Vol.2022-CG-188 No.2
Vol.2022-DCC-32 No.2
Vol.2022-CVIM-231 No.2
2022/11/18

形容詞・名詞ペアを学習したモデルの蒸留を用いた画像の感情分析

齋藤 優輝[1,a)]　数藤 恭子[1,b)]

概要：画像の感情分析は，マーケティングにおける推薦システムやコミュニケーションへの応用が期待され盛んに研究されている．しかし, データセットの作成コストが大きいためノイズも発生しやすく，画像から感情への変換の際にギャップが生じてしまうという問題がある．これを改善するために，中間表現である形容詞・名詞ペア (ANPs) を出力する SentiBank が提案されている．しかし，ANPs に変換する際に元画像から色情報や輝度情報などの多くの情報が失われる可能性がある．本研究ではノイズ低減と画像情報の維持を両立したモデルの作成を目的とする．具体的には，画像を画像内のオブジェクト毎に求めた ANPs を自然言語処理モデルに学習させて感情分析し．その出力を，蒸留を用いて ResNet に統合する手法を提案する．今回の提案モデルを用いた実験により，ANPs に分割と蒸留を適用したことによる精度の向上を確認した．

1. はじめに

近年，SNS などの発展に伴い画像に対してどのような感情を抱くか分析することに大きな意義が生まれている．例としては，推薦システムや広告などといったマーケティング, 画像の共有・検索の円滑化などコミュニケーション等への応用が期待されている．機械学習による画像の感情分析における課題として, 二つ大きなものが挙げられる．一つが感情という主観的なものをラベル付けするため, 信頼できるデータセットを作成するのには多くのアノテーターが必要であり，作成コストが大きく，適切でないアノテーションが付与されてしまうようなノイズも発生しやすいこと．もう一つが画像は感情情報よりも多くの情報を含むため，感情情報への変換の際にギャップが生じてしまうことである．ノイズを低減しギャップを埋めるための効果的な手法としては, 画像から形容詞・名詞の組である Adjective Noun Pairs(ANPs) を出力する SentiBank[1] がある．ANPs は画像と感情の中間表現として用いることができ, ギャップを低減できる．しかし，画像から ANPs へと変換する際に，形容詞・名詞の組からなる文章情報へと変換されるため，元画像の印象を左右する色情報，輝度情報などが失われる可能性がある．そのため近年の画像の感情分析の研究では，画像の部分領域ごとに画像情報から抽出した中間表現を取得する手法がより大きな成果を上げている [2]．一方で，近年成果の出ている手法に Knowledge Distillation(蒸留)[4] がある．蒸留はモデル圧縮手法として提案されたが，モデル出力を他のモデルの学習に使用するこの手法は，同じタスクの別種のネットワーク同士の出力の統合にも応用できることが示されている [5][6]. 本研究では蒸留による統合と ANPs を用いて，画像情報の維持とノイズ低減を両立したモデルを提案する．具体的には ANPs に不足してしまう画像情報を多く取り入れるため，画像内のオブジェクト毎に分割して ANPs に変換した後，自然言語処理モデルを学習させ，蒸留を用いてそのモデルを ResNet へ統合することで，ANPs の感情特徴の抽出能力を ResNet に適用した．実験を行った結果, 先行研究の最も精度の高い研究には及ばなかったものの, 蒸留による統合でそれぞれのモデル単独より高い精度が出ることが確認でき，ANPs による画像の感情特徴の抽出と画像情報の維持の両立ができたと言える結果になった．

2. 関連研究

2.1 形容詞・名詞ペア (ANPs) の活用

Borth らは，24 の人間の感情から 3000 以上の概念を含む visual sentiment ontology を構築し，それを元に画像を ANPs に変換する SentiBank を提案した [1]．画像の感情分析の分野では，形容詞の付与を目的として使用される [7] 他，SentiBank の出力である 1200 のベクトルを感情のベクトルとして使用する [2] など，画像から感情特徴を抽出するために幅広く使用されている．

[1] 東邦大学
Toho University
[a)] 6521005s@st.toho-u.jp
[b)] kyoko.sudo@sci.toho-u.ac.jp

