外部知識及び可視化の利用による マルチモーダルフェイクニュース検出の説明性改善

田邉耕太*1　野田五十樹*1　小山聡*1*2

*1北海道大学　　*2名古屋市立大学（2023年4月1日以降）

人工知能学会全国大会2023

背景・目的

・SNSの発達により**画像とテキストからなるフェイクニュース**が増加

・深層学習を用いたフェイクニュース検出の研究が注目されており，精度向上が進んでいる

・しかし，**フェイクニュース検出モデルの説明性**はあまり検討されていない

・**DBpedia**から得られる**外部知識**を活用するモデルを提案

・モデルに**Grad-CAM**，**Attention**の可視化を適用し，さらに外部知識を提示することによる説明性の改善・考察を行う

・提案モデルの性能比較も行う

利用手法

提案モデル

・VGG19，XLNetから得られた画像特徴量，テキスト特徴量を連結し，DNNに入力してフェイクニュースを検出する

・外部知識は，テキストにBERTによる固有表現抽出を適用し，DBpediaで検索をかけて得る．得られた外部知識はSEPトークンで連結し，テキストと共にXLNetに入力する（図1）

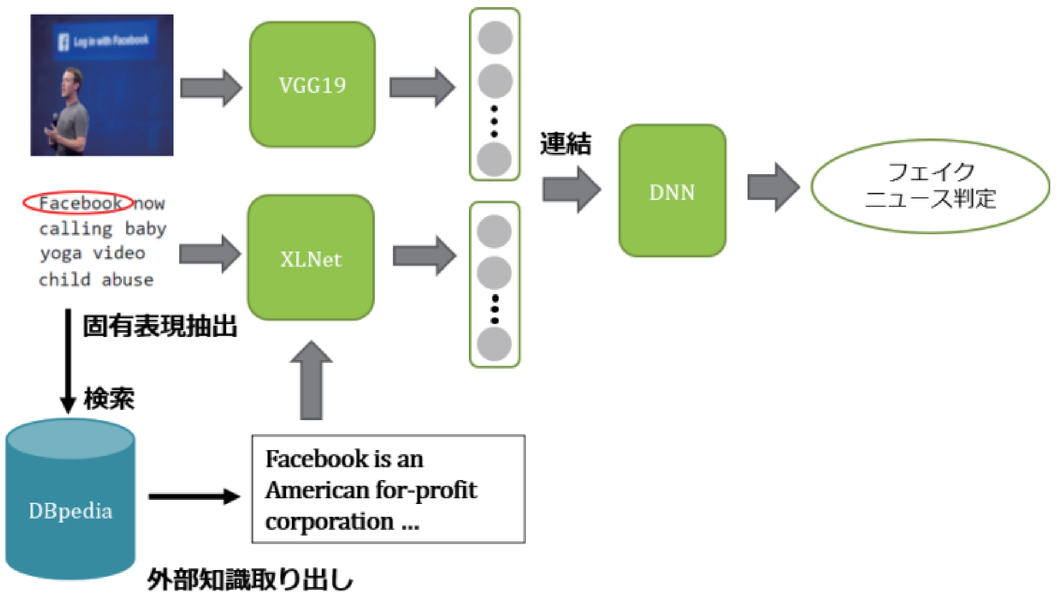


図1　モデルの全体概要

可視化手法

・Grad-CAMによりVGG19の注目領域をヒートマップで可視化する

・XLNetの注目部分は，最終層におけるCLSトークンの各トークンに対するAttention Weightが大きいほど強くハイライトされるように可視化する

実験・結果

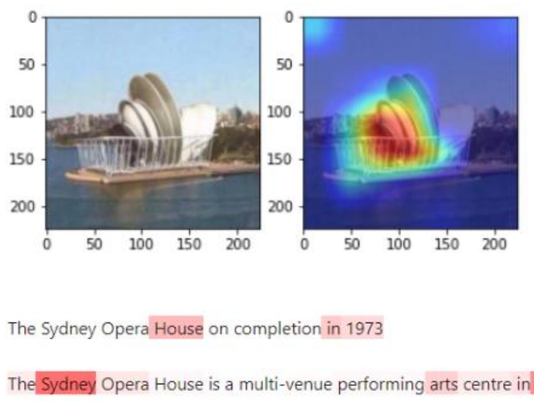
実験

・データセットにはFakedditを利用する．Fakedditは画像とテキストのペアが約53万枚含まれ，6種類のラベルが割り当てられている．画像の真意と関連しないテキストを持つFalse Connectionや，意図的に操作されたコンテンツであるManipulated Content等がある

・テキストは「title」と「clean_title」の2種類があり，小文字化を適用しているかという違いがある

・訓練データ，評価用データ，テストデータを8:1:1の比で分割し，モデルの学習は10epochで行った．

可視化結果

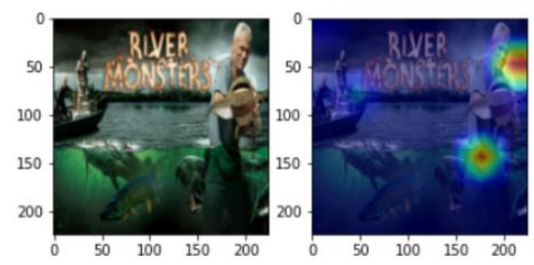


・画像とテキストで異なる意味を持つ部分に注目している

・”Sydney Opera House”の外部知識を提示できている

・外部知識にも注目している

図2　False Connectionのペアに対して正しく予測できた例．画像はオペラハウスに類似したオブジェであり，テキストはオペラハウスの説明である（1行目）．2行目はオペラハウスの外部知識である



・”Jeremy Wade”の外部知識を提示

・外部知識にも注目している

・**画像中の文字列に注目できていない**

図3　Manipulated Contentのペアに対してTrueと誤って予測した例．画像は「ジェレミー・ウェイドのRiver Monster」を表すが，テキストではReddit Monstersと事実と異なる内容が含まれる．2行目はジェレミー・ウェイドの外部知識である

性能比較

・clean_title，titleいずれの場合でも，外部知識の活用により性能が向上する結果となった

表1　モデルの性能比較

手法	Accuracy	Macro-F1 Score
ベースライン（clean_title）	0.792	0.605
ベースライン + 外部知識（clean_title）	0.794	0.618
ベースライン（title）	0.839	0.711
ベースライン + 外部知識（title）	0.843	0.714

まとめ

・DBpediaを活用した外部知識の提示により説明性の向上が期待できることを示した

・外部知識によりモデルの性能が向上した

・説明性が改善されたかに関して，アンケートを取るなど客観的な評価が必要

・モデルが画像中の文字列を活用できていない