applied sciences

MDPI

Article

Sentiment Interaction Distillation Network for Image Sentiment Analysis

Lifang Wu, Sinuo Deng, Heng Zhang and Ge Shi *

Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China; lfwu@bjut.edu.cn (L.W.); dsn0w@emails.bjut.edu.cn (S.D.); hz6902420@gmail.com (H.Z.)
* Correspondence: tinkersxy@gmail.com

Abstract: Sentiment is a high-level abstraction, and it is a challenging task to accurately extract sentimental features from visual contents due to the "affective gap". Previous works focus on extracting more concrete sentimental features of individual objects by introducing saliency detection or instance segmentation into their models, neglecting the interaction among objects. Inspired by the observation that interaction among objects can impact the sentiment of images, we propose the Sentiment Interaction Distillation (SID) Network, which utilizes object sentimental interaction to guide feature learning. Specifically, we first utilize a panoptic segmentation method to obtain objects in images; then, we propose a sentiment-related edge generation method and employ Graph Convolution Network to aggregate and propagate object relation representation. In addition, we propose a knowledge distillation framework to utilize interaction information guiding global context feature learning, which can avoid noisy features introduced by error propagation and a varying number of objects. Experimental results show that our method outperforms the state-of-the-art algorithm, e.g., about 1.2% improvement on the Flickr dataset and 1.7% on the most challenging subset of Twitter I. It is demonstrated that the reasonable use of interaction features can improve the performance of sentiment analysis.

Keywords: sentiment classification; knowledge distillation; visual sentiment analysis; convolutional neural networks

Citation: Wu, L.; Deng, S.; Zhang, H.; Shi, G. Sentiment Interaction Distillation Network for Image Sentiment Analysis. *Appl. Sci.* **2022**, *12*, 3474. https://doi.org/10.3390/app12073474

Academic Editor: Krzysztof Koszela

Received: 2 March 2022
Accepted: 28 March 2022
Published: 29 March 2022

1. Introduction

Psychological research has proved that visual content (such as images and videos) can evoke various sentimental responses in human observers [1]. Therefore, with the potential applications on sentiment image retrieval and opinion mining, understanding the sentiment of a given image is of great significance [2–4].

Sentiment analysis is a challenging task because of the diverse set of objects involved and the complex interactions among them. Some researchers have attempted to explore human visual principles using multimodal information and made progress [5,6], but single-modal image emotional classification tasks have developed less. As shown in Figure 1, images that express the same sentiment contain entirely different entities. Simultaneously, as shown in Figure 1, images containing similar objects such as "Teddy bear" express opposite sentiment due to differences in other objects. This indicates that the interaction among objects plays a vital role in visual sentiment. Previous studies focused on how to extract the sentimental features effectively but neglected the interactions. Yang et al. [7] put forward the "Affective Regions", which are components that convey significant sentiments information, and utilized three fusion strategies for the features from the global context and "Affective Regions". Alternatively, Wu et al. [8] enhanced the local features of images, employed a saliency detection model, and improved the classification performance by no small margin.

Local objects effectively enrich the fine-grained features, but they treat the objects as individuals and may see the forest trees. Meanwhile, to supplement the missing global

context in local features, most of the study merges the local branch with a global feature branch by feature concatenation or pool [7]. However, as a high-level abstraction, sentiment contains various objects, and mapping local object features directly to sentiment may generate noise, limiting the performance improvement of the model.

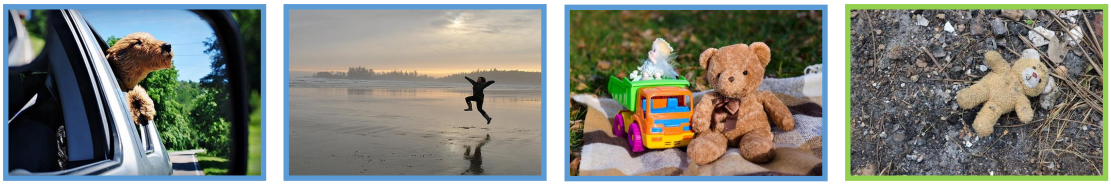

Figure 1. Examples of image sentiment. The blue border represents positive sentiment, while the green border represents negative. We show the categories of objects obtained by panoramic segmentation under the images.

To solve these problems, we proposed a Sentiment Interaction Distillation Network with two branches to exploit the sentimental interaction and transformation of objects. An object branch captures sentimental object interaction as relational knowledge. Then, we employ knowledge distillation [9] strategy to merge it with a global context branch. More effective than directly merging, the sentimental relationships provide more sufficient and general information of the feature distribution and make the distilled knowledge guide the global branch with a different architecture from its teacher, which can use the feature extraction ability of convolution neural network to suppress noise in sentimental relationship information. Compared with the previous study, knowledge distillation achieves a smooth regularization on logits, which means the model can learn robust features [10]. In the test step, only the global context branch is used to generate the predicted sentiment.

This paper introduces the "sentimental graph" to model the object interaction in sentiment space. Specifically, with instance segmentation, we build a graph on the input image, where nodes represent objects and edges describe the sentimental correlations among them. To accurately and appropriately describe the sentimental relationships among objects, we design the adjacency matrices on the base of SentiWordNet [11], which is commonly used in natural language sentiment analysis tasks, and annotation the polarity and strength of words. Then, we employ Graph Convolution Networks to update and aggregate the graph representation, which is used as the distilled knowledge to improve the performance of the context branch. We conduct experiments on five public affective datasets, and our model achieves a better result than the state-of-the-art approaches.

