解釈性向上のための注意機構と損失勾配に対する関連損失の導入

北田 俊輔, 彌冨 仁
shunsuke.kitada.8y@stu.hosei.ac.jp
法政大学 HOSEI University

Background

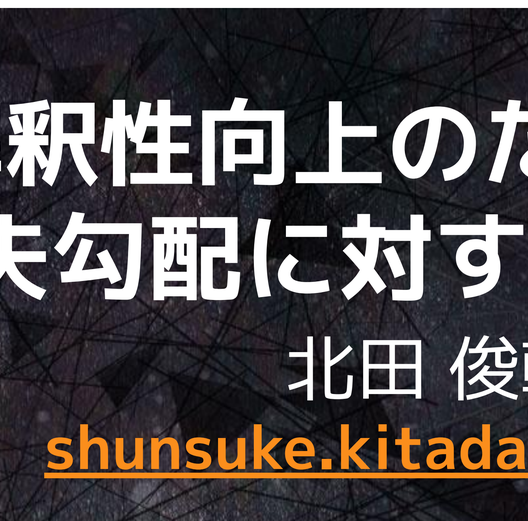
自然言語処理における注意機構 (attention)

入力される複数のベクトルに対して、どのベクトルを重要視するかの重み付け

- 予測精度の向上に寄与
- 予測根拠の提示に利用

予測根拠の提示手法

損失計算時の勾配 (gradient) を元に根拠を提示 [Ross+, AAAI17]



注意機構と勾配の関係

- 注意機構における単語の重要度
- 損失勾配ベースの単語の重要度

注意機構の重みは勾配ベースのスコアと関連すべき
➜ 必ずしも関連するとは限らない [Jain and Wallance, NAACL19]

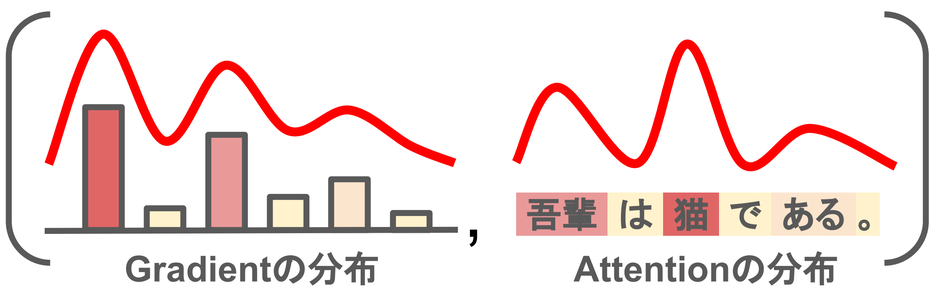
注意機構の分布と損失勾配を基にした値の分布を
関連するよう学習する**関連損失**を新たに導入

Method

関連損失 $\mathcal{L}_r$

双方向LSTMでエンコードして得られる隠れ層 $\mathbf{h}$ の勾配 $\mathbf{g_h}$ の分布と注意機構 $\boldsymbol{\alpha}$ の分布の距離を最小化

$$\mathcal{L}_r$$



訓練時: 通常の損失に加えて提案手法の **関連損失** を導入

$$\mathcal{L}_{\text{all}} = \mathcal{L}\left(\mathbf{x}^{(n)}, y^{(n)}\right) + \lambda\mathcal{L}_r\left(\sigma\left(\mathbf{g}_h^{(n)}\right), \boldsymbol{\alpha}^{(n)}\right)$$

λ : 関連損失の効果をコントロールするハイパーパラメータ

隠れ層の勾配$\mathbf{g_h}$を活性化関数σに通して確率値に変換

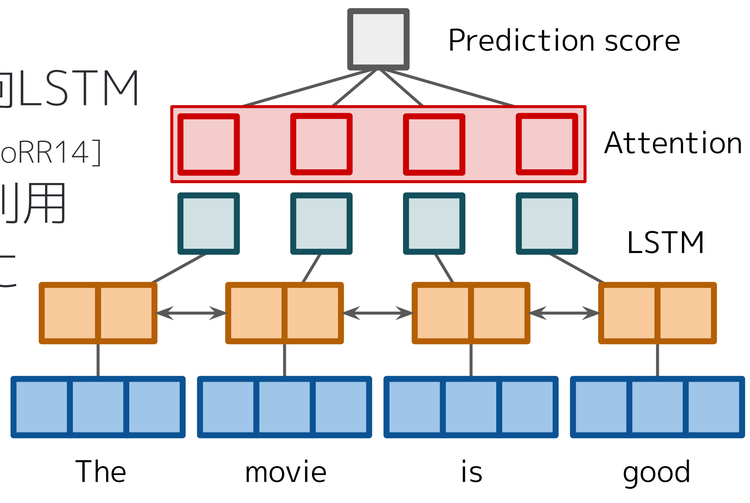
$$\mathbf{g}_h^{(n)} = \frac{\partial y^{(n)}}{\partial \mathbf{x}^{(n)}}\mathbf{h}^{(n)}, \quad \sigma\left(\mathbf{g}_h^{(n)}\right) = \text{softmax}\left(|\mathbf{g}_h^{(n)}|\right)$$