Vol.2022-CG-188 No.2
Vol.2022-DCC-32 No.2
Vol.2022-CVIM-231 No.2
2022/11/18

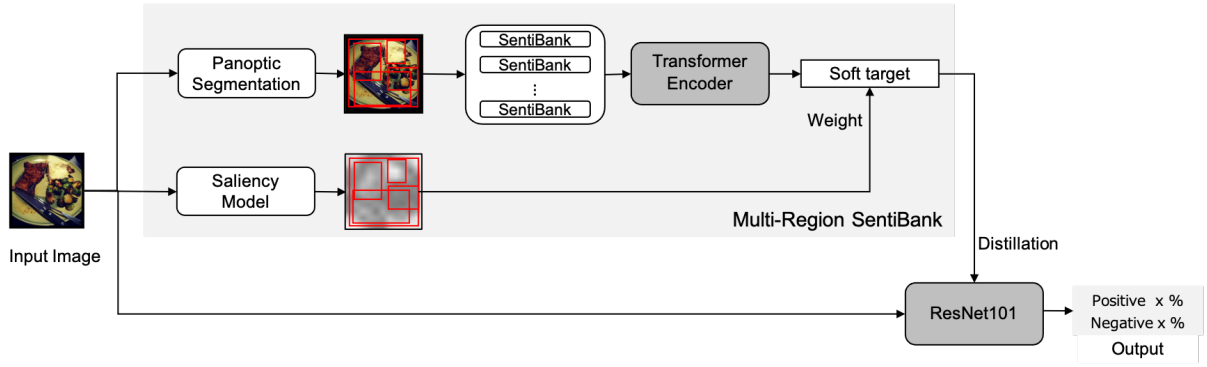

図 1 提案手法の概要

2.2 オブジェクト検出による画像の感情分析

画像内のオブジェクトや，一部領域に注目することで感情分析の精度が向上することが確認されている．Wu らの研究 [3] では，画像全体を感情分析して大域的な感情特徴を取得した後，画像内にある顕著物体が 1 つ以上検出できた場合その領域の局所的な感情特徴を取得し，大域的な感情特徴と統合することで最終的な感情特徴を出力する手法が提案された．Yang らの研究 [2] では，モデルを用いて領域ごとにオブジェクトの存在確率や感情スコアを算出し，CNN を通して得た画像の感情特徴値と統合する手法が提案されている．

2.3 Knowledge Distillation(蒸留)

Knowledge Distillation(蒸留) は異なるモデルへのデータ圧縮として,Hinton らの研究 [4] で提案された手法である．教師モデルとなる大規模モデルの，softmax 層の出力であるクラス分類の確率分布を, 生徒モデルの画像の soft target として使用することで，hard target である one-hot ベクトルラベルのみで学習した場合よりも意味的な学習が可能になる．この soft target の形でのモデル性能の抽出は，異なる抽出能力を持つモデル同士の統合にも活用できる．主観を含むため一意にラベルが決まらない感情認識の研究においては，one-hot ベクトルラベルでない soft target による学習の効果が高いこともあり，この蒸留を用いた統合手法が幾つかの研究で使用されている．Wei らの研究 [6] では，音声・画像・テキストを入力として算出した感情出力の蒸留により, それぞれの入力の相関関係が強まり，いずれか 1 つの入力が欠けた場合でも高い感情分析精度が出ることが確認されている．

3. 提案手法

3.1 全体構成

提案手法の 1 つ目として，画像を分割し SentiBank を通して得た ANPs から感情の確率を出力するモデル，Multi Region SentiBank を提案する．Multi Region SentiBank はまず，疑似的なデータ拡張と画像情報の詳細化を目的として，入力された画像を Panoptic Segmentation を用いてオブジェクト毎に分割する．それら分割画像を SentiBank へ入力して分割画像ごとの ANPs を得る．その後，Transformer-encoder を通して得た各分割画像の感情の確率 P_{TE} を，各分割画像の Saliency map 画素値の総計割合で補正をかけて合算し，分割前画像の感情の確率 P_{SB} として出力する．