Our contributions are as follows:

- We propose the SID network, which makes comprehensive use of the sentiment interaction among objects rather than directly integrating the visual features. To accurately describe the sentimental interaction, we design a "sentiment graph" to convert images to graph and demonstrate the effectiveness of sentiment relation knowledge.
- We put forward a knowledge distillation method to enhance the global context feature learning. Scene feature learning can obtain better supervision with object interaction constraint by knowledge distillation from sentiment relation.

2. Related Work

2.1. Image Sentiment Prediction

Existing image sentiment analysis methods can be summarized as two groups: dimensional emotion spaces (DES) and categorical emotion states (CES). DES methods usually employ valence–arousal–dominance space [12] or activity–weight–heat space [13] to represent human emotion. On the contrary, CES methods model emotions with categories [14,15], which is more intuitive, and our work falls into categorical states.

Some researchers have devoted themselves to discovering sentimental features in the images and bridging the "affective gap", which can be defined as a lack of strong connection between visual features and sentiment [16]. On the basis of psychological research, Machajdik and Hanbury [15] proposed to utilize the low-level features of images, such as texture, color, and composition, to achieve the image sentiment classification.

With the development of deep learning, You et al. [17] employed AlexNet to achieve the classification of emotions. Sun et al. [18] proposed sentimental regions based on object proposal method and employed the corresponding features to achieve emotion classification. With the help of attention mechanism, You et al. [19] achieved a higher performance by enhancing the local features in the image, which proves that local features have a promoting effect on image sentiment classification.

Further, Yang et al. [7] proposed the "Affective Region" and designed three fusion strategies to utilize the features of "Affective Region". Wu et al. [8] proposed to leverage the salient region of images and made efforts to fuse the features of local and global, which achieve a large performance improvement. Recently, through the use of multimodal information, some researchers have proposed more methods combined with the attention mechanism [20–22], which further promote the development of emotion analysis tasks. These methods make efforts in extracting visual features to improve performance while ignoring the interaction information among objects. In contrast, we propose utilizing the sentimental interaction information and realize the sentiment analysis task.

2.2. Graph Convolutional Network (GCN)

Gori et al. [23] proposed the idea of graph neural networks, which is further developed by Scarselli et al. [24]. However, due to the limitation of computing power, it is prohibitively expensive to realize these methods on massive datasets. Further, Bruna et al. [25] proposed the GCN, which attracted a lot of attention from researchers, and many articles have been published [26,27].

Different to the CNN model, the graph virtually describes the interactions among nodes by modeling the relationship. Based on this, Chen et al. [28] utilized GCN with a multilabel image recognition task to mine object relations from labels. They constructed the correlation matrix by calculating the co-occurrence probability of the labels and obtained a better performance than previous works. However, this method relies on the human-annotated object information.

In this paper, we utilize GCN to capture and explore sentimental interaction information. Specifically, we designed a method to build the sentiment graph for existing image sentiment datasets automatically and extract interaction information among objects in the sentimental space.

2.3. Knowledge Distillation

Knowledge distillation is proposed by Hinton et al. [9] to transfer knowledge from a large model into a smaller, distilled model by minimizing the KL divergence between their logit distributions. The main idea is utilizing soft targets (i.e., the logits distribution of large model) to optimize the small model, as it contains more label distribution information than the one-hot label. Later, Lopez-Paz et al. [29] introduced privileged information in distillation learning, which is additional information available during the training period but not available during testing.

In this paper, local interaction feature learning may capture a fine-grained but biased representation. In contrast, global context feature learning can capture rounded representation over the entire image and produce credible label distribution knowledge. We regard local interaction features as privileged information and exploit this information by distilling it into the global context branch during the training period. During testing, only the global context branch needs to be executed.

3. Methods

An overview of our proposed network architecture is shown in Figure 2. The upper network in blue is the teacher network, while the lower orange one is the student network. Given an image, we aim to condense it into a representation that genuinely captures the sentimental object interaction, which is completed via the object interaction branch. The interaction information is then distilled into another global context branch through a knowledge distillation mechanism. During testing, only the global context branch is kept to generate textual descriptions. We will describe each part in detail as follows.

