シングルモーダル学習モデルとマルチモーダル学習モデルにおける判断根拠の分析

ER19021 岡島滉樹　　指導教授：藤吉弘亘

1. はじめに

深層学習によるマルチモーダル学習モデルは，複数のモーダルを用いて学習し，分類する処理を行う．各モーダルの特徴を捉えることで，精度向上することが知られている．しかし，画像とテキストを用いたマルチモーダル学習モデルは，シングルモーダル学習モデルと比べて，テキストがどのように画像認識モデルに作用しているのか明らかではない．そこで本研究では，シングルモーダル学習モデルとマルチモーダル学習モデルにおける判断根拠を可視化し，モーダルを追加することによる判断根拠への影響について分析する．

2. マルチモーダル学習

モーダルとは，テキスト・音声・静止画・動画などの入力される情報のことである．シングルモーダルとは，画像，テキスト，音声などの 1 つの情報のことである．また，1 つの情報のみを用いて学習して，分類するモデルをシングルモーダル学習モデルと呼ぶ．一方，マルチモーダルとは，複数の情報のことであり，複数のモーダルを用いて学習して，分類するモデルをマルチモーダル学習モデルと呼ぶ．マルチモーダル学習モデルは，複数の情報を考慮して問題を解くことができるため，シングルモーダル学習モデルと比べて高い認識性能を発揮することが知られている．代表的な手法の 1 つに Gallo らの Early fusion モデル [1] がある．Early fusion モデルのネットワーク構造を図 1 に示す．Early fusion モデルは，画像とテキストを入力データとし，各モーダル用のモデルで抽出した特徴量を連結して分類器に入力することで，最終予測を出力する．

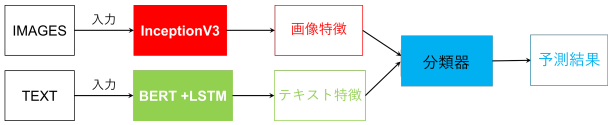


図 1：Early fusion のネットワーク構造

3. マルチモーダル学習モデルの判断根拠の分析

Early fusion モデルは，テキストのモーダル情報がどのように画像認識モデルに作用しているのか明らかではない．そこで，本研究ではシングルモーダル学習モデルとマルチモーダル学習モデルにおける判断根拠の可視化に Grad-CAM [2] を用いて分析する．

3.1. 実験概要

画像とテキストを用いる Early fusion モデルと画像のみを用いる InceptionV3 の判断根拠がどのように変化するかを調査する．また，Early fusion モデルのテキストを変更した場合に判断根拠へどのような影響を与えるのかを比較する．InceptionV3 の学習回数は 10 エポック，バッチサイズは 75，学習率 0.01 である．Early fusion の学習回数は 15 エポック，バッチサイズは 8，学習率は 0.01 である．データセットには UPMC Food-101 を使用する．UPMC Food-101 [3] は 101 クラスの食べ物の分類を行うデータセットであり，画像とテキストをペアで構成される．

3.2. 実験結果と可視化の比較

判断根拠の可視化に使用する各モデルの精度を表 1 に示す．画像とテキストの情報を用いて学習することで，画像のみを用いた場合と比べて，精度が 19.23 ポイント向上した．

表 1：各モーダルモデルの分類精度

手法	画像	テキスト	精度 [%]
Inception V3	✓		70.67
Early Fusion	✓	✓	89.90

InceptionV3 と Early fusion のアテンションマップの可視化結果を表 2 に示す．アテンションマップは，色が赤色に近いほどその領域を注視していることを表す．Early fusion モデルと比べて，分類精度が低い InceptionV3 は画像の中心領域を注視する傾向がある．多くの画像で画像の中央に食べ物があるため，画像のみを使用して学習すると注視領域が中央に偏ったと考えられる．一方で，画像とテキストを使用する Early fusion モデルでは物体を注視した判断根拠が出力されている．このことから，テキスト情報を追加することで，画像とテキストの内容を紐づけて学習し，テキストの内容に応じた箇所に注視したアテンションマップを獲得できるようになったと考えられる．

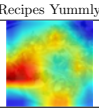
表 2：アテンションマップの可視化結果

クラス名	shrimp and grits	french onion soup	chocolate mousse	tuna tartare	deviled eggs
入力画像					
入力テキスト	Wild Patagonian Shrimp and Grits Cooking and Recipes	french onion soup In Jennie's Kitchen	Christmas Dessert Recipes: Chocolate Mousse Pie CHOW	Ahi Ahi Tuna Tartare Recipes Yummly	Mexican Deviled Eggs Recipe MyRecipes.com
InceptionV3					
Early fusion					

3.3. テキスト入力の変化による可視化の比較

InceptionV3 と Early fusion モデルでアテンションマップが大きく異なったデータについて，テキストを変更してモデルへ入力した場合のアテンションマップの可視化結果を表 3 に示す．これより，テキストの変更内容に応じてアテンションマップが変化することが確認できた．色に関するテキストに変更した場合，より広い領域を注視するアテンションマップとなることが確認できた．細かい位置を指定するテキスト変更した場合，色に関するテキストに変更したときに比べ，アテンションマップがわずかに変化することが確認できた．この結果から Early fusion モデルは，位置を指定したテキスト情報に比べ，色に関するテキスト情報を重要視するモデルであると言える．

表 3：テキスト変更時のアテンションマップ

	変更前	変更後：色に着目	変更後：位置に着目
入力テキスト	Ahi Ahi Tuna Tartare Recipes Yummly	red tomato ketchup	top left white plate
attention map			

4. おわりに

本稿では，シングルモーダルで学習したモデルとマルチモーダルで学習したモデルのアテンションマップを比較した．画像のみを用いた学習と比べて，画像とテキストを用いた学習は，テキストに応じたアンテンションマップを獲得できることを確認した．今後は，シングルモーダル学習モデルとマルチモーダル学習モデルの種類を増やし，異なるモデル構造においても同様の傾向にあるのか調査を行う．

参考文献

[1] I.Gallo, *et. al.*, “Image and Text fusion for UPMC Food-101 using BERT and CNNs”, IVCVZ, 2020.

[2] Ramprasaath R. Selvaraju, *et. al.*, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”, ICCV, 2019.

[3] X.Wang, *et. al.*, “Recipe recognition with large multimodal food dataset”, ICMEW, 2015.