International Journal of Computer Vision (2023) 131:1073–1090
https://doi.org/10.1007/s11263-023-01752-7

Sentimental Visual Captioning using Multimodal Transformer

Xinxiao Wu[1,2] · Tong Li[1]

Received: 29 December 2021 / Accepted: 6 January 2023 / Published online: 6 February 2023

Abstract

We propose a new task called *sentimental visual captioning* that generates captions with the inherent sentiment reflected by the input image or video. Compared with the stylized visual captioning task that requires a predefined style independent of the image or video, our new task automatically analyzes the inherent sentiment tendency from the visual content. With this in mind, we propose a multimodal Transformer model namely *Senti-Transformer* for sentimental visual captioning, which integrates both content and sentiment information from multiple modalities and incorporates prior sentimental knowledge to generate sentimental sentence. Specifically, we extract prior knowledge from sentimental corpus to obtain sentimental textual information and design a multi-head Transformer encoder to encode multimodal features. Then we decompose the attention layer in the middle of Transformer decoder to focus on important features of each modality, and the attended features are integrated through an intra- and inter-modality fusion mechanism for generating sentimental sentences. To effectively train the proposed model using the external sentimental corpus as well as the paired images or videos and factual sentences in existing captioning datasets, we propose a two-stage training strategy that first learns to incorporate sentimental elements into the sentences via a regularization term and then learns to generate fluent and relevant sentences with the inherent sentimental styles via reinforcement learning with a sentimental reward. Extensive experiments on both image and video datasets demonstrate the effectiveness and superiority of our Senti-Transformer on sentimental visual captioning. Source code is available at https://github.com/ezeli/InSentiCap_ext.

Keywords Sentimental visual captioning · Visual sentiment analysis · Transformer

1 Introduction

Visual captioning aims to generate textual descriptions for images or videos and has made great progress based on the encoder-decoder framework (Vinyals et al., 2015; Anderson et al., 2018; Luo et al., 2021; Yao et al., 2015; Yang et al., 2021). It mainly focuses on describing the visual content in an objective and neutral manner, while ignoring the linguistic style of sentences. Therefore, stylized visual captioning has been proposed (Mathews et al., 2016; Guo et al., 2019; Zhao et al., 2020b; Wu et al., 2022), which incorporates a specified linguistic style into natural language descriptions. However, this task assumes that the linguistic style is predefined, which may not hold in real applications. Moreover, the given style may not be consist with the underlying emotion of the image or video. For example, the emotion expressed in Fig. 1(a) is happiness, so it is inappropriate to generate a specified negative caption. Therefore, exploring the inherent sentiments within images or videos is non-trivial and critical for generating more reasonable and sentimental captions.

In this paper, we propose a new task, called sentimental visual captioning, to generate captions that embody the underlying sentiment expressed by images or videos. This new task relaxes the assumption of style independence in existing stylized visual captioning methods and has wide applications in real scenarios, including helping people with visual impairments and infants in early education to better understand images and videos from more perspectives, and

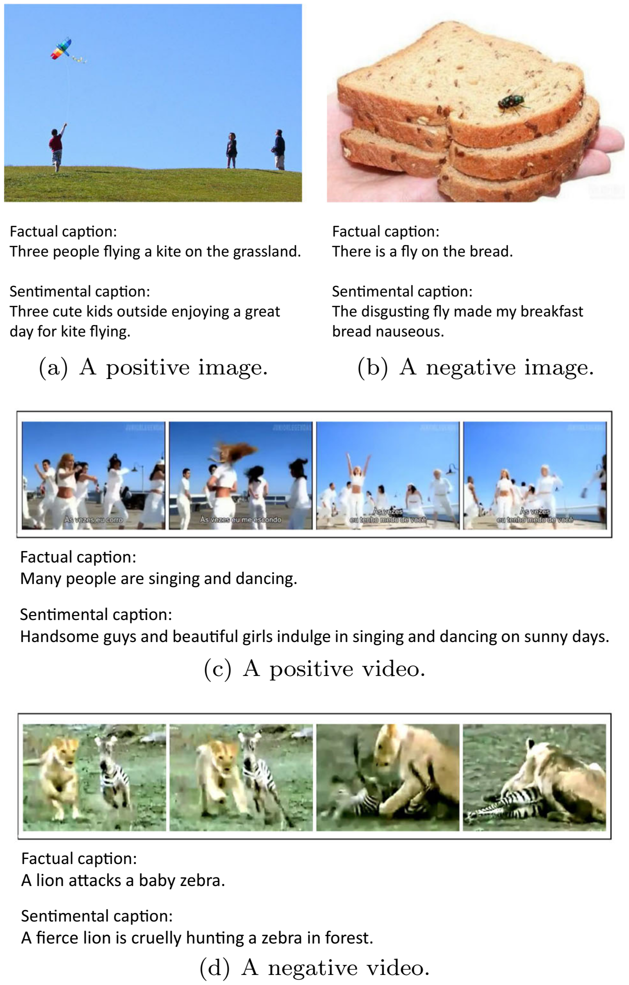
Communicated by Shin'ichi Satoh.

✉ Xinxiao Wu
wuxinxiao@bit.edu.cn

Tong Li
litong11@bit.edu.cn

[1] Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China

[2] Guangdong Provincial Lab of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, Shenzhen 518172, China



Factual caption:
Three people flying a kite on the grassland.

Sentimental caption:
Three cute kids outside enjoying a great day for kite flying.

(a) A positive image.

Factual caption:
There is a fly on the bread.

Sentimental caption:
The disgusting fly made my breakfast bread nauseous.

(b) A negative image.

Factual caption:
Many people are singing and dancing.

Sentimental caption:
Handsome guys and beautiful girls indulge in singing and dancing on sunny days.

(c) A positive video.

Factual caption:
A lion attacks a baby zebra.

Sentimental caption:
A fierce lion is cruelly hunting a zebra in forest.

(d) A negative video.

Fig. 1 Examples that reflect positive and negative sentiments. **a** and **b** show a factual caption and a caption with the image sentiment, respectively. **c** and **d** are examples of positive and negative videos

assisting social platforms to automatically generate appropriate captions for the images and videos uploaded by users. However, the sentimental visual captioning is very challenging since it requires not only understanding the visual content comprehensively, but also analyzing the intrinsic visual sentiment deeply as well as incorporating the sentimental elements into captions appropriately. Moreover, we have no access to large-scale pairs of image or video and sentimental caption for training, and also it is extremely time-consuming and labor expensive to collect them.

To address the challenging issues, we propose a multimodal Transformer called Senti-Transformer for sentimental visual captioning. It first extracts the content and sentiment information from multiple modalities and incorporates prior knowledge from sentimental corpus, then encodes these information by a multi-head Transformer encoder, and finally generates the sentimental description via a decomposed multimodal decoder.

To be specific, we extract multimodal information by performing visual feature extraction, sentiment analysis, concept detection and audio feature extraction (only for video). We construct a prior knowledge base from extra sentimental corpus to infer sentiment words related to the visual content according to the extracted concept words. To encode these multimodal information, we propose a multi-head Transformer encoder, where a head encoder is designed for one specific sentiment to encode visual features since different sentiments attend different aspects of visual content, and additional standard Transformer encoders are utilized for encoding other features. In decoding the encoded multimodal features for sentimental caption generation, we decompose the attention layer in the middle of Transformer decoder to focus on most important and discriminative features of each modality, then we design an intra- and inter-modality fusion mechanism for word prediction.