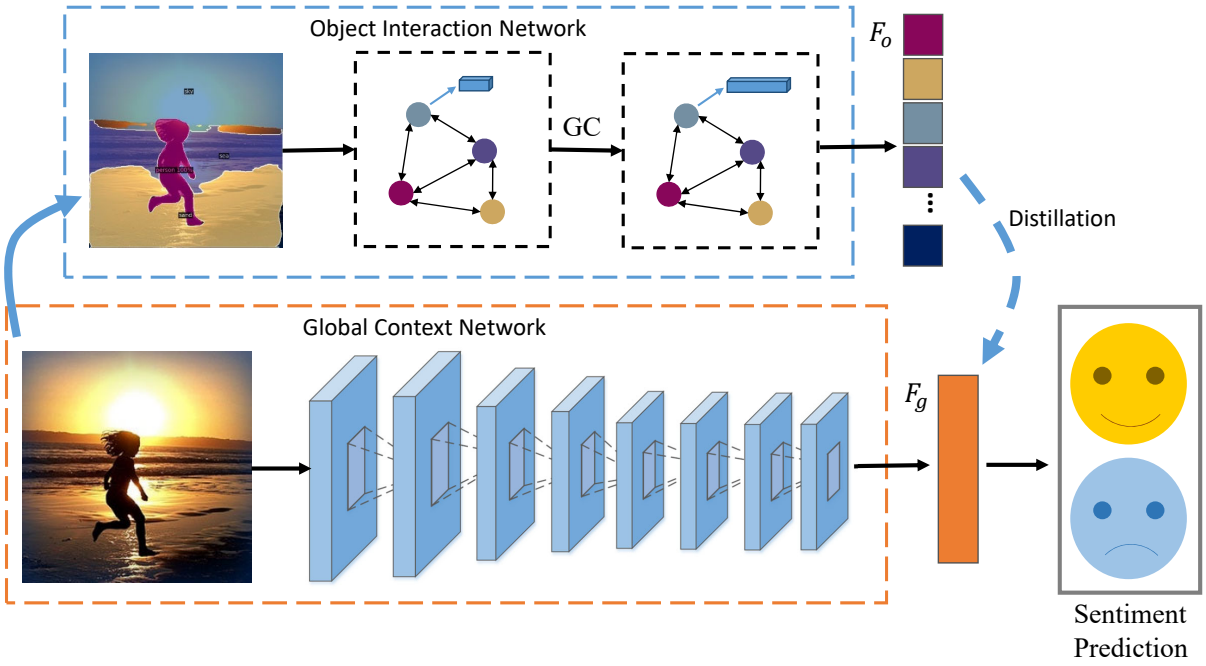

Figure 2. Overview of the Sentiment Interaction Distillation Network. During training, the object branch captures sentiment interaction information through the sentiment graph model, while the global context branch provides the context information. Then, the object information is distilled into the context feature. For testing, only the global context branch is needed for sentiment prediction.

3.1. Sentiment Graph

Sentiment is a intricate logical response, and the sentimental interactions among objects have a vital contribution to it. To accurately describe the sentimental interactions, we constructed a unique undirected sentiment graph (the sentimental relationships among objects) to define the interaction features. The goal of the sentiment graph is to capture interactions among sentiment-related objects, and the relationships are formulated as an adjacent matrix of sentiments. Figure 3 shows an example of a sentimental graph. Inspired by [7], we employed the panoptic segmentation algorithm as an object detector and took the objects as nodes. However, it is challenging to properly describe sentimental relations without annotation because of the gap between object semantics and sentiment. We propose utilizing the objects' semantic relationship in sentimental space as the edges.

Given the nodes, we employ SentiWordNet [11] to label each node with sentimental polarity and strength. SentiWordNet is a lexical resource that annotates three sentiment scores: positivity, objectivity, and negativity to each synset of WordNet. We retrieved nouns and adjectives related to the node of SentiWordNet and determined the strength with the average value of related words. In particular, we defined sentimental polarity in terms of the strength of positive and negative sentiments. For example, the strength of the word "cat" is (0.8, 0.5), and the sentiment polarity is positive of 0.3.

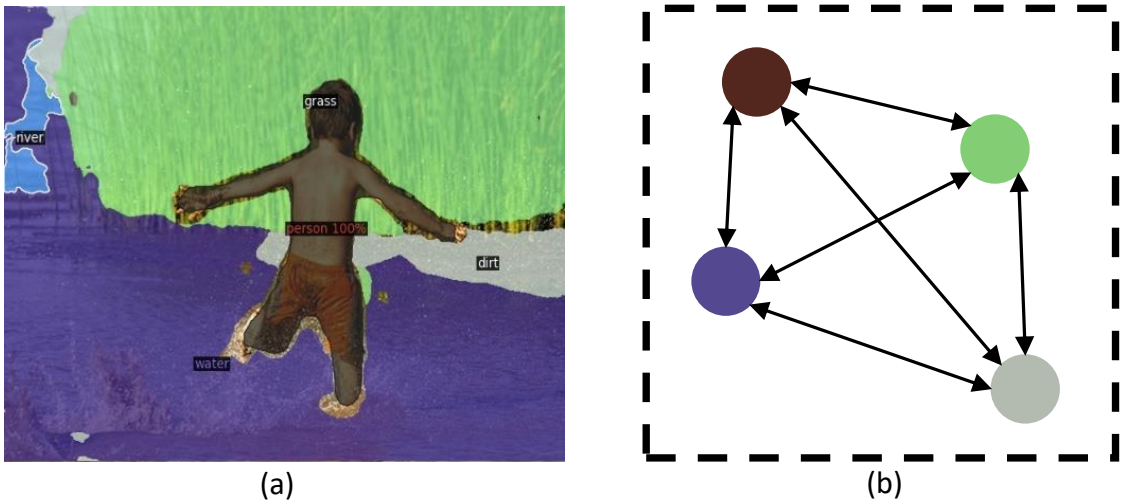

Figure 3. An example of sentiment graph: (**a**) An object segmentation result, where the objects are distinguished by different color overlay. (**b**) A sentiment graph structure, where the nodes represent objects of corresponding color and edges reflect the similarity of nodes in the sentiment space.