注意機構の分布と勾配の分布とのKL距離を最小化

$$\mathcal{L}_r\left(\sigma\left(\mathbf{g}_h^{(n)}\right), \boldsymbol{\alpha}^{(n)}\right) = \mathcal{D}_{\text{KL}}\left(\sigma\left(\mathbf{g}_h^{(n)}\right), \boldsymbol{\alpha}^{(n)}\right)$$

ベースモデル

- 注意つき1層 双方向LSTM 加法注意 [Bahdanau+, CoRR14] を注意機構として利用
- 埋込み層の初期化に学習済みfastText
- 最適化手法としてAdamを使用



Experiment

実験用データセット

2クラスになるよう前処理し、訓練 / 開発 / 評価に分割
使用したテキスト分類データセットは以下の3つ:

- 20 Newsgroups (News)
- Stanford Sentiment Treebank (SST)
- IMDB Movie Review Corpus (IMDB)

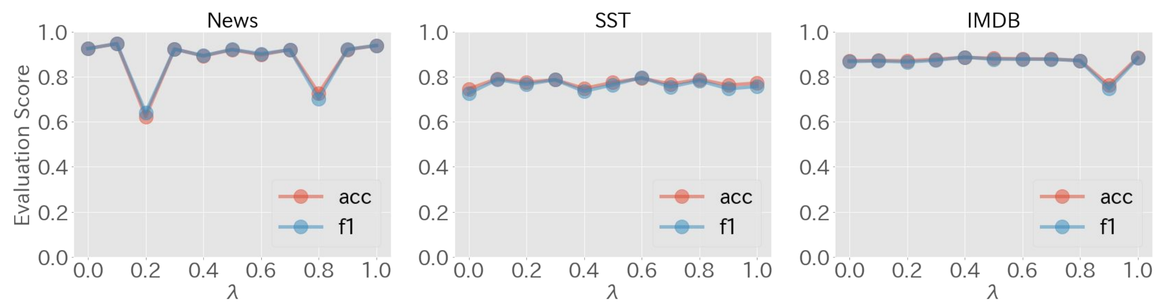
テキスト分類による精度比較

関連損失の有無による精度の違いを正解率とF1で評価

データセット	Accuracy [%]		F1 score [%]	
	関連損失 なし	関連損失 **あり**	関連損失 なし	関連損失 **あり**
News	92.51	**94.57**	92.39	**94.68**
SST	75.78	**79.13**	74.13	**78.75**
IMDB	87.10	**87.19**	86.70	**86.95**

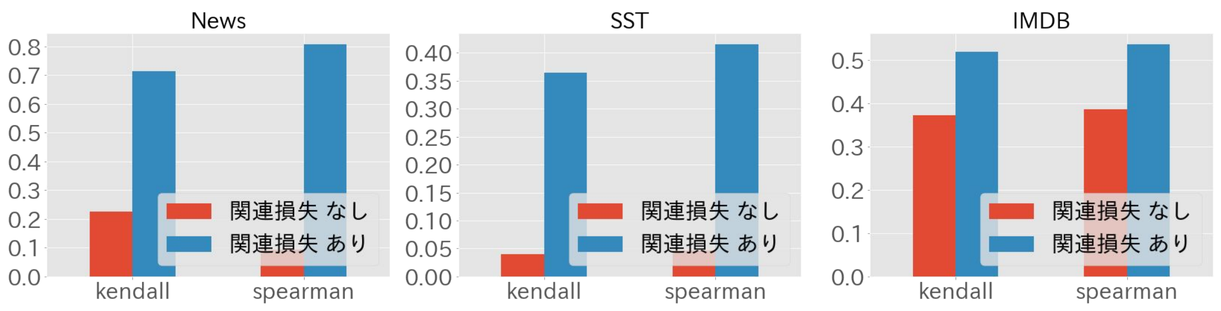
関連損失を導入したモデルはそれぞれのデータセットでベースラインを超えるスコアを達成

ハイパーパラメータの差異による精度比較



ハイパーパラメータλに関わらずほぼ一定の精度を達成

関連損失の有無による注意機構と勾配の関係



各順位相関係数において、注意機構と勾配がより相関

Conclusion & Future work

関連損失を新たに導入

注意機構と損失勾配を関連するように学習させることでベースラインを超える精度を達成

今後の展望

- 敵対的なattentionに対するモデルの頑健性の確認
- より解釈性の高い注意機構の可視化方法の検討

Reference

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "*Neural machine translation by jointly learning to align and translate.*" In CoRR2014.

Jain, Sarthak, and Byron C. Wallace. "*Attention is not Explanation.*" In NAACL2019

Ross, Andrew, Michael C. Hughes, and Finale Doshi-Velez. "*Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations.*" In AAAI2017