To effectively train our Senti-Transformer using the pairs of image or video and factual caption as well as the independent sentimental corpus, we propose a two-stage training strategy. In the first stage, to enable the model to incorporate sentimental elements into sentences, we propose a sentimental regularization term defined by reconstructing sentimental sentences in the corpus, thus benefiting faster and better model training in the next stage. In the second stage, to encourage the model to pay more attention to the sentimental part of a sentence, we introduce reinforcement learning and propose a sentimental reward calculated by the sentence evaluations on sentimentality and fluency, further improving the quality of the generated sentences.

The main contributions of this paper are:

- We propose a novel task, sentimental visual captioning, that aims to generate language descriptions of images or videos with the inherent sentiment style reflected by the visual content, making it more practical and general in real-world scenarios.
- We propose a Senti-Transformer for the sentimental visual captioning, where a multi-head Transformer encoder and a decomposed Transformer decoder are newly designed to generate sentimental descriptions by effectively utilizing the content and sentiment information from multiple modalities and the prior knowledge from extra sentimental corpus.
- We propose a two-stage training strategy where a new sentimental regularization is designed for pre-training to incorporate the sentimental elements to captions, and a new sentimental reward is designed for fine-tuning to further improve the fluency and relevance of captions.
- Experiments on the COCO (image) (Lin et al., 2014) and MSR-VTT (video) (Xu et al., 2016) datasets evaluate the effectivenss and superiority of our Senti-Transformer.

We extend a preliminary version of this work (Li et al., 2021b) in the following aspects: (1) proposing a novel Senti-Transformer to replace LSTM-based model, thus further enhancing the ability of handling multi-modal information; (2) applying our Senti-Transformer to sentimental video captioning by integrating more modalities and modifying the encoding and decoding model parts; (3) designing a perplexity reward function in the reinforcement learning stage to improve the sentimentality and fluency of the generated sentences.

2 Related Work

2.1 Visual Captioning

Since the neural image caption model (Vinyals et al., 2015) is proposed, the encoder-decoder framework has become the mainstream of image captioning methods. Afterwards, many researchers explore the use of attention mechanism to generate more accurate and richer captions. You et al. (2016) use a semantic attention to dynamically focus on visual concepts. Anderson et al. (2018) propose a combined bottom-up and top-down attention mechanism to focus on salient objects in images. Wang et al. (2019) propose a hierarchical attention network to learn a feature pyramid by leveraging patch features, object features and text features. Compared with images, videos are more complex due to large variations on both static appearance and dynamic motion, therefore the video captioning task is more challenging. Yao et al. (2015) consider both the local and global temporal structure of videos using an attention mechanism. Yu et al. (2016) exploit both temporal- and spatial-attention mechanisms to selectively focus on visual elements during generation. Chen et al. (2018) propose a plug-and-play PickNet to select informative frames to reduce computational cost. Fang et al. (2019) design a coarse-to-fine and inherited attention structure to focus on more useful visual information.

In recent years, Transformer (Vaswani et al., 2017) has been successfully applied to visual captioning. For image captioning, Huang et al. (2019) model the relevance between attention results and queries through an attention on attention module. Guo et al. (2020) propose a normalized and geometry-aware self-attention network to promote the performance of image captioning. Cornia et al. (2020) explore low- and high-level features by connecting multilayer encoders and decoders through a mesh scheme. Luo et al. (2021) introduce a novel dual-level collaborative Transformer network to show the complementary advantages of grid features and region features. For video captioning, Suin and Rajagopalan (2020) and Pan et al. (2020) use Transformer as the basic model to generate video descriptions. Lei et al. (2020) use a memory module to equip the network architecture with the ability of modeling the previous history of video segments and sentences. Yang et al. (2021) employ a bi-directional self-attention based network to achieve a non-autoregressive and coarse-to-fine captioning procedure.

All these methods focus on generating the factual descriptions of images or videos without taking into account the linguistic styles of sentences. In contrast, our method explores the intrinsic sentiment of input visual content and incorporates it into caption generation appropriately, achieving both factual description and emotion expression.

2.2 Stylized Visual Captioning

Stylized visual captioning aims to generate captions with specified linguistic styles for images or videos and has attracted increasing attention in recent years. Mathews et al. (2016) first propose a switching RNN with word-level regularization for generating positive or negative captions. Gan et al. (2017) design a factored LSTM to extract style factors from the stylized corpus. Chen et al. (2019) use a novel layer normalization strategy to disentangle the language styles from the content. Li et al. (2021a) propose an extract-retrieve-generate data augmentation framework to expand the insufficient paired stylized data for training the captioning model.

All these methods are under a single-style setting, that is, one model is trained for one style. In order to achieve multi-style visual captioning, i.e. a single model is trained to generate sentences in multiple styles, Guo et al. (2019) propose an adversarial learning network that can handle multiple styles simultaneously. Zhao et al. (2020b) present a MemCap method, where a style memory module is designed to memorize the style knowledge learned from corpus. Wu et al. (2022) propose a multi-pass decoding process method for stylized image captioning, where multiple cooperative neural modules are trained under a reinforcement learning paradigm.

Rather than assuming that the sentence style is predefined in the aforementioned methods, our method automatically analyzes the underlying emotion of the input image or video and treats it as the linguistic style for captioning.

2.3 Visual Sentiment Analysis

As CNN have achieved remarkable success in many computer vision tasks, it has also been used for visual sentiment analysis. For the image sentiment analysis, (You et al., 2015; Campos et al., 2017) use CNNs to extract image features and then add several fully connected layers to recognize image sentiment. (You et al., 2017; Yang et al., 2018b, a) not only use the global information of the image, but also consider the local information of the image. Lin et al. (2020) develop an MS-GAN framework to adapt the visual sentiment from multiple source domains to a target domain. For the video sentiment analysis, Bargal et al. (2016) compute deep features using three CNN-based networks and then train a SVM for emotion classification. Nguyen et al. (2018) propose a new feature-level fusion approach based on a bilinear pooling strategy to combine the visual and audio feature vectors. Zhao et al. (2020a) integrate spatial, channel-wise and temporal attentions into a visual 3D CNN and temporal attentions into an audio 2D CNN for emotion recognition in videos.

3 Senti-Transformer

3.1 Overview

We first define the new sentimental visual captioning task and distinguish it from the stylized visual captioning task. For the stylized visual captioning, the input is an image or a video with a fixed style class, and the output is a sentence of the corresponding style, defined as

$$Y = \mathrm{M}_{sty}(I, s), \tag{1}$$

where M_{sty} represents the stylized visual captioning model, I represents the input image or video, s represents the style class, and Y represents the generated stylized sentence.

For the sentimental visual captioning, only the input image or video is required, and the generated sentimental sentence with the sentiment class is output, defined as

$$(Y, s) = \mathrm{M}_{sen}(I), \tag{2}$$

where M_{sen} represents the sentimental visual captioning model, and s represents the sentiment class of the input image or video.