Based on this, we quantified the relationships among objects in terms of differences in sentimental strength. As shown in Equation (1), we took the absolute value of the difference in sentimental strength as the relationship between nodes. When the two nodes have opposite sentimental polarities, we set a greater sentimental distance to reflect the difference in sentimental relationships.

$$A_{ij} = \begin{cases} ||S_i| - |S_j|| + 1, & if\ S_i * S_j > 0 \\ 0.5, & if\ S_i = 0, S_j = 0 \\ ||S_i| - |S_j||, & otherwise \end{cases} \quad (1)$$

On the basis of describing interactive information in the graph model, nodes in the graph model correspond to the visual features of each object. We selected handcrafted features, containing texture features and the brightness distribution chart, as the representation of objects. Inspired by [15], we observed the image intensity characteristics on the EmotionROI and the Flickr and Instagram (FI) datasets. Specifically, we quantified the brightness values of the HSI color space to 0–10 and obtained the brightness distribution chart. As shown in Figure 4, the brightness distribution can distinguish the sentimental polarity in some degree. In particular, positive sentiment has a higher distribution than negative when the brightness is 4–6, and negative sentiment is higher at 1–2. At the same time, to supplement details of the image, we utilized the Gray Level Co-occurrence Matrix (GLCM) to describe the texture feature of all objects.

3.2. Convolutions on the Sentiment Graph

To simulate sentimental interactions, we select GCN to propagate and converge representation of objects by the supervised of sentimental relationships. Specifically, we employ a stacked GCN, in which the input of all layers of H^l is the output from the previous layer and the output is a new node feature H^{l+1}.

For example, Equation (2) shows the feature update process of layer l, where $\tilde{A}$ describes the relationship among nodes. The previous layer's output is H^l, and the current layer's is H^{l+1}. The current layer is formed as a weight matrix W^l, and σ is the nonlinear activation function.

$$H^{l+1} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^l W^l) \quad (2)$$

In addition, $\tilde{D}$ is the degree matrix of $\tilde{A}$, which is obtained by Equation (3). H^0 is the input of the first layer, which has 512 dimensions generated from the GLCM introduced above and the brightness histogram.

$$\tilde{D}_{ii} = \sum_j \tilde{A}_{ij} \tag{3}$$

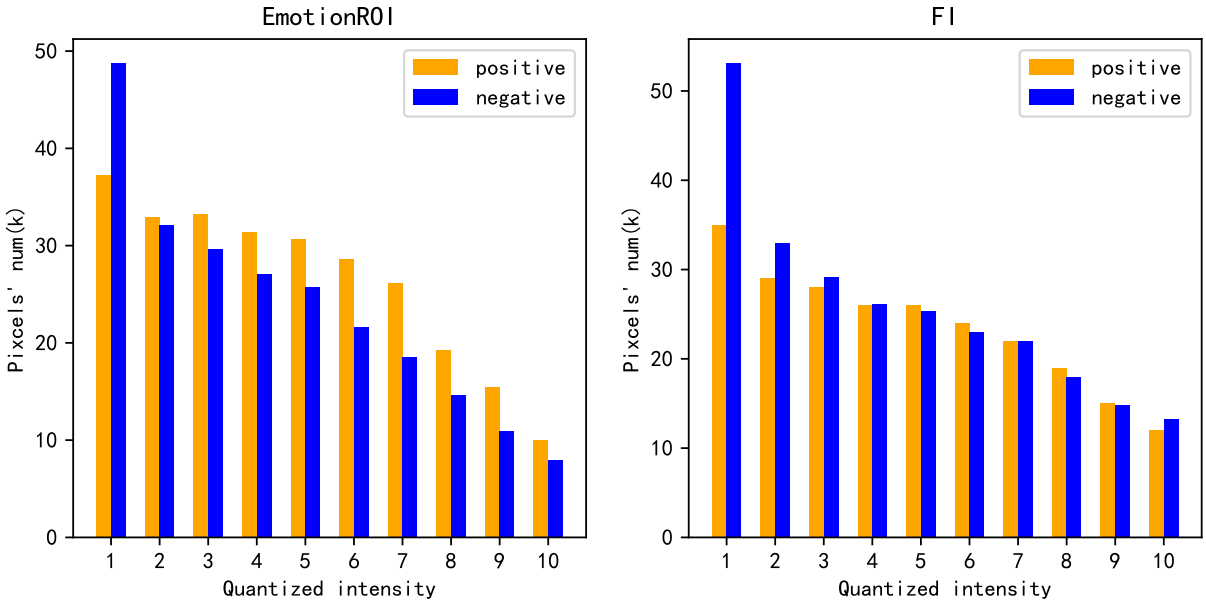

Figure 4. Distribution bar chart displaying the number of brightness pixels of different emotion categories in the EmotionROI and FI dataset.

3.3. Global Context Branch

Similar to previous works [7,8,17–19], we also employed the CNN model to capture the deep visual features of images. To highlight the effect of sentiment interaction information and make a fair comparison with previous works, we keep the scene branch as simple as possible. Previous studies have demonstrated the sentimental feature extraction capability of VGGNet with 16 layers [30]; we selected this as the backbone to supplement the global context information missing in the interactive features. Besides, to effectively extract visual features, we retained the FC layer and changed the last fully connected layer from 4096 to 2048.