提案手法の 2 つ目として，Multi Region SentiBank の出力である P_{SB} を蒸留し，ResNet101 と統合するモデルを提案する．Multi Region SentiBank から得た感情情報を蒸留して ResNet101 の学習に使用することで，画像情報に感情情報を疑似的に付与し，感情の確率 P_{RN} を出力した．これにより，ANPs 文章情報と画像情報を蒸留を用いて統合した場合の精度向上効果の確認をする．

3.2 Multi Region SentiBank

Panoptic Segmentation[8] は，画素ごとにクラス分類を行い, 同じクラスの物体でも複数ある場合は別のオブジェクトとして検出するセグメンテーション手法である．本研究ではオブジェクトの検出のために使用し, 各オブジェクトのバウンディングボックス内の画素を分割画像として SentiBank に入力する．SentiBank の出力である 1200 次元の ANPs 確率分布のうち，先行研究 [7] に倣い上位 4 つの形容詞・名詞ペアを抜き出した．ANPs 自体には感情極性が付与されておらず、自然言語処理モデルを通して感情推定の必要があるため、positive,negative 両方に使われる可能性の高い名詞は感情推定にあまり寄与しない．よって，ANPs の形容詞部のみそのまま入力し，名詞部を Panoptic Segmentation によって検出した際のラベル名に置き換えてテキスト化し，Transformer-encoder への入力とした (**図 2**).

Transformer-encoder では，入力された各テキストの末尾にトークンを付与し，テキストを embedding 層でベクトル化して出力した後，各ベクトルの末尾の出力のみを全結合層へ入力した．出力である P_{TE} は，分割前画像の感情の確率へと統合する必要がある．そのため，P_{TE} に補正

Vol.2022-CG-188 No.2
Vol.2022-DCC-32 No.2
Vol.2022-CVIM-231 No.2
2022/11/18

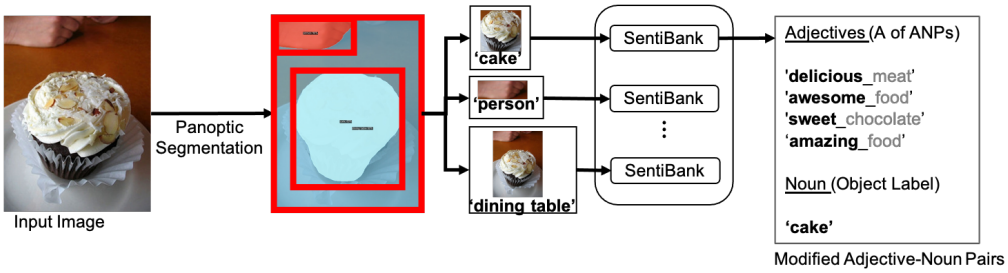

図 2 Panoptic Segmentation による画像分割を利用したテキスト化処理

をかけて合算し，平均して P_{SB} を出力した．この際，補正となる値は分割画像の Saliency map 画素値合計を使用する．Saliency map は公開されている Saliency map 生成モデル [9] を用いて生成した．

3.3 蒸留による統合

事前学習された Multi Region SentiBank から出力される P_{SBi} は，実ラベル $\hat{y}_i$ と併用し ResNet101 の学習時に仮のラベルとして使用する．画像を入力として ResNet が出力した感情の確率を P_{RNi} とし，学習サンプル数を N とすると，損失の計算式は以下のようになる．

$$L_{sp} = -\frac{1}{N}\sum_{i=1}^{N} P_{SBi}\log P_{RNi} + (1-P_{SBi}) * \log(1-P_{RNi}) \quad (1)$$

$$L_r = -\frac{1}{N}\sum_{i=1}^{N} \hat{y}_i \log P_{RNi} + (1-\hat{y}_i) * \log(1-P_{RNi}) \quad (2)$$

$$L = \frac{(\alpha L_{sp} + L_r)}{1+\alpha} \quad (3)$$

それぞれ P_{SBi} と P_{RNi}，$\hat{y}_i$ と P_{RNi} の交差エントロピー誤差であり，最終的な損失は L である．α は損失 L における L_{sp} の割合を決めるためのハイパーパラメータである．