We propose a Senti-Transformer for sentimental visual captioning, which integrates both content and sentiment information from multiple modalities and incorporates prior sentimental knowledge to generate sentimental descriptions of images or videos. As illustrated in Fig. 2, it consists of four parts: multimodal information extraction, sentimental prior knowledge incorporation, multi-head Transformer encoding and decomposed multimodal decoding. The multimodal

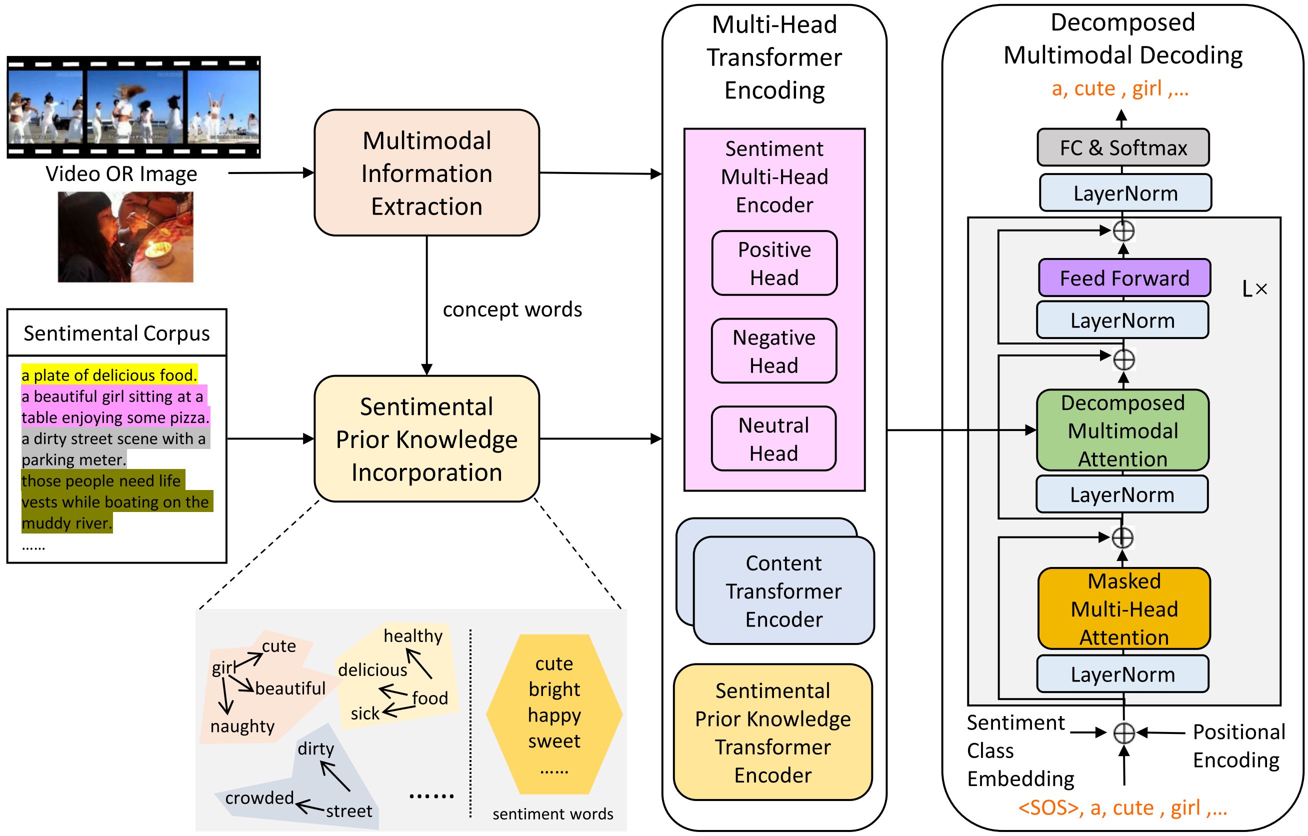

Video OR Image
Multimodal Information Extraction
concept words
Sentimental Corpus
a plate of delicious food.
a beautiful girl sitting at a table enjoying some pizza.
a dirty street scene with a parking meter.
those people need life vests while boating on the muddy river.
......
Sentimental Prior Knowledge Incorporation
cute
girl
beautiful
naughty
healthy
delicious
food
sick
dirty
crowded
street
......
cute
bright
happy
sweet
......
sentiment words
Multi-Head Transformer Encoding
Sentiment Multi-Head Encoder
Positive Head
Negative Head
Neutral Head
Content Transformer Encoder
Sentimental Prior Knowledge Transformer Encoder
Decomposed Multimodal Decoding
a, cute , girl ,...
FC & Softmax
LayerNorm
Feed Forward
LayerNorm
L×
Decomposed Multimodal Attention
LayerNorm
Masked Multi-Head Attention
LayerNorm
Sentiment Class Embedding
Positional Encoding
<SOS>, a, cute , girl ,...

Fig. 2 The framework of our proposed multimodal Senti-Transformer. Firstly, extract both content and sentiment multimodal information from the image or video and incorporate prior knowledge obtained from the sentimental corpus as inputs. Subsequently, these information are respectively encoded by independent encoders. Finally, the decoder performs attention fusion on the encoded features to generate a caption word by word

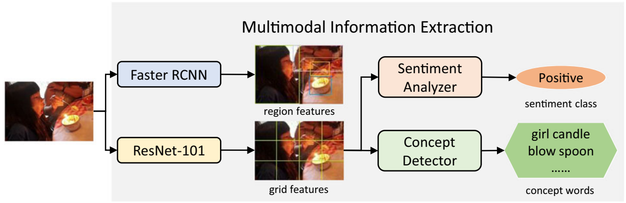


Fig. 3 Multimodal information extraction

information extraction refers to visual feature extraction, sentiment analysis, concept detection and audio feature extraction (only for video), which provide a wealth of information about content and sentiment. The sentimental prior knowledge incorporation refers to an importance calculation method, which is used to dig important object-sentiment words pairs in the corpus as prior knowledge. In the multi-head Transformer encoding, a multi-head encoder is designed to encode multimodal features for each sentiment and other inputs are individually feature-enhanced by independent transformer encoders. In the decomposed multimodal decoding, we decompose the middle attention layer of Transformer decoder to dynamically focus on important features of each modality, then the decoder fuses these attended features using an intra- and inter-modality fusion mechanism to generate a sentence word by word.

We train our model in two stages. In the first stage, we pre-train the model using paired image/video-factual caption data by combining a new sentimental regularization and a standard cross-entropy loss. The sentimental regularization is formulated by sentence reconstruction in the sentimental corpus. In the second stage, we fine-tune the model via reinforcement learning with a new sentimental reward calculated by a sentence sentiment classifier and a language model.

Compared with our preliminary InSentiCap, the new contributions of the proposed Senti-Transformer are: (1) adding the multi-head Transformer encoding part to encode the extracted multimodal features; (2) proposing the decomposed multimodal decoding part to replace the original LSTM-based decoder.

We will present our Senti-Transformer in the following Sects. 3.2, 3.2.2, 3.4, 3.5 and 3.6 by taking the image captioning for example. The video captioning will be elaborated in Sect. 3.7.

3.2 Multimodal Information Extraction

We extract visual features, sentiment class label and concept words from the image to provide content and sentiment information for captioning, as shown in Fig. 3.

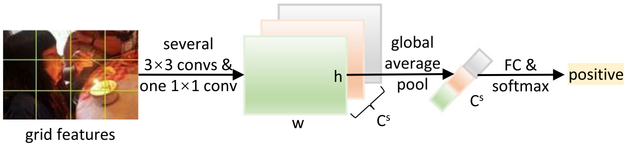


Fig. 4 Sentiment analyzer

3.2.1 Visual Feature Extraction

We employ Faster R-CNN in conjunction with ResNet-101 (Anderson et al., 2018) to detect salient objects in the image and extract the visual features of the object regions as the visual content information for captioning. The extraction process is as follows:

$$\boldsymbol{V}^c = \mathrm{M}_R(I), \tag{3}$$

where I represents the input image and M_R is Faster R-CNN in conjunction with ResNet-101. $\boldsymbol{V}^c \in \mathbb{R}^{H^c \times W^c \times D^c}$ represent the extracted object region features, where H^c, W^c and D^c are the height, width and dimension of region features, respectively. When encoding, the two dimensions of H^c and W^c are flattened to one dimension M^c for use.

