H-033

マルチモーダル機械学習における学習法の検討
Investigation of Learning Algorithm for Multimodal Machine Learning

藤森 真綱† 遠藤 伶† 河合 吉彦† 望月 貴裕†
Naotsuna Fujimori Rei Endo Yoshihiko Kawai Takahiro Mochizuki

1. はじめに

放送局では近年、動画、音声、字幕が含まれる放送映像や、テキストとともに画像や動画が投稿される SNS といった、複数の要素（モダリティ）からなる情報を分析し、活用する重要性が高まっている。そのような複数種類の情報を入力するマルチモーダル機械学習は、画像とテキストを用いた分類や発話認識などをはじめとする様々なタスクに適用され、近年盛んに研究されている。

NHK 技研では、SNS に投稿された自然災害や人的災害の情報を活用し、取材を迅速化するため、投稿のテキスト情報からニュース性の有無を判定するソーシャルメディア分析システム[1]を開発している。テキストのみでの判断が難しい場合には、テキストともに投稿された画像の利用が有効だと考えられる。そこで本稿では、Twitter に投稿された画像付きツイートにマルチモーダル機械学習を適用してニュース性のある投稿を抽出・分類する手法を提案する。

マルチモーダル機械学習においては、複数のモダリティを用いた分類モデルと単一のモダリティによる分類モデルを同時に学習するマルチタスク学習を行うことで、モデルの汎化性能が向上し、分類精度が向上することが知られている。しかし、Twitter の投稿の中には、画像のみ、もしくはテキストのみでは情報が不十分であり分類が難しいため、単一のモダリティによる分類モデルの学習に有用でないものが存在する。そのようなデータはモデルの過学習を引き起こし、分類精度を低下させる要因となる。そこで本稿では、学習データの有用度を考慮した学習法およびモデルを提案する。また、Twitter の投稿を用いて構築したデータセットを用いた実験により、従来手法に比べ提案手法の分類精度が向上することを確認する。

2. データセットの構築

2016 年および 2017 年に Twitter に投稿された画像付きの投稿のうち、災害に関するものの一部を収集した。マルチモーダル機械学習により分類モデルを学習するための従来のデータセットには、複数のモダリティからなるデータの組に対して、正解ラベルが一つ付与されている。本稿では、各モダリティの有用度を考慮した学習を行うため、モダリティ毎のラベルを追加したデータセットを構築した。具体的には、画像とテキストの両方を考慮して付与した従来の正解ラベルに加え、画像のみを考慮した画像ラベル、テキストのみを考慮したテキストラベルを付与した。正解ラベルの種別は、「火事」「列車事故」「洪水」「交通事故」「噴火・火山灰」「土砂災害」「ニュース性なし」の 7 カテゴリであり、画像ラベルおよびテキストラベルには上記の 7 カテゴリに加え、単一のモダリティでは情報が十分でないためカテゴリを決定できないことを意味する「不明」ラベルを許容した。さらに、各モダリティの有用度を示す「有用度ラベル」を、第 1（および第 2）要素が、正解ラベルと画像 （およびテキスト）ラベルが等しければ 1、異なれば 0 である長さ 2 のベクトルと定義する。このようにすることで、「不明」ラベルが付けられたモダリティの有用度は 0 となり、単一のモダリティだけでも正しく分類が可能なモダリティの有用度は 1 となる。