3.4. Sentimental Interaction Knowledge Distillation

The problem of merging two branches by concatenating features or pooling [7] is that images contain a variable number of objects, which interferes with the feature learning. This is caused by direct merging, which imposes hard constraints on features from two essentially different spaces. We applied soft regularization only to affective responses, which are essentially interactive knowledge, thus ensuring a robust feature learning process and exploiting object information simultaneously. We aligned sentiment by knowledge distilling, and not just fusion as in direct feature merging. Concretely, we minimized the $L1$ normalization between feature vectors from the two branches. Let $F_o(i)$ be the interaction feature across the objects from the object branch and $F_g(i)$ be the global context branch. We minimized the distillation loss, as shown in Equation (4), where N is the scale of training set.

$$L_{distill} = \frac{\sum_{i=1}^{N} (F_o(i) - F_g(i))}{N} \tag{4}$$

After distillation, the scene feature is sent into the FC layer to achieve the mapping between sentimental polarity and features. The cross entropy function is taken as the

supervision, which is shown in Equation (5), N is the scale of training set, y_i is the labels, and $\hat{y}_i$ is the prediction value of y_i.

$$L_{class} = -\frac{1}{N}\sum_{i=1}^{N}(y_i * log\hat{y}_i + (1 - y_i) * log(1 - log(1 - \hat{y}_i))) \quad (5)$$

The overall loss function consists of two parts:

$$L = \lambda L_{distill} + L_{class} \quad (6)$$

4. Experiment

4.1. Datasets

We evaluated our method on five public datasets: FI [17], Flickr [31], EmotionROI [32], Twitter I [33], and Twitter II [31]. FI was collected from Flickr and Instagram. By querying with eight emotion categories (i.e., amusement, anger, awe, contentment, disgust, excitement, fear, sadness) as keywords, researchers obtained a raw dataset containing approximately 90,000 noisy images. Then, the researchers employed 225 Amazon Mechanical Turk (AMT) workers to label the images, which resulted in 23,308 images receiving more than three agrees and each emotion category having at least 1000 images. The Flickr dataset contains 484,258 images in total, which are labeled by corresponding ANP automatically, meaning its labels are weak. EmotionROI has 1980 sentiment images with six emotions (i.e., joy, surprise, anger, disgust, fear, and sadness), which were collected from Flickr and annotated manually with 15 regions that can evoke emotions. Twitter I was annotated with two sentiment categories (positive and negative) by AMT workers, consisting of 1296 images. Besides, following [34], we implemented the three subsets of Twitter I separately, including "At least three agree", "At least four agree" and "Five agree", which are divided by the number of agrees. For example, "Three agree" refers to obtaining three same sentiment labels from AMT workers. Twitter II contains 603 images that are annotated as positive or negative.

4.2. Baselines

To demonstrate the validity of our proposed method, we first compared our approach against several previous works, including methods using handcrafted features, the CNN-based methods, and deep-learning-based methods with local feature branch.

- Researchers extracted the low-level features from some small-scale datasets, including the local color histogram features (LCH)), which comprise the 64-bin RGB histogram after first dividing into 16 blocks, and the global color histograms (GCH), which comprise the 64-bin RGB histogram [35].
- SentiBank was proposed by Borth et al. [31], and can use 1200 adjective–noun pairs (ANPs) to describe the sentiment concept and performs better for images that have rich semantics.
- DeepSentibank [36] employs CNN to achieve visual sentiment classification and discovering ANPs. We employed the pretrained DeepSentiBank to obtain the 2089 dimension features as mid-level representations from the last FC layer and applied LIBSVM to realize sentiment image classification.
- You et al. [33] proposed a potentially cleaner dataset and designed the PCNN, which is a progressive framework based on CNN. They used large volumes of weakly supervised images to train the model and achieved a generalization improvement.
- Yang et al. [7] utilized an object detection algorithm to label the "Affective Regions" and employed three different fusion strategies to complete the final classification.
- Wu et al. [8] employ a saliency detection method to improve the salient features, and achieved a significant performance boost. Besides, they used an ensemble strategy, which may help improve performance.

4.3. Implementation Details

Following [7], we selected the VGG16 [32] as the backbone and initialized it with the parameters pretrained on the ImageNet. Specifically, the input images were randomly cropped and resized into 224 × 224. During the training period, the images were also flipped random horizontal for data enhancement. We chose SGD as the optimizer on the FI dataset and set the momentum to 0.9. The initial learning rate was set to 0.01, which is 0.1 times per 20 epochs.

We employed two stacked GCN layers in the sentiment interaction branch whose dimensions of output are 1024 and 2048. We expressed the features of each input node in the graph using a 512-dimensional vector. To extract object features, we applied the panoptic segmentation model (Resnet-101 with FPN) [37] pretrained by Detectron2 to detect object boundary and category. The confidence score threshold of detection was set to 0.6. Given the boundaries of output, we applied intensity and GLCM to obtain features from the corresponding regions. Specifically, we quantized the brightness to 0–255 and obtained a 256-dimensional intensity distribution chart. Then, we computed GLCM at 45°, which was reshaped to 256. Finally, we concatenated the intensity features and texture features, which result an object feature of 1×512.

Referring to previous work [7], we applied the same split and test strategy for these datasets without specific division. Specifically, the multilabel datasets, FI and EmotionROI, need to be divided into positive and negative to execute the sentiment polarity classification. The EmotionROI dataset has six emotion categories: joy, surprise, anger, disgust, fear, and sadness. We relabel the joy and surprise as positive and relabel the anger, disgust, fear, and sadness as negative. In FI dataset, eight emotion categories are divided into binary labels based on Mikel's emotion wheel model [38]—amusement, awe, contentment, and excitement are labeled as positive, and anger, disgust, fear, and sadness are labeled as negative. For small-scale datasets, including EmotionROI, Twitter I, and Twitter II, we referred to the commonly used strategy in sentiment analysis [7,8,34] and set the initial weights with the model parameters trained on FI, then fine-tuned the model on small datasets.

For the trade-off parameter in the loss function, we set λ to 0.1, which is tuned on the FI validation set.