Since grid features provide more rich and detailed information (e.g., color and texture) that reflects image emotions, we use them as the visual sentiment information, formulated by

$$\boldsymbol{V}^s = \mathrm{M}_G(I), \tag{4}$$

where M_G is the last convolutional layer of the pre-trained ResNet-101 (He et al., 2016). $\boldsymbol{V}^s \in \mathbb{R}^{H^s \times W^s \times D^s}$ represent the grid features where H^s, W^s and D^s denote the height, width and dimension, respectively. When encoding, the two dimensions of H^s and W^s are flattened to one dimension M^s for use. The extracted grid features also serve as the input of the sentiment analyzer and the concept detector.

3.2.2 Sentiment Analysis

We design a sentiment analyzer to predict the sentiment class as textual sentiment information. As shown in Fig. 4, after the grid features pass through several convolutional layers with ReLU function, a 1 × 1 convolution with C^s filters is first used for dimension reduction, where C^s represents the number of sentiments. After that, through the global average pooling, we get an C^s-dimensional vector. Finally, the sentiment class s_i is obtained through a fully connected layer and a softmax operation.

We train the sentiment analyzer using the public image sentiment analysis datasets (Peng et al., 2016; Machajdik and Hanbury, 2010; You et al., 2015; Borth et al., 2013) and apply it to the image captioning dataset (Lin et al., 2014). Since there exists a domain gap between these two different kinds of image datasets, we propose a threshold filtering method on sentiment classification scores to reduce the domain gap. That is, only when the classification score exceeds a threshold λ_{ss}, the image is classified into the corresponding sentiment class, otherwise it is considered as neutral.

3.2.3 Concept Detection

We design a concept detector by three fully connected layers and a sigmoid operation to extract objects and their relationships in the image. The concept detector takes the global grid feature as input, and outputs the nouns and verbs that correspond to objects and relationships, respectively. It is formulated by

$$\boldsymbol{P} = \text{Sigmoid}(\text{FC}_{\times 3}(\boldsymbol{v}_g^s)), \tag{5}$$

where $\boldsymbol{v}_g^s = \frac{1}{M_s}\sum_{i=1}^{M_s} \boldsymbol{v}_i^s$ represents the global grid feature, generated by performing the mean operation on the local grid features $\boldsymbol{v}_i^s \in \mathbb{R}^{D^s}$. $\boldsymbol{P} \in \mathbb{R}^M$ is the probability distribution on M concept words.

During training, each image is assigned to multiple concept words and the loss function is given by

$$\mathcal{L} = -\frac{1}{M}\sum_{i=1}^{M}[p_i^* \log(p_i) + (1 - p_i^*)\log(1 - p_i)], \tag{6}$$

where p_i^* is 1 if the image is assigned to the i-th concept word, otherwise it is set to 0. During inference, N_c concept words are detected as the textual content information, represented by the word embedding features $\boldsymbol{T}^c \in \mathbb{R}^{N_c \times D^m}$.

3.3 Sentimental Prior Knowledge Incorporation

As we know, human beings can still describe images or videos with proper emotions even when some visual emotion cues are not accurately perceived, owing to their background knowledge about sentiments of objects and their relationships in real-world scenarios. This inspired us to explore sentimental commonsense from the sentimental corpus as prior knowledge and to incorporate it into captioning. As an augmentation of the global sentiment information that is directly extracted from the image or video by the sentiment analysis described in Sect. 3.2.2, the sentimental elements inferred from the prior knowledge provide more detailed emotion information of local objects. For example, the pair $< baby - cute >$ obtained from the sentimental corpus expresses that people often praise babies with cute and the pair $< fly - disgusting >$ reflects the aversion to flies. Therefore, it is beneficial to use the prior knowledge of "object-sentiment" word pairs as the augmented textual sentiment information for generating rich and emotional captions.

We first employ the NLTK tool (Bird et al., 2009) to mark the part of speech of sentences, and extract the noun-adjectives pairs as candidates of the object-sentiment word pairs. Then an importance evaluation method is proposed to filter out the noun-adjectives pairs that have nothing to do with sentiment. It consists of two parts. The first part measures the closeness of a noun-adjective pair, given by

$$CL_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,k}}, \tag{7}$$

where $n_{i,j}$ represents the number of occurrences of noun i and adjective j. The second part calculates the emotionality of an adjective, given by

$$EM_j = \text{MAX}(\frac{n_j^s}{n_j}),\ s \in \{positive, negative\}, \tag{8}$$

where n_j^s is the number of adjective j in the corpus with sentiment s, n_j is the number of j in all corpus, and MAX$(\cdot)$ means to traverse all sentiments and then take the maximum value. Finally, the importance score of a noun-adjective pair is calculated by

$$IS_{i,j} = CL_{i,j} \times EM_j. \tag{9}$$

Finally, the correspondence with a lower score is directly discarded.

We extract 2651 object-sentiment word pairs from the sentimental corpus (excluding neutral sentiment corpus) as the sentimental prior knowledge where each object is associated with six sentiment words on average. Table 1 shows several examples. When captioning, the top N_s sentiment words are selected from the sentimental prior knowledge by ranking the IS scores of sentiment words paired with the concept words extracted by the concept detector in Sect. 3.2.3. These selected sentiment words are the textual sentiment information into the encoder, represented by the word embedding features $\boldsymbol{T}^s \in \mathbb{R}^{N_s \times D^m}$.

3.4 Multi-Head Transformer Encoding

After obtaining the content and sentiment information from visual and textual modalities, we encode these information. We propose a sentiment multi-head encoder to encode the visual sentiment information, where one head deals with one sentiment category. The other information of textual sentiment, visual content and textual content is separately encoded by the standard Transformer encoder (Vaswani et al., 2017).

Table 1 Examples of sentimental prior knowledge

Object	Sentiment Words
Man	Nice, happy, great, cool, bad, lonely
Woman	Beautiful, pretty, abused, great, crazy
Food	Tasty, bad, healthy, delicious, awesome, fancy, rotten
Street	Busy, lonely, nice, charming, dirty, calm, dead, poor
Dog	Funny, adorable, lazy, beautiful, stupid, scared, friendly, dangerous

3.4.1 Sentiment Multi-head Encoding

Since different sentiments have their distinctive visual representations, we design an individual head module for each sentiment. Each head consists of four convolutional layers with ReLU function and a fully connected layer followed by layer normalization, and takes the visual sentiment information (i.e. grid features $\boldsymbol{V}^s$) as input. Given the image sentiment class s_i that is predicted by the sentiment analyzer, the corresponding head module Head_{s_i} is formulated by

$$\begin{aligned}\boldsymbol{E}^{vs} &= \text{Head}_{s_i}(\boldsymbol{V}^s)\\ &= \text{LayerNorm}(\text{FC}(\text{Flatten}(\text{conv}_{\times 4}(\boldsymbol{V}^s)))),\end{aligned} \tag{10}$$

where $\boldsymbol{E}^{vs} \in \mathbb{R}^{M^{vs} \times D^m}$ is the encoded visual sentiment features.

3.4.2 Other Information Encoding

The visual content information (i.e., object region features $\boldsymbol{V}^c$) and the textual content information (i.e., concept word embedding features $\boldsymbol{T}^c$) are both decoded by the standard Transformer encoder, shown as the content Transformer encoder in Fig. 2. The augmented textual sentiment information (i.e., sentiment word embedding features $\boldsymbol{T}^s$) is also encoded by the standard Transformer encoder, shown as the sentimental prior knowledge Transformer encoder in Fig. 2.

