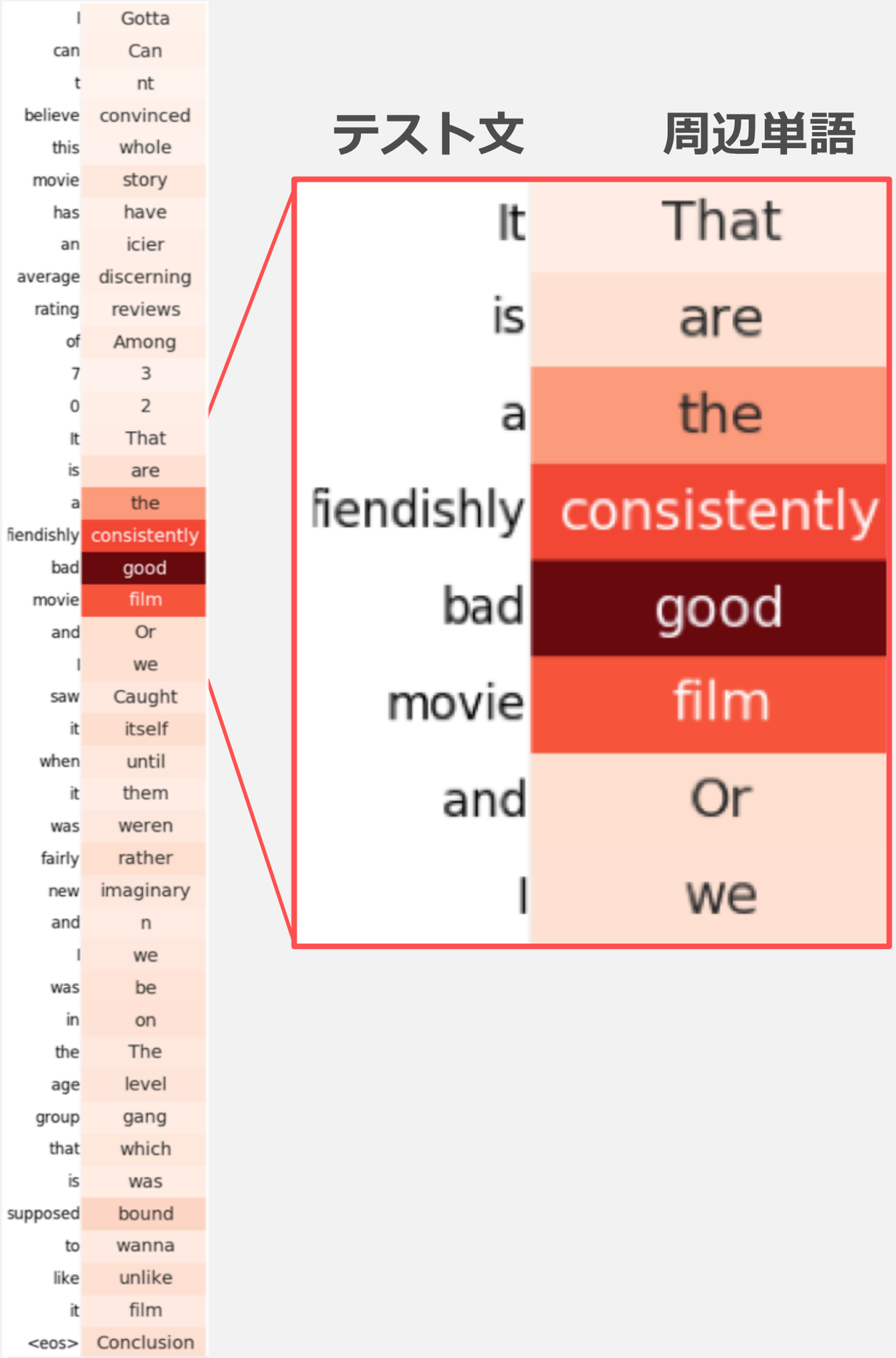
分類器の予測が
Positive → **Negative**となる摂動を求める.

摂動の**方向**と**大きさ**を可視化した.
方向　：どの単語が存在するか
大きさ：摂動のL2ノルム
※摂動の重みは最大値を用いた

左軸： 入力文（テストデータ）
Positiveな文

右軸：摂動の方向に存在する単語
（単語の置き換え）

“**better**” → “**worse**”
と単語ベクトルに摂動を加えると
Positive → **Negative**となりやすいことが可視化で分かる



分類器の予測が
Negative → **Positive**となる摂動を求める.

摂動の**方向**と**大きさ**を可視化した.
方向　：どの単語が存在するか
大きさ：摂動のL2ノルム
※摂動の重みは最大値を用いた

左軸： 入力文（テストデータ）
Negative な文

右軸：摂動の方向に存在する単語
（単語の置き換え）

“**bad**”→ “**good**”と単語ベクトルに摂動を加えると
Negative→**Positive**となりやすいことが可視化で分かる

予測を変化させる入力文の作成

- **摂動のノルム**が大きい単語を置き換えて，
  分類器の予測結果が変わるか確かめる．

テストデータ文

予測結果: Positive

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>

敵対的サンプル

予測結果: Negative

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>

“**better**” → “**worse**”と置き換えると予測結果が反転した
（**文の意味も変化している**）

予測を変化させる入力文の作成

テストデータ文

予測結果: Negative

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>

敵対的サンプル

予測結果: Positive

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>

“**this**” → “**that**”と置き換えると予測結果が反転した
（**文の意味は変化していない**）

予測を変化させる入力文の作成

“**better**” → “**worse**”と置き換えると予測結果が反転した
（**文の意味も変化している**）

“**this**” → “**that**”と置き換えると予測結果が反転した
（**文の意味は変化していない**）

このような入力を人手や辞書のコストを掛けず，
勾配情報から求めることができるのが利点

発表の概要

1. 背景・目的
   - 敵対的摂動・敵対的サンプル
   - 自然言語処理における敵対的摂動
2. 関連研究
3. 提案手法
4. 実験
5. 提案手法の分析
6. **まとめ**

まとめ

- **単語ベクトル**に対して**摂動**を加える際に
  単語が存在する方向に制約を加える手法を提案した.
- 既存の手法に比べ，同等程度の性能を得た
- **可視化**することでモデルの解釈性が上がった
- **敵対的な入力文**を**勾配**から求めることを示した.