4.4. Results

We compared our approach with previous works from published papers [7,8,33]. As shown in Table 1, on four of the five datasets, our method obtains the best outperform, e.g., about 1.2% improvement on the Flickr dataset, 0.97% on EmotionROI, and 1.7% on the most challenging subset of Twitter I. On the FI, however, the performance of our method is second place by only 0.33% difference. We summarize in the following reasons: (1) As the most extensive dataset manually annotated among the five datasets, FI contains more image styles, object types, and contents compared with small-scale datasets. This leads to more frequent classification errors and region division errors in the panoptic segmentation model. Consequently, in our proposed sentiment map, it is more difficult to capture the emotional interaction information between objects. (2) Wu et al. employed a strategy similar to ensemble, which may improve the performance and make it unfair to compare directly with them.

Table 1. Sentiment classification performance on the FI, Flickr, EmotionROI, Twitter I, and Twitter II datasets. The best results are indicated in bold.

Method	FI	Flickr	EmotionROI	Twitter I			Twitter II
				Twitter I-5	Twitter I-4	Twitter I-3	
LCH [35]	-	-	64.29	70.18	68.54	65.93	75.98
GCH [35]	-	-	66.53	67.91	97.20	65.41	77.68
SentiBank [31]	-	-	66.18	71.32	68.28	66.63	65.93
DeepSentiBank [36]	61.54	57.83	70.11	76.35	70.15	71.25	70.23
VGGNet [30]	70.64	61.28	72.25	83.44	78.67	75.49	71.79
PCNN [33]	75.34	70.48	73.58	82.54	76.50	76.36	77.68
Yang [7]	86.35	71.13	81.26	88.65	85.10	81.06	80.48
Wu [8]	**88.84**	72.39	83.04	89.50	86.97	81.65	80.97
Ours	88.71	**73.59**	**84.01**	**89.86**	**87.04**	**83.31**	**81.11**

4.5. Ablation Study

Here, we study how each design in our model influences the overall performance. Local features enrich fine-grained features but concatenate or pool ignored object relationships. Zheng et al. [39] explored the causal relationship between content and emotion in the image. We extend this idea to study the object interaction in an image and implement our experiment. The method we propose consists of two main components at a high level: sentimental interaction and knowledge distillation.

To demonstrate the effectiveness of these two high-level components, we evaluate the performance of several variants to verify the effectiveness of interaction. Firstly, we evaluate (1) Global Branch Only, which only the global context branch is used on; (2) Full Model + Concat, where both branches are used, we fixed the number of nodes in the graph model, where we set the nonexistent nodes as 0 and the fusion of the two branches is completed by concatenation of features directly before passing into a fully connected layer; (3) Full Model + Distill, which minimizes the L1 distance between features for distillation.

Ablation study results of the five datasets are shown in Table 2. Compared with the Global Branch Only, "Full Model + Concat" has an 4.2% average performance improvement, which indicates the effectiveness of sentiment interaction features in image sentiment classification tasks. Thus, "Full Model + Concat" performs worse than the "Full Model + Distill"; this suggests that applying hard constraints on features can exploit valuable object-level information, but may also decrease performance by interfering with the model with noisy features.

Table 2. The model performance comparison across image datasets.

Method	FI	Flickr	EmotionROI	Twitter I			Twitter II
				Twitter I-5	Twitter I-4	Twitter I-3	
Global Branch Only	83.05	70.12	77.02	84.35	82.26	76.75	76.99
Full Model + Concat	88.12	72.31	83.62	89.24	85.19	81.25	80.59
Full Model + Distill	**88.72**	**73.59**	**84.01**	**89.86**	**87.04**	**83.31**	**81.01**

4.6. Qualitative Analysis

To verify that our model can indeed perform a better visual basis after distilling knowledge from object branches, we plotted the saliency maps of the images from EmotionROI. As shown in Figure 5, we observe that the "Full model + Distill" can focus on key regions in the emotion stimulus map better than the corresponding "Global Branch only" part. For the first image, "Full model + Distill" pays major attention to the strawberry as well as the cat, while "Global Branch only" focuses on the strawberry only. In the second image, "Full Model + Distill" put its attention on the lighting and surrounding

areas, while the "Global Branch only" is attention-diffused. Similarly, "Full Model + Distill" focuses the weights more accurately on the Ferris wheel, and "Global Branch only" only focuses on two key regions. This further demonstrates that our proposed sentiment graph as well as the knowledge distillation mechanism endow the model with better visual grounding capabilities.