In practice, following GPT-2 (Radford et al., 2019), the layer normalization is moved to the input of multi-head attention block and feed forward block, and an additional layer normalization is set after the final block. The formulation of the multi-head attention block is given by

$$\begin{aligned}&\text{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Concat}(\text{head}_1, \cdots, \text{head}_h)\boldsymbol{W}^O\\ &\text{head}_i = \text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V),\\ &\text{Attention}(\boldsymbol{Q}', \boldsymbol{K}', \boldsymbol{V}') = \text{Softmax}(\frac{\boldsymbol{Q}'\boldsymbol{K}'^T}{\sqrt{D^m}})\boldsymbol{V}',\end{aligned} \tag{11}$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$ represent query, key, and value in Transformer, respectively. Here, they are actually the input features. $\boldsymbol{W}^O \in \mathbb{R}^{D^m \times D^m}$, $\boldsymbol{W}_i^Q \in \mathbb{R}^{D^m \times D^k}$, $\boldsymbol{W}_i^K \in \mathbb{R}^{D^m \times D^k}$ and $\boldsymbol{W}_i^V \in \mathbb{R}^{D^m \times D^k}$ ($D^m = D^k \times h$) are parameter matrices. The encoded visual content features are denoted as $\boldsymbol{E}^{vc}$, the encoded textual content features as $\boldsymbol{E}^{sc}$, and the encoded textual sentiment features as $\boldsymbol{E}^{ss}$.

3.5 Decomposed Multimodal Decoding

The decoder takes the sum of the previously generated word embedding, positional encoding and sentiment class embedding as input. The sentiment class is predicted by the sentiment analyzer and used to enable the decoder to decide which sentiment information to be focused on. The output is the conditional distribution over all possible words calculated by a linear transformation and a softmax operation. The decoder consists of a masked multi-head attention block, a decomposed multimodal attention block and a feed-forward network block, where the masked multi-head attention block and the feed-forward network block are the same as the standard Transformer decoder (Vaswani et al., 2017).

In the decoding process of generating captions word by word, in order to attend the important elements of each encoded feature (i.e., $\boldsymbol{E}^{vc}$, $\boldsymbol{E}^{vs}$, $\boldsymbol{E}^{sc}$ and $\boldsymbol{E}^{ss}$), we design the decomposed multimodal attention block that decomposes the attention layer in the middle of the Transformer decoder into four parts: a visual content attention module, a visual sentiment attention module, a textual content attention module and a textual sentiment attention module, as illustrated in Fig. 5. In each decomposed attention module, we adopt a multi-head attention block, where the input encoded features are treated as the key and value, and the output from the masked multi-head attention block is taken as the query. Take the visual sentiment attention module as an example to illustrate. At the current time step, both key and value are the encoded visual sentiment features, i.e, $\boldsymbol{E}^{vs} \in \mathbb{R}^{M^{vs} \times D^m}$, and the query is the output from the masked multi-head attention block, i.e., $\boldsymbol{q} \in \mathbb{R}^{1 \times D^m}$. The visual sentiment attention module is formulated by

$$\boldsymbol{v}^s = \text{MultiHead}(\boldsymbol{E}^{vs}, \boldsymbol{E}^{vs}, \boldsymbol{q}), \tag{12}$$

where $\boldsymbol{v}^s \in \mathbb{R}^{1 \times D^m}$ is the output visual sentiment vector. In the same way, the remaining visual content attention module, textual content attention module and textual sentiment attention module process the encoded visual content features $\boldsymbol{E}^{vc}$,

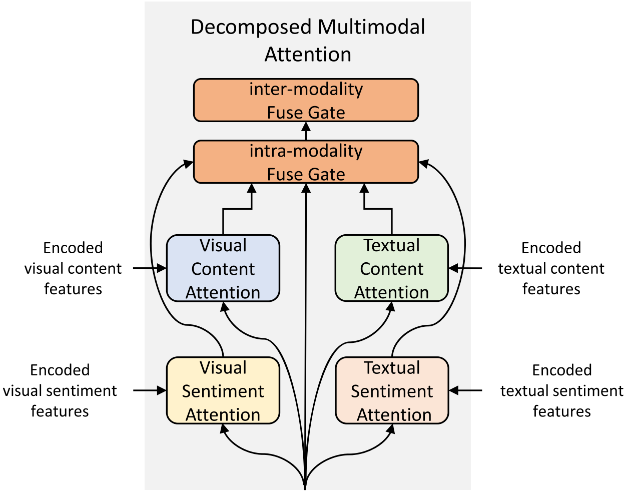

Decomposed Multimodal Attention
inter-modality Fuse Gate
intra-modality Fuse Gate
Visual Content Attention
Textual Content Attention
Visual Sentiment Attention
Textual Sentiment Attention
Encoded visual content features
Encoded textual content features
Encoded visual sentiment features
Encoded textual sentiment features

Fig. 5 Decomposed multimodal attention block

the encoded textual content features $\boldsymbol{E}^{sc}$ and the encoded textual sentiment features $\boldsymbol{E}^{ss}$ to output the visual content vector $\boldsymbol{v}^c$, the textual content vector $\boldsymbol{t}^c$, and the textual sentiment vector $\boldsymbol{t}^s$, respectively.

Then the attended feature vectors (i.e., $\boldsymbol{v}^s$, $\boldsymbol{v}^c$, $\boldsymbol{t}^s$ and $\boldsymbol{t}^c$) from the four decomposed attention modules are fused through an intra- and inter-modality fusion mechanism.

3.5.1 Intra-modality Fusion

The content and sentiment features of the same modality are fused by attention. Taking the visual modality for example, at the current time step, the query vector $\boldsymbol{q}$ (i.e., output of the masked multi-head attention block) is used to determine whether the visual content vector $\boldsymbol{v}^c$ or the visual sentiment vector $\boldsymbol{v}^s$ should be paid attention to, given by

$$\begin{aligned}
&a^c = \boldsymbol{v}^c \cdot \boldsymbol{q}^T, \\
&a^s = \boldsymbol{v}^s \cdot \boldsymbol{q}^T, \\
&[\alpha^c, \alpha^s] = \mathrm{softmax}([a^c, a^s]), \\
&\boldsymbol{f}^v = \alpha^c \boldsymbol{v}^c + \alpha^s \boldsymbol{v}^s,
\end{aligned} \tag{13}$$

where $\boldsymbol{f}^v \in \mathbb{R}^{1\times D^m}$ represents the fused visual feature vector. The fused textual feature vector $\boldsymbol{f}^s$ is obtained by $\boldsymbol{t}^s$ and $\boldsymbol{t}^c$ in a similar way.

3.5.2 Inter-modality Fusion

Different modalities are merged through a mean operation and then the residual connection is performed with the query vector $\boldsymbol{q}$, formulated by

$$\boldsymbol{f}^o = \boldsymbol{q} + \frac{\boldsymbol{f}^v + \boldsymbol{f}^s}{2}. \tag{14}$$

3.6 Model Training

3.6.1 Pre-training

At the pre-training stage, the captioning model is trained using the paired images or videos and factual sentences. In order to enable the model to learn how to incorporate sentimental elements into sentences, we propose a new regularization term defined by the reconstruction of sentimental sentences in the corpus using the Transformer decoder described in Sect. 3.5. When reconstructing sentimental sentences, the nouns and verbs are first extracted from the sentimental sentences as the concept words, and the sentiment words are obtained from the prior knowledge. Then they are encoded as the input into the decoder. During decoding, the output of decomposed multimodal attention module in Eq. 14 only consists of texture fusion vector: $\boldsymbol{f}^o = \boldsymbol{q} + \boldsymbol{f}^s$.

