ISS-A-076

深層ニューラルネットワークに基づく テキストと画像間のクロスメディア検索

MAIERHABA AIERKEN† 鳥井 浩平† 松本 和幸†† 吉田 稔†† 北 研二††
† 徳島大学大学院先端技術科学教育部 †† 徳島大学大学院社会産業理工学研究部

1. はじめに

近年，画像やテキストを含むマルチメディアデータがInternet 上に急速に出現している．このような状況では，検索エンジンやマルチメディアデータ管理のための技術として，クロスメディア検索が不可欠となっている．クロスメディア検索とは，異なるメディアタイプのデータ間の意味的な関連性を見つけることで，複数のメディアタイプの検索結果を提供することができる技術である．たとえば，ユーザーはテキストクエリを送信して，クエリを最も適切に示す関連画像または動画を取得できる．本研究では，深層ニューラルネットワークに基づくテキストと画像間のクロスメディア検索を実現することを目的とする．

2. 実験データ

本研究で使用したデータセットは MSCOCO2014 で，トレーニングセットに 82,783 枚(13.5GB)，検証セットに 40,504 枚(6.6GB)，合計 123,287 枚の画像を使用した．また評価のために MSCOCO データセットのテストデータから単語が含まれる画像 200 枚と画像が含まれるテキストデータを用意した．

3. 提案手法

画像と単語それぞれの特徴抽出には深層ニューラルネットワークを用いる．画像の特徴量抽出にはVGG16 のフル結合層(FC)を用いる，また単語に対する特徴量抽出には word2vec を用いる．最終的に正準相関分析(CCA)を用いて画像ベクトルと単語ベクトル間の相関を求める(図 1)．正準相関分析(Canonical Correlation Analysis：CCA)は，2 つの変数の集合の間の関係性を調べる解析手法である．

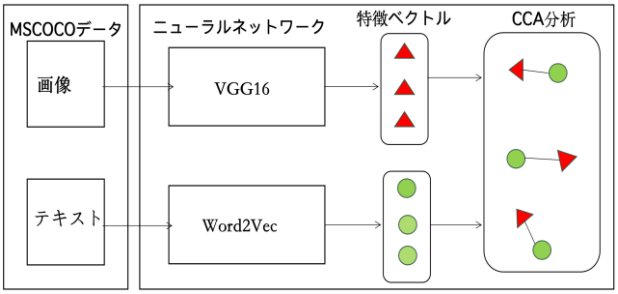

図 1. 画像とテキスト特徴抽出の流れ

4. 実験結果および考察

単語から画像を検索した場合の平均精度は 45.45%，画像から単語を検索した場合の平均精度は 74.94%であった．単語から画像検索では bus，airplane，train などの精度が良かった．しかし backpack，frisbee などの精度は低かった．画像から単語検索では，結果を評価するために，5 人(この研究に直接参加していないメンバー)に，各画像に対して推定された 5 つの単語を検証するように依頼した．その結果，精度は図 2 のようになった．ただし縦軸は正解画像の割合，横軸は正解とする画像の正解単語個数を表す．

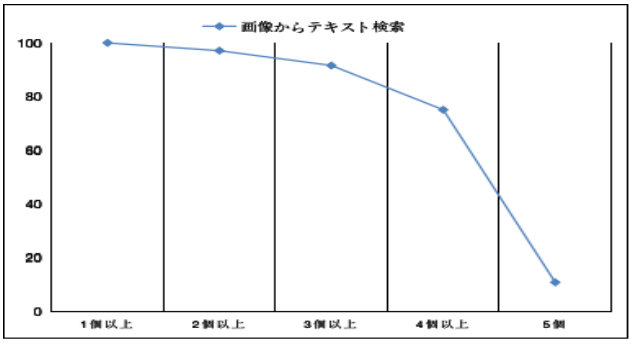

図 2. 画像からの単語検索に対する評価結果

5. おわりに

本稿では，深層ニューラルネットワークに基づくテキストと画像間のクロスメディア検索について述べた．正準相関分析(CCA)を用いることにより，画像特徴量とテキスト特徴量を同一の潜在的意味空間にマッピングすることが可能となった．単語から画像を検索した場合の精度は 45.45%，また画像から単語を検索した場合の精度は 74.94%であった．今後は，データを追加することで精度をさらに向上させたいと考えている．

参考文献

[1] L. Zhang, B. Ma, G. Li et al., “Generalized Semi-supervised and Structured Subspace Learning for Cross-Modal Retrieval”. IEEE Trans. on Multimedia, Vol. 20, Issue 1, pp. 128-141, 2017.

[2] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. “A multi-view embedding space for modeling internet images, tags, and their semantics”. IJCV, 106(2), pp. 210-233, 2014.