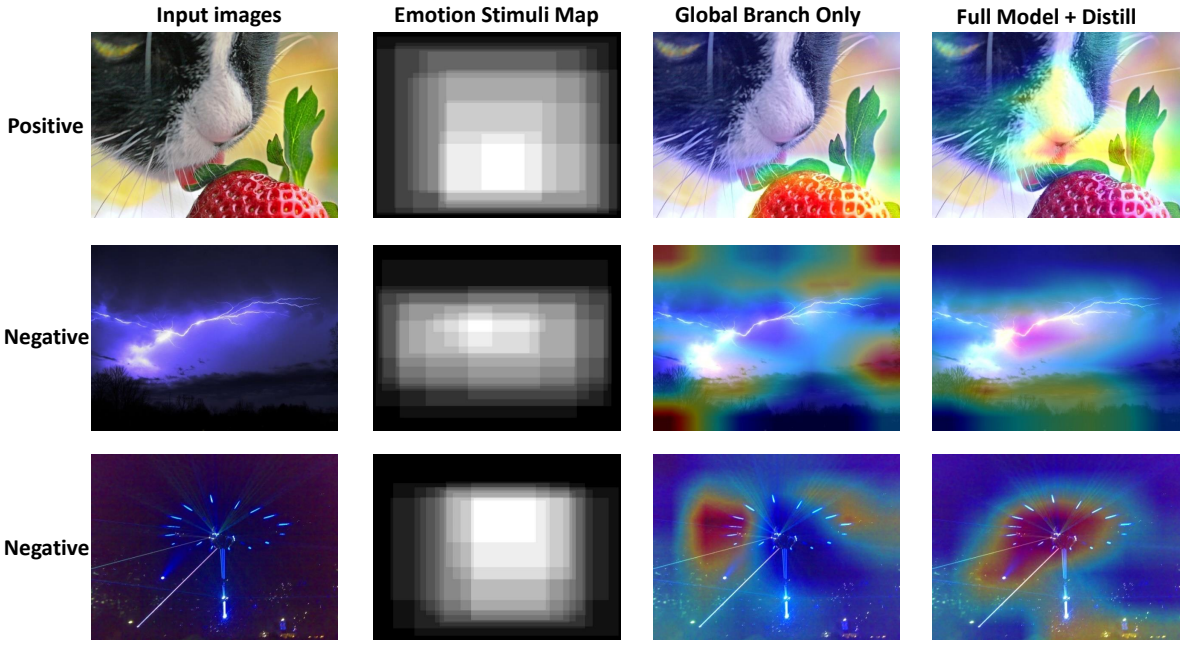

Figure 5. Qualitative results on 3 images from EmotionROI. For each image, the first column from the left is the original input image and the second column is the Emotion Stimuli Map annotated by person, while the third and forth columns are saliency maps from the two models: "Global Branch only" and "Full Model + Distill". Specifically, blue color indicates low attention weights, while red means the opposite.

5. Conclusions

This paper deals with the problem of image sentiment analysis by utilizing sentiment interaction information among objects. Particularly, we studied the problem from the viewpoint of employing sentiment reasoning and relation distillation. To verify this, we presented the Sentiment Interaction Distillation Network to model sentimental interaction information, which consists of two branches: object branch and global context branch. Specifically, we proposed "sentiment graph" to model the sentiment relationship among objects without human annotation, which describes objects with their appearance feature and defines edges with sentimental similarity. Simultaneously, we use stacked GCN models to aggregate and update node features and obtain expressions of emotional interaction.

Further, we employ a knowledge distillation mechanism to avoid the noise caused by segmentation error and the variable number of objects, in which the interaction information is used to supervise global context feature learning. The experimental results demonstrate the effectiveness of our approach on five popular datasets. This work explores the interaction of relational information in visual emotions with visual features; however, more effective use of object interaction information remains a challenging problem. In the future, we will continue exploring the method of integrating abstract affective relational information with specific visual features, which will play an essential role in achieving alignment of different levels of affective information.

Author Contributions: Conceptualization, S.D. and H.Z.; methodology, H.Z.; software, S.D. and H.Z.; validation, H.Z.; formal analysis, S.D. and H.Z.; investigation, S.D. and H.Z.; resources, S.D. and H.Z.; data curation, S.D. and H.Z.; writing—original draft preparation, H.Z. and S.D.; writing—review and editing, L.W. and G.S.; visualization, H.Z.; supervision, L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by National Natural Science Foundation of China grant number 61976010, 62106010, 62106011, 62176011.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing is not applicable to this paper as no new data were created or analyzed in this study.

Acknowledgments: We thank the anonymous reviewers for their valuable comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Detenber, B.H.; Simons, R.F.; Bennett, G.G., Jr. Roll 'em!: The effects of picture motion on emotional responses. *J. Broadcast. Electron. Media* **1998**, *42*, 113–127. [CrossRef]
2. Habernal, I.; Ptáček, T.; Steinberger, J. Supervised sentiment analysis in Czech social media. *Inf. Process. Manag.* **2014**, *50*, 693–707. [CrossRef]
3. Kumar, A.; Srinivasan, K.; Cheng, W.H.; Zomaya, A.Y. Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. *Inf. Process. Manag.* **2020**, *57*, 102141. [CrossRef]
4. Balahur, A.; Jacquet, G. Sentiment analysis meets social media–Challenges and solutions of the field in view of the current information sharing context. *Inf. Process. Manag.* **2015**, *51*, 428–432. [CrossRef]
5. Peng, W.; Hong, X.; Zhao, G. Adaptive Modality Distillation for Separable Multimodal Sentiment Analysis. *IEEE Intell. Syst.* **2021**, *36*, 82–89. [CrossRef]
6. Stappen, L.; Baird, A.; Cambria, E.; Schuller, B.W. Sentiment analysis and topic recognition in video transcriptions. *IEEE Intell. Syst.* **2021**, *36*, 88–95. [CrossRef]
7. Yang, J.; She, D.; Sun, M.; Cheng, M.M.; Rosin, P.L.; Wang, L. Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Trans. Multimed.* **2018**, *20*, 2513–2525. [CrossRef]
8. Wu, L.; Qi, M.; Jian, M.; Zhang, H. Visual Sentiment Analysis by Combining Global and Local Information. *Neural Process. Lett.* **2019**, *51*, 2063–2075. [CrossRef]
9. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
10. Liu, Y.; Cao, J.; Li, B.; Yuan, C.; Hu, W.; Li, Y.; Duan, Y. Knowledge distillation via instance relationship graph. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7096–7104.
11. Esuli, A.; Sebastiani, F. Sentiwordnet: A publicly available lexical resource for opinion mining. In Proceedings of the LREC, Genoa, Italy, 22–28 May 2006; Volume 6, pp. 417–422.
12. Nicolaou, M.A.; Gunes, H.; Pantic, M. A multi-layer hybrid framework for dimensional emotion classification. In Proceedings of the 19th ACM International Conference on Multimedia, Scottsdale, AZ, USA, 28 November–1 December 2011.
13. Xu, M.; Jin, J.S.; Luo, S.; Duan, L. Hierarchical movie affective content analysis based on arousal and valence features. In Proceedings of the 16th ACM International Conference on Multimedia, Vancouver, BC, Canada, 27–31 October 2008.
14. Zhao, S.; Gao, Y.; Jiang, X.; Yao, H.; Chua, T.S.; Sun, X. Exploring principles-of-art features for image emotion recognition. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014.
15. Machajdik, J.; Hanbury, A. Affective image classification using features inspired by psychology and art theory. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010.
16. Hanjalic, A. Extracting moods from pictures and sounds: Towards truly personalized TV. *IEEE Signal Process. Mag.* **2006**, *23*, 90–100. [CrossRef]
17. You, Q.; Luo, J.; Jin, H.; Yang, J. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
18. Sun, M.; Yang, J.; Wang, K.; Shen, H. Discovering affective regions in deep convolutional neural networks for visual sentiment prediction. In Proceedings of the ICME, Seattle, WA, USA, 11–15 July 2016.
19. You, Q.; Jin, H.; Luo, J. Visual sentiment analysis by attending on local image regions. In Proceedings of the AAAI, San Francisco, CA, USA, 4–9 February 2017.
20. Huang, F.; Wei, K.; Weng, J.; Li, Z. Attention-based modality-gated networks for image-text sentiment analysis. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2020**, *16*, 1–19. [CrossRef]
21. Zhang, K.; Zhu, Y.; Zhang, W.; Zhang, W.; Zhu, Y. Transfer correlation between textual content to images for sentiment analysis. *IEEE Access* **2020**, *8*, 35276–35289. [CrossRef]