The regularization term is actually a cross-entropy loss using the sentimental sentences, defined by

$$\mathcal{L}_{re} = -\frac{1}{T}\sum_{t=1}^{T} \log p_\theta(y_t^s | y_{1:t-1}^s), \tag{15}$$

where $p_\theta(y_t^s|y_{1:t-1}^s)$ denotes the predicted probability of the ground-truth word y_t^s given the previous word sequence $y_{1:t-1}^s$. The final loss for pre-training is given by

$$\begin{aligned}
&\mathcal{L}_{Pt} = \mathcal{L}_{XE} + \mathcal{L}_{re}, \\
&\mathcal{L}_{XE} = -\frac{1}{T}\sum_{t=1}^{T} \log p_\theta(y_t^f | I, y_{1:t-1}^f),
\end{aligned} \tag{16}$$

where $\mathcal{L}_{XE}$ is the cross-entropy loss using the pairs of images or videos and factual sentences, and $p_\theta(y_t^f|I, y_{1:t-1}^f)$ denotes the predicted probability of the ground-truth word y_t^f given the image I and the previous word sequence $y_{1:t-1}^f$.

3.6.2 Fine-Tuning

At the fine-tuning stage, we introduce reinforcement learning to further improve the quality of the generated captions. Our Senti-Transformer can be viewed as an agent that interacts with an external environment. Here, the environment is the extracted multimodal information, sentimental prior knowledge, and the previously generated words. The caption generation process is formulated as a markov decision process (MDP), denoted as $\mathcal{M} =< S, A, P, R, \gamma >$, where S is a set of states, A is a set of actions, P is the state transition probability, R is the reward function and γ is a discount factor.

Specifically, at the t-th time step, given the current state $s_t \in S$, the agent takes an action $a_t \in A$ (i.e., generates a word

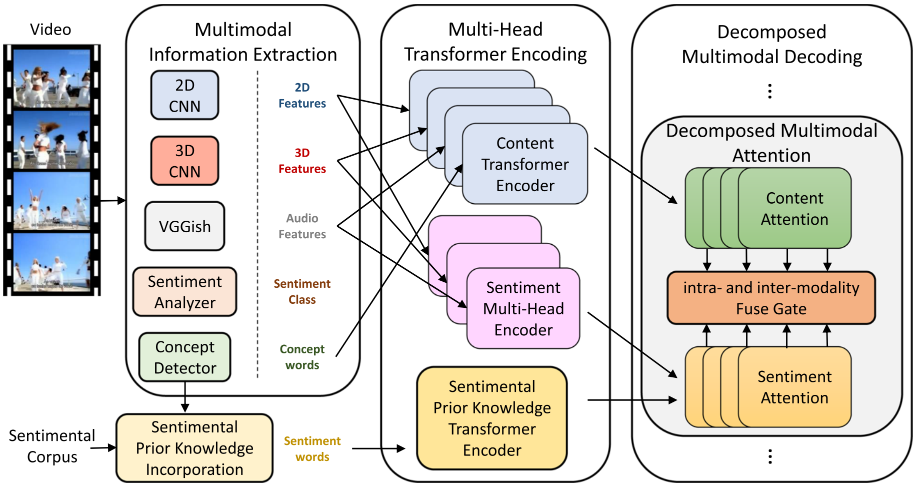

Video
Multimodal Information Extraction
2D CNN
3D CNN
VGGish
Sentiment Analyzer
Concept Detector
2D Features
3D Features
Audio Features
Sentiment Class
Concept words
Sentimental Corpus
Sentimental Prior Knowledge Incorporation
Sentiment words
Multi-Head Transformer Encoding
Content Transformer Encoder
Sentiment Multi-Head Encoder
Sentimental Prior Knowledge Transformer Encoder
Decomposed Multimodal Decoding
⋮
Decomposed Multimodal Attention
Content Attention
intra- and inter-modality Fuse Gate
Sentiment Attention
⋮

Fig. 6 Multimodal Senti-Transformer for sentimental video captioning. The ellipses of the decomposed multimodal decoding part indicate the same as in Fig. 2

w_t) according to the policy $\pi_\theta(a|s_t)$. The policy $\pi_\theta(a|s_t)$ is defined as the conditional probability distribution of all the actions $a \in A$ (i.e., all the words) given the state s_t, and is implemented by the multi-head Transformer encoding module and the decomposed multimodal decoding module, where θ denotes the parameters of the two modules. The current state s_t includes the extracted multimodal information MI, the sentiment prior knowledge PK, and the previously generated words PW_t (i.e. $\{w_1, w_2, \ldots, w_{t-1}\}$), denoted as $s_t = \{MI, PK, PW_t\}$.

After the agent performs the action a_t, the action is appended to the state s_t to form a new state $s_{t+1} = \{MI, PK, PW_{t+1}\}$, that is, the probability of state transition is always equal to 1, denoted as $P(s_{t+1}|s_t, a_t) \equiv 1$.

The agent observes a discounted future reward $R(s_t, a_t)$ from the environment after the state transition, denoted as

$$R(s_t, a_t) = \sum_{l=t}^{T} \gamma^{l-t} r_l(s_l, a_l), \tag{17}$$

where $r_l(s_l, a_l)$ is the reward obtained at the l-th time step, and T represents the length of the whole generated sentence. Since the reward can only be calculated after the whole sentence is generated in this task, the discount factor γ is set to 1 and $r_l(s_l, a_l)$ is set to 0 before generating the sentence end token $< EOS >$, given by

$$r_l(s_l, a_l) = \begin{cases} 0, & l < T \\ r_T(s_T, a_T), & l = T \end{cases}, \tag{18}$$

Since the reward obtained through the interaction between the agent and the environment in this task only depends on the performed actions, $r_T(s_T, a_T)$ can be simplified to $r(y^s)$, where $y^s = \{a_1, a_2, \ldots, a_T\}$ represents all the actions performed by the agent (i.e., the generated whole sentence). Therefore, $R(s_t, a_t)$ is derived as

$$R(s_t, a_t) = r(y^s). \tag{19}$$

The goal of reinforcement learning is to minimize the negative expected discounted reward, and the loss $\mathcal{L}_R$ of reinforcement learning is formulated as

$$\begin{aligned} \mathcal{L}_R &= -\sum_{t}^{T} \mathbb{E}_{a_t \sim \pi_\theta}[R(s_t, a_t)] \\ &= -\sum_{t}^{T} \mathbb{E}_{a_t \sim \pi_\theta}[r(y^s)]. \end{aligned} \tag{20}$$

By using the REINFORCE algorithm (Williams, 1992), the gradient of $\mathcal{L}_R$ is computed as

$$\nabla_\theta \mathcal{L}_R(\theta) = -\sum_{t}^{T} \mathbb{E}_{a_t \sim \pi_\theta}[r(y^s)\nabla_\theta \log \pi_\theta(a_t)]. \tag{21}$$

In practice, the gradient is approximated using a single Monte-Carlo sampling from the policy π_θ:

$$\nabla_\theta \mathcal{L}_R(\theta) \approx -\sum_{t}^{T} r(y^s)\nabla_\theta \log \pi_\theta(a_t). \tag{22}$$