† NHK 放送技術研究所

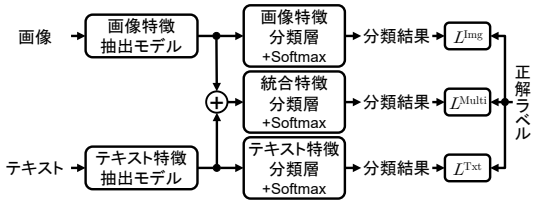

図 1　従来手法のモデル構造

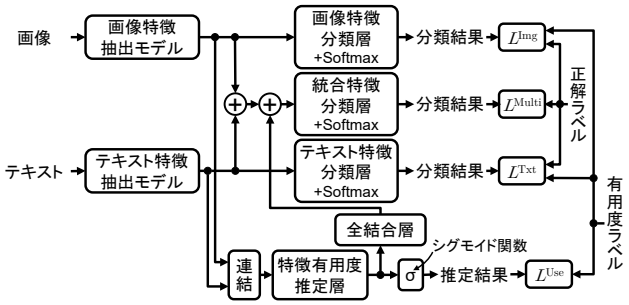

図 2　提案手法のモデル構造

3. 従来手法

図 1 に、従来手法であるマルチタスク学習で用いられるモデル構造を示す。画像特徴抽出モデルおよびテキスト特徴抽出モデルから抽出された特徴ベクトルをそれぞれV^{Img}、V^{Txt}とする。統合特徴ベクトル$V = V^{\mathrm{Img}} + V^{\mathrm{Txt}}$を統合特徴分類層に入力し、Softmax 関数で正規化することで分類結果を得ると同時に、V^{Img}およびV^{Txt}をそれぞれ画像特徴分類層およびテキスト特徴分類層に入力し、Softmax 関数で正規化することで単一のモダリティによる分類結果を得る。3 つの分類結果と正解ラベルから計算されるクロスエントロピーロスをそれぞれL^{Multi}、L^{Img}、L^{Txt}とすると、全体の損失を$L = L^{\mathrm{Multi}} + L^{\mathrm{Img}} + L^{\mathrm{Txt}}$としてマルチタスク学習を行う。

4. 提案手法

従来手法では、単一のモダリティによる分類には有用でないデータによる損失がL^{Img}およびL^{Txt}に含まれており、過学習による精度低下を引き起こす可能性がある。また、統合特徴ベクトルを用いて分類する際にも、各モダリティ

Copyright © 2020 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

の有用度については考慮されていない。本稿では、上記の課題を解決するため、ラベル選別学習および特徴有用度推定を提案する。図 2 に提案手法のモデル構造を示す。以下で、それぞれの手法について説明する。

4.1 ラベル選別学習

ラベル選別学習では、単一のモダリティによる分類モデルの学習を有用なデータのみを用いて行うことで、過学習による精度低下を防ぐ。本手法では、単一のモダリティによる分類結果に対する損失関数を以下のように計算する。

$$L^m = \frac{1}{\sum_{n=1}^{N} U_n^m} \sum_{n=1}^{N} U_n^m \mathrm{CE}(X_n^m, Y_n)$$

ただし、Nはバッチサイズ、mはモダリティの種別（Img, Txt）、$\{X_n^m\}_{n=1}^N$ はモダリティmの特徴分類層による分類結果、$\{Y_n\}_{n=1}^N$は正解ラベル、$\{U_n^m\}_{n=1}^N$はモダリティmの有用度ラベル、$\mathrm{CE}(X,Y)$はXとYのクロスエントロピーを表す。また、有用度ラベルがすべて 0 の場合、損失は 0 とする。

4.2 特徴有用度推定

V^{Img}およびV^{Txt}を連結したベクトルを特徴有用度推定層に入力し、出力をシグモイド関数で要素ごとに正規化することで、各モダリティの有用度を推定する。推定結果と有用度ラベルから計算されるバイナリクロスエントロピーロスをL^{Use}とし、全体の損失を$L = L^{\mathrm{Multi}} + L^{\mathrm{Img}} + L^{\mathrm{Txt}} + L^{\mathrm{Use}}$としてマルチタスク学習を行う。また、特徴有用度推定層の出力を 1 層の全結合層に入力して統合特徴ベクトルVと同じ次元に変換し、Vに足し合わせたうえで統合特徴分類層に入力することで、有用度の情報を分類結果に反映させることができる。