22. Xu, J.; Li, Z.; Huang, F.; Li, C.; Philip, S.Y. Social image sentiment analysis by exploiting multimodal content and heterogeneous relations. *IEEE Trans. Ind. Inf.* **2020**, *17*, 2974–2982. [CrossRef]
23. Gori, M.; Monfardini, G.; Scarselli, F. A new model for learning in graph domains. In Proceedings of the 2005 IEEE International Joint Conference on Neural Networks, Montreal, QC, Canada, 31 July–4 August 2005; Volume 2, pp. 729–734.
24. Scarselli, F.; Gori, M.; Tsoi, A.C.; Hagenbuchner, M.; Monfardini, G. The graph neural network model. *IEEE Trans. Neural Netw.* **2008**, *20*, 61–80. [CrossRef] [PubMed]
25. Bruna, J.; Zaremba, W.; Szlam, A.; LeCun, Y. Spectral networks and deep locally connected networks on graphs. In Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, 14–16 April 2014; pp. 61–80.
26. Li, M.; Li, S.; Wang, Z.; Huang, L.; Cho, K.; Ji, H.; Han, J.; Voss, C. Future is not One-dimensional: Graph Modeling based Complex Event Schema Induction for Event Prediction. *arXiv* **2021**, arXiv:2104.06344.
27. Liu, F.; Cheng, Z.; Zhu, L.; Liu, C.; Nie, L. A2-GCN: An Attribute-aware Attentive GCN Model for Recommendation. *IEEE Trans. Knowl. Data Eng.* **2020**. [CrossRef]
28. Chen, Z.M.; Wei, X.S.; Wang, P.; Guo, Y. Multi-label image recognition with graph convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5177–5186.
29. Lopez-Paz, D.; Bottou, L.; Schölkopf, B.; Vapnik, V. Unifying distillation and privileged information. *arXiv* **2015**, arXiv:1511.03643.
30. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
31. Borth, D.; Ji, R.; Chen, T.; Breuel, T.; Chang, S.F. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain, 21–25 October 2013.
32. Peng, K.C.; Sadovnik, A.; Gallagher, A.; Chen, T. Where do emotions come from? predicting the emotion stimuli map. In Proceedings of the ICIP, Phoenix, AZ, USA, 25–28 September 2016.
33. You, Q.; Luo, J.; Jin, H.; Yang, J. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In Proceedings of the Twenty-ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.
34. Yang, J.; She, D.; Lai, Y.K.; Rosin, P.L.; Yang, M.H. Weakly supervised coupled networks for visual sentiment analysis. In Proceedings of the CVPR, Salt Lake City, UT, USA, 18–23 June 2018.
35. Siersdorfer, S.; Minack, E.; Deng, F.; Hare, J. Analyzing and predicting sentiment of images on the social web. In Proceedings of the 18th ACM international conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 715–718.
36. Chen, T.; Borth, D.; Darrell, T.; Chang, S.F. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv* **2014**, arXiv:1410.8586.
37. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
38. Mikels, J.A.; Fredrickson, B.L.; Larkin, G.R.; Lindberg, C.M.; Maglio, S.J.; Reuter-Lorenz, P.A. Emotional category data on images from the International Affective Picture System. *Behav. Res. Methods* **2005**, *37*, 626–630. [CrossRef]
39. Zheng, H.; Chen, T.; You, Q.; Luo, J. When saliency meets sentiment: Understanding how image content invokes emotion and sentiment. In Proceedings of the ICIP, Beijing, China, 17–20 September 2017.