Following (Rennie et al., 2017), we use the reward of a greedily-decoded sentence $\hat{y}$ as the baseline to reduce the variance of the reward. The gradient is approximated by

$$\nabla_\theta \mathcal{L}_R(\theta) \approx -\sum_{t}^{T} r(y^s, \hat{y})\nabla_\theta \log \pi_\theta(a_t). \tag{23}$$

The entire reward $r(y^s, \hat{y})$ consists of a content reward r_{con} and a newly proposed sentimental reward r_{sen}:

$$r(y^s, \hat{y}) = r_{con} + r_{sen}. \tag{24}$$

The content reward r_{con} is given by

$$r_{con} = \lambda_{con}(\text{CIDEr}(y^s) - \text{CIDEr}(\hat{y})), \tag{25}$$

where $\text{CIDEr}(y)$ is the CIDEr (Vedantam et al., 2015) score of the sentence y, which is commonly used to evaluate captioning quality. The new sentimental reward r_{sen} is calculated by the evaluations of sentence sentiment and language fluency. A classifier is designed to evaluate the sentence sentiment and consists of a LSTM and a fully connected layer. A language model is designed to evaluate the sentence sentiment and fluency, which is trained separately using the corpus of each sentiment. So the sentimental reward r_{sen} is given by

$$\begin{aligned} r_{sen} &= \lambda_{cls} r_{cls} + \lambda_{ppl} r_{ppl}, \\ r_{cls} &= \mathbb{I}_{(s_i = s_{cls})}, \\ r_{ppl} &= \text{sign}(\text{ppl}(\hat{y}) - \text{ppl}(y^s)), \end{aligned} \tag{26}$$

where r_{cls} is the classification reward provided by the classifier, and r_{ppl} is the perplexity reward provided by the language model. Specifically, for r_{cls}, if the image sentiment class s_i is consistent with the sentence sentiment classification result s_{cls}, $\mathbb{I}_{(s_i = s_{cls})}$ takes 1 otherwise it takes -1. For r_{ppl}, $\text{sign}(\cdot)$ represents the sign function, and the $\text{ppl}(y)$ represents the perplexity score calculated by the language model trained using the sentences with the sentiment class s_i.

3.7 Sentimental Video Captioning

In sentimental video captioning, the motion and audio information play a vital role in understanding the video content and sentiment, so we modify three parts of the proposed Senti-Transformer to allow more multimodal input for captioning, including the multimodal information extraction, multi-head Transformer encoding, and decomposed multimodal attention, as shown in Fig. 6.

3.7.1 Multimodal Information Extraction

We use 2D CNN (He et al., 2016) and 3D CNN (Hara et al., 2018) to extract static and dynamic visual features of the video, and VGGish (Hershey et al., 2017) to extract audio features. The video concept detector consists of four fully connected layers with ReLU activation functions and a sigmoid function, and takes the concatenation of global visual and audio features (generated by averaging the feature set) as input. The sentiment analyzer first encodes the visual and audio features through a Transformer encoder, then averages the encoded features, and finally performs sentiment classification via a fully connected layer.

3.7.2 Multi-head Transformer Encoding

For the sentiment information, we design a multi-head Transformer encoder to encode the visual and audio features and design a standard Transformer encoder to encode the textual features, shown as the sentiment multi-head encoder in Fig. 6. For the content information, we use a standard Transformer encoder for each modality, shown as the content Transformer encoder in Fig. 6. The sentimental prior knowledge Transformer encoder in Fig. 6 is the same as that in sentimental image captioning.

3.7.3 Decomposed Multimodal Attention

The attention layer in the middle of the decoder is decomposed into four content attention modules and four sentiment attention modules. The intra-modality fusion is the same as the image captioning model and the inter-modality fusion is given by

$$\boldsymbol{f}^o = \boldsymbol{q} + \frac{\boldsymbol{f}^{2d} + \boldsymbol{f}^{3d} + \boldsymbol{f}^a + \boldsymbol{f}^s}{4}, \tag{27}$$

where $\boldsymbol{f}^{2d}$, $\boldsymbol{f}^{3d}$, $\boldsymbol{f}^a$ and $\boldsymbol{f}^s$ respectively represent the fused features of 2D visual, 3D visual, audio, and textual modalities via the intra-modality fusion.

4 Experiments

4.1 Datasets

4.1.1 Captioning dataset

We choose COCO (Lin et al., 2014) dataset for image captioning and use the Karpathy splits (Karpathy and Fei-Fei, 2015) for the model validation and offline evaluation, where 113, 287 and 5000 images with five factual captions are for training and validation, respectively, and 5000 images are for test. We choose MSR-VTT (Xu et al., 2016) dataset for video captioning, where 6513 and 497 videos with twenty factual captions are for training and validation, respectively, and 2990 videos are for test.

4.1.2 Sentiment Dataset

As shown in Table 2, we use the EmotionROI (Peng et al., 2016), ArtPhoto (Machajdik and Hanbury, 2010), Twitter I (You et al., 2015) and Twitter II (Borth et al., 2013) datasets of image sentiment analysis field for training the sentiment

Table 2 Image sentiment dataset

Dataset	#positive	#negative	#neutral
EmotionROI	330	1320	330
ArtPhoto	206	428	172
Twitter I	689	427	153
Twitter II	470	133	0
Sum	1695	2308	655

analyzer. For the Twitter I dataset, we use the “At Least Four Agree” result which indicates that at least 4 AMT workers gave the same sentiment class for a given image. The EmotionROI and ArtPhoto datasets contain multiple sentiments. In this paper, we only focus on the positive and negative sentiments, so we reclassify images into positive, negative and neutral categories. For the video sentiment dataset, four annotators perform sentimental annotations on some videos of the MSR-VTT dataset, and we adopt the results agreed by at least three of them as the new video sentiment dataset (called Senti-MSR-VTT) in which the numbers of positive, negative and neutral videos are 170, 63 and 220.

4.1.3 Sentimental Corpus

We employ the SentiCap (Mathews et al., 2016) dataset as the sentimental corpus, which includes 4892 positive sentences and 3977 negative sentences. For the neutral category, we select the sentences from visual captioning dataset that do not contain sentiment words.

4.2 Implementation Details

The hyperparameter λ_{ss} of the threshold filtering method in Sect. 3.2.2 is set to 0.7. In Sect. 3.2.3, the most common 2000 concept words (i.e. M=2000) are selected from the captioning dataset to form the vocabulary and the number of concept words N_c is set to 5. The number of sentiment words N_s in Sect. 3.3 is set to 10. The decoder and all encoders are composed of a stack of 4 same layers (1 layer for encoders in video captioning model). We employ eight heads (i.e., $h = 8$) in multi-head attention block and the parameter dimension of the attention layers D^m is set to 512 in Eq. 11. The hidden layer dimension of the feed-forward network is set to 2048. The embedding dimensions of words and sentimental labels are both 512. The hyperparameters λ_{cls} and λ_{ppl} in Eq. 26 and the λ_{con} in Eq. 25 are set to 0.5, 0.1 and 1, respectively. In the pre-training stage, the learning rate is set to 4×10^{-4}, and in the fine-tuning stage, the learning rate is set to 4×10^{-5}. We employ the Adam optimizer (Kingma and Ba, 2015) for training.

4.3 Visual Sentiment Analysis Performance

To evaluate the performance of the image sentiment analyzer, we collect a new image sentiment dataset based on the COCO dataset, called Senti-COCO, in which the numbers of positive, negative and neutral images are 81, 35, and 137. The experimental result is that if the sentiment analyzer is directly used (that is, the hyperparameter λ_{ss} is set to 0), the analyzer’s accuracy is only 62.9%, but when the λ_{ss} is set to 0.7, the accuracy is improved to 65.6%. This shows that threshold filtering on sentiment scores is effective in solving the domain gap between the images in the image captioning field and the image sentiment analysis field.