4. 実験

4.1 使用データセット

EmotionROI[10] データセットと Twitter データセット [1] を使用する．EmotionROI データセットは，Emotion6[12] のデータベースを基に作られた計 1980 枚の画像と 6 分類の感情ラベルからなるデータセットである．先行研究 [2] に倣い，6 感情中の anger，disgust，fear，sadness の 4 つを negative に，surprise，joy の 2 つを positive にマージして実験に使用した (neg:1320,pos:660)．Twitter データセットは，3 名のアノテーター—によってラベル付けされたデータセットであり，全アノテーター—の付与ラベルが一致した 603 組の画像と文章と 2 分類の感情ラベルからなる (neg:133,pos:470)．両データセットとも実験には画像とラベルのみを使用した．学習にはデータの 8 割を訓練データとして使用し，残り 2 割をテストデータとして使用する．

4.2 実験詳細

4.2.1 実験項目

以下の 2 つの項目について実験を行った．

(1) 蒸留による統合の効果の確認

ResNet101 と Multi Region SentiBank，蒸留を用いてそれらを統合した提案手法の計 3 つのモデルで精度と F 値を算出し比較する．Multi Region SentiBank は saliency 補正による統合後の確率 P_{SB} と実ラベルを用いて精度と F 値を算出した．

(2) 先行研究との精度比較

Yang らの手法 [2]，Wu らの手法 [3] を比較対象とする．

4.2.2 事前学習・ハイパーパラメータ

ResNet101，および Panoptic Segmentation 出力モデルは，COCO データセットを用いて事前学習済みのものを使用した．Panoptic Segmentation 出力モデルは画像検出手法ライブラリである Detectron2[13] を用いて実装した．Panoptic Segmentation による画像分割と，SentiBank による分割画像の ANPs 生成は事前に行った．Transformer-encoder は ResNet101 の学習に用いる訓練データと同じ画像から出力された分割画像の ANPs を使用して事前に学習した．この際の単語の埋め込み次元数は 10 であり，バッチ数は 20，初期学習率は 0.003 である．損失は交差エントロピー誤差であり，この学習の際は OutputIntegration による統合を行わず，各分割画像 ANPs に対して分割前画像と同様のラベルを疑似的な正解ラベルとして学習した．ResNet 入力時の画像サイズは 255 × 255 にリサイズし，バッチ数は 16，初期学習率は 0.0001，蒸留統合の際の損失 L における L_{sp} の係数 α は 1 にして実験を行った．

4.3 実験結果

蒸留による統合効果の結果は**表 1** のようになった．EmotionROI，Twitter データセットの双方で，Multi Region SentiBank の精度と ResNet101 の精度がほぼ同等なのに対し，蒸留を用いてそれらを統合した ANPs Distillllation は高い精度を示した．一方で，F 値に関しては統合したことによる大きな変化は見られなかった．

先行研究との精度比較の結果は**表 2** の通りである．EmotionROI データセットに関しては Yang らの手法を上回る

表 1 蒸留前後の acurracy/F 値比較

	EmotionROI		Twitter	
手法	精度	F 値	精度	F 値
ResNet101	78.43	0.73	73.39	**0.61**
提案手法 (MR SentiBank)	78.57	**0.75**	74.74	0.53
提案手法 (ResNet101+蒸留)	**83.08**	0.74	**77.80**	0.59

表 2 2つのデータセットでの先行研究との比較

手法	EmotionROI	Twitter
Yang らの手法	81.26	80.48
Wu らの手法	83.04	**80.97**
提案手法 (ResNet101+蒸留)	**83.08**	77.83

ことができ，Wu らの手法とほぼ同等の精度を出すことができた．一方，Twitter データセットでは大きく下回ってしまった．

5. 考察

表 1 の結果から，統合前の 2 モデルである Multi Region SentiBank(proposed）と ResNet101 の結果を統合後のモデルが上回っていることがわかる．よって，蒸留は画像の感情分析において，ANPs 経由の出力と，画像特徴量から直接導いた出力を統合する手法として有効であると考えられる．