5. 実験

5.1 学習モデル

画像特徴抽出モデルには 18 層の ResNet[2]を用いた。Global Average Pooling 層の出力をV^{Img}とし、ImageNet[3]で学習済みの重みを初期値として用いた。テキスト特徴抽出モデルには GRU[4]を双方向化したものを 2 層用い、アテンション機構を導入した。V^{Img}およびV^{Txt}は 512 次元とした。また、統合特徴分類層、画像特徴分類層、テキスト特徴分類層、特徴有用度推定層には 3 層の全結合層を用いた。

5.2 学習データセット

第 2 節で構築したデータセットのうち、2016 年に投稿された 22,620 ツイートを学習データとし、2017 年に投稿された 2,383 ツイートおよび 21,313 ツイートをそれぞれ検証データおよびテストデータとして用いた。

5.3 学習処理

入力画像は、短辺が 256 ピクセルとなるようにリサイズした後、256×256 の中心領域を切り出し、学習時はランダムに 224×224 の領域を、検証時およびテスト時は 224×224 の中心領域を切り出した。入力テキストは、上記の学習データセットとは別にランダムに収集したツイートを用いて学習した SentencePiece[5]により分割した。語彙サイズは 8000、最大入力長は 75 トークンとした。

学習アルゴリズムには確率的勾配降下法を用い、学習率を 0.01、モーメンタムを 0.9 とした。200 エポック学習し、1 エポックごとに検証データを用いて正答率を計算した。正答率が最も高かったエポック数のモデルの精度を、テストデータを用いて評価した。

	F 値
Baseline	$0.943 \pm 7.3\times10^{-4}$
MT	$0.947 \pm 3.7\times10^{-3}$
MT + LS	$0.951 \pm 4.5\times10^{-4}$
MT + LS + UE (proposed)	**0.952** $\pm 7.8\times10^{-4}$

表 1　分類結果の比較

5.4 結果

結果を表 1 に示す。評価には、全カテゴリの F 値（適合率と再現率の調和平均）の、正解ラベルのデータ数による重み付き平均を用いた。各モデルについて 4 回学習を行い、平均および標準偏差を比較した。Baseline は損失としてL^{Multi}のみを用いたマルチタスク学習を行わないモデルを、MT、LS、US はそれぞれ、マルチタスク学習（Multi-Task Learning）、ラベル選別学習（Label Selection Learning）、特徴有用度推定（Usefulness Estimation）を示す。MT、LS、US を全て行った提案手法が最も精度が高いことがわかる。

また、マルチタスク学習のみを行った場合は、ほかに比べ標準偏差が大きい。これは、単一モダリティによる分類に有用でないデータを学習に用いたことにより、学習が安定しなかったことが原因であると考えられる。

6. おわりに

本稿では、学習データの有用度を利用したマルチモーダル機械学習のモデルおよび学習法を提案した。Twitter に投稿された災害に関するツイートから、正解ラベルに加えてモダリティ毎のラベルを付与したデータセットを構築し、有用度ラベルを定義した。構築したデータセットを用いた実験では、学習データの有用度を考慮することにより分類精度が向上し、提案手法の有効性が確かめられた。

参考文献

[1] J. Goto, T. Miyazaki, Y. Takei, K. Makino, “Automatic Tweet Detection based on Data Specified through News Production”, Proceedings of the 23rd ACM International Conference on Intelligent User Interfaces Companion, pp. 1:1-1:2 (2018).
[2] K. He, X. Zhang, S. Ren, J. Sun, “Deep Residual Learning for Image Recognition”, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778 (2016).
[3] O. Russakovsky, J. Deng, H. Su, J. Krause, et al., “ImageNet Large Scale Visual Recognition Challenge”, International Journal of Computer Vision, vol. 115, no. 3, pp. 211-252 (2015).
[4] K. Cho, B. Merriënboer, C. Gulcehre, D. Bahdanau, et al., “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation”, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724-1734 (2014).
[5] T. Kudo, J. Richardson, “SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing”, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 66-71 (2018).

Copyright © 2020 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.