表 1 を見ると，統合により精度が向上し，F 値は統合前の 2 モデルの F 値の中間の値となった．その理由として，今回用いたデータセットはいずれも positive/negative の枚数に偏りがあり，統合によって，枚数の多いラベルの正解が増え，枚数の少ないラベルの正解が減少したためだと考えられる．

Yang らの手法と提案手法の比較では，EmotionROI データセットに対する結果では提案手法の精度が上回った。しかし，Twitter データセットに対する結果では提案手法の方が低かった．この原因として，本研究では Yang らの手法や Wu らの手法で用いられているような，画像感情分析の大規模データセットによる事前学習を行っていないことが挙げられる．Twitter データセットはデータ数が 603 枚と小規模であり，また，positive のラベルが付いた画像が 8 割弱を占めるという偏りがあるため，充分な精度が得られなかったと考えられる．

6. まとめ

本研究では，ANPs の画像と感情の中間表現としての性能を活かし，部分画像群から得られる形容詞・名詞ペア (ANPs) を入力として感情極性を出力する自然言語処理モデルと，画像を入力として positive/negative の確率を出力できるように学習した CNN モデルを，蒸留を用いて統合する手法を提案した．ANPs に反映できない画像情報を，部分画像からの抽出と CNN への蒸留を用いた統合により取り込んだモデルの提案と実験を行った．その結果，精度は先行研究をやや下回る結果となったものの，統合による精度向上を確認した．

参考文献

[1] D.Borth, R.Ji, T. Chen, T. Breuel and S. Chang: *Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs*,In Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain (2013).

[2] J. Yang, D. She, M. Sun, M. Cheng, P. L. Rosin and L. Wang: *Visual Sentiment Prediction based on Automatic Discovery of Affective Regions*, IEEE Transactions on Multimedia (2018).

[3] L. Wu, M. Qi, M. Jian and H. Zhang: *Visual Sentiment Analysis by Combining Global and Local Information*, Springer(2020).

[4] G. Hinton, O. Vinyals and J. Dean, *Distilling the Knowledge in a Neural Network*, arXiv (2015).

[5] C. Du, H. Sun,J. Wang,Q. Qi, J. Liao:*Adversarial and Domain-Aware BERT for Cross-Domain Sentiment Analysis*,Proceedings of the 58th annual meeting of the Association for Computational Linguistics.(2020)

[6] W. Peng, X. Hong, G. Zhao, *Adaptive Modality Distillation for Separable Multimodal Sentiment Analysis*, IEEE Intelligent Systems 36, 82-89 (2021).

[7] Z. Li, Q. Sun, Q. Guo, H. Wu, L. Deng, Q. Zhang, J. Zhang, H. Zhang, Y. Chen: *Visual Sentiment Analysis based on Image Caption and Adjective-noun-pair Description*, Soft Computing : 1-13.(2021)

[8] A. Kirillov, R. Girshick, K. He ,P. Dollar: *Panoptic Feature Pyramid Networks*, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019).

[9] L. Itti, C. Koch, E. Niebur:*A Model of Saliency-Based Visual Attention for Rapid Scene Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence*,Vol. 20, No. 11, pp. 1254-1259, Nov(1998).

[10] K. Peng, A. Sadovnik, A. Gallagher, T. Chen. : *Where Do Emotions Come from? Predicting the Emotion Stimuli Map.*, IEEE International Conference on Image Processing (ICIP), (2016).

[11] T. Chen, D. Borth,T. Darrell,S. Chang:*DeepSentiBank: Visual Sentiment Concept Classification with Deep Convolutional Neural Networks*,arXiv preprint arXiv:1410.8586 (2014).

[12] K. Peng, T. Chen, A. Sadovnik, A. Gallagher:*A mixed bag of emotions: Model, predict, and transfer emotion distributions*,In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 860-868).(2015).

[13] Y. Wu and A. Kirillov, F. Massa , W. Lo, R. Girshick:"*Detectron2*", https://github.com/facebookresearch/detectron2(2019).