To evaluate the performance of the video sentiment analyzer, we use the newly collected Senti-MSR-VTT dataset to train the video sentiment analyzer, and the analyzer achieves 68.3% accuracy.

4.4 Evaluation Metrics

We evaluate the quality of the captions generated by our model from two aspects: content relevance and sentiment consistency. To verify the content relevance, we report the widely used automatic evaluation metrics, i.e., BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014) and CIDEr (Vedantam et al., 2015). These metrics are measured based on n-gram overlap with factual ground-truth captions as references. However, for the sentimental visual captioning, we should also take into account the sentimentality of generated sentences in word selection, so these content metrics are not well suited for sentimental purpose.

We measure the sentiment consistency by the sentiment classification accuracy (denoted as cls.) and the average perplexity (denoted as ppl.). The accuracy of sentiment classification is defined as the percentage of generated sentences that correctly express the sentiment of the image or video. We train the sentence sentiment classifier using the sentimental corpus and achieve 99% accuracy. The average perplexity is calculated by the trained language model corresponding to the sentiment of the generated sentence. Specifically, for each sentiment, we use the SRILM toolkit (Stolcke, 2002) to train a tri-gram based statistical language model on the corresponding sentimental corpus. The lower score indicates that the generated sentences are more fluent and more in line with the corresponding linguistic style.

4.5 Comparison Methods

In order to evaluate the performance of our Senti-Transformer on describing visual content, we compare with several factual visual captioning methods, including Up-Down (Anderson et al., 2018), M2 Transformer (Cornia et al., 2020) and DLCT (Luo et al., 2021) for image captioning, and Pick-

Table 3 Comparison with our preliminary InSenti-Cap and the stylized image captioning methods on the COCO dataset

Sentiment	Method	Bleu-1	Bleu-3	METEOR	CIDEr	ppl.(↓)	cls.(%)
Positive	MemCap*	50.8	17.1	16.6	54.4	13.0	99.8
	ERG(Up-Down)*	52.4	24.4	18.1	77.7	10.3	**100**
	ERG(VinVL)*	53.4	23.4	18.0	75.9	12.4	**100**
	MSCap	46.9	16.2	16.8	55.3	19.6	92.5
	MemCap	51.1	17.0	16.6	52.8	18.1	96.1
	InSenti-Cap	59.7	25.3	20.9	61.3	13.0	98.5
	Our Senti-Transformer	**65.1**	**32.8**	**22.5**	**87.3**	**9.9**	**99.3**
Negative	MemCap*	48.7	19.6	15.8	60.6	14.6	93.1
	ERG(Up-Down)*	52.6	21.6	18.0	68.3	**8.5**	**100**
	ERG(VinVL)*	53.1	21.2	18.6	70.0	13.7	**100**
	MSCap	45.5	15.4	16.2	51.6	19.2	93.4
	MemCap	49.2	18.1	15.7	59.4	18.9	98.9
	InSenti-Cap	59.1	24.3	19.4	53.3	12.3	95.5
	Our Senti-Transformer	**68.6**	**38.9**	**23.0**	**94.6**	**9.9**	**99.8**
Neutral	InSenti-Cap	73.5	41.2	24.7	97.5	8.4	98.9
	Our Senti-Transformer	**74.7**	**43.2**	**25.7**	**108.3**	**6.8**	**99.9**
Overall	Our Senti-Transformer	71.6	40.2	24.5	102.8	8.7	99.7

The best results are highlighted in bold

*Indicates that a model can only generate captions of one sentiment, and others can generate multiple sentiment sentences using a single model. “overall” in the sentiment column represents the overall model performance over all sentiment classes. For ppl metric, the smaller value is better, and for other metrics, the larger value is better

Net (Chen et al., 2018), STG-KD (Pan et al., 2020) and NACF (Yang et al., 2021) for video captioning.

In order to evaluate the quality of generated sentimental captions, we compare with our previous work InSenti-Cap (Li et al., 2021b) and the state-of-the-art methods of stylized image captioning, including MSCap (Guo et al., 2019), MemCap (Zhao et al., 2020b) and ERG (Li et al., 2021a).

- InSenti-Cap is a sentimental image captioning method, which is based on the LSTM with an attention mechanism.
- MSCap uses an adversarial learning network to generate captions of multiple specified styles for a image.
- MemCap explicitly encodes the knowledge about linguistic styles with memory mechanism.
- ERG performs data augmentation for the small-scale paired stylized data, and then trains the captioning model, which achieves good results based on the methods of Up-Down (Anderson et al., 2018) and VinVL (Zhang et al., 2021).

4.6 Comparison Results

4.6.1 Sentimental Image Captioning

As shown in Table 3, the models with the symbol * are trained under single-style setting, where a model is trained for each style, and others are multi-style captioning methods, i.e. a model is trained to generate sentences in multiple styles. Our Senti-Transformer generates multiple sentimental captions using a single model and utilizes the image (or video) sentiment detection result as the corresponding sentiment.

We have the observations as follows:

- Compared with the multi-style captioning (below the dotted line of each sentiment), our method dramatically surpasses all previous methods in both content and sentiment metrics, which illustrates that our model can make an effective trade-off between content and sentiment.
- Compared with the single-style captioning (above the dotted line), most metrics of our method improves significantly, validating the superiority of our method on capturing multiple sentimental knowledge for captioning.
- Compared with our preliminary InSentiCap, our Senti-Transformer achieves significant improvements on all the metrics, clearly indicating that the newly added parts of Senti-Transformer, i.e., the multi-head Transformer encoding part and the decomposed multimodal decoding part, benefits a lot to handling multimodal information.

4.6.2 Sentimental Video Captioning

The results of sentimental video captioning on the MSR-VTT dataset are shown in Table 4. Compared with the sentimental

Table 4 Sentimental video captioning results on the MSR-VTT dataset

Method	Sentiment	Bleu-1	Bleu-3	METEOR	CIDEr	ppl.(↓)	cls.(%)
Our Senti-Transformer	positive	67.7	29.2	21.1	35.3	12.3	100
	negative	60.5	23.9	17.2	20.5	8.9	99.9
	neutral	80.0	58.2	29.1	48.1	5.2	100
	Overall	71.0	39.7	23.3	39.5	9.0	100

"overall" in the sentiment column represents the overall model performance over all sentiment classes. For ppl metric, the smaller value is better, and for other metrics, the larger value is better

Table 5 Comparison with the factual image captioning methods on the COCO dataset

Method	Bleu-4	METEOR	CIDEr
Up-Down	36.3	27.7	120.1
M2 Transformer	39.1	29.2	131.2
DLCT	39.8	29.5	133.8
Ours (neutral)	31.8	25.7	108.3

For fair comparison, only the neutral sentiment results of our Senti-Transformer are reported

Table 6 Comparison with the factual video captioning methods on the MSR-VTT dataset

Method	Bleu-4	METEOR	CIDEr
PickNet	41.3	27.7	44.1
STG-KD	40.5	28.3	47.1
NACF	42.0	28.7	51.4
Ours (neutral)	46.5	29.1	48.1

For fair comparison, only the neutral sentiment results of our Senti-Transformer are reported

image captioning, the performance of sentimental video captioning have the same trend: the results of neutral sentiment generally exceed the results of positive and negative sentiments, probably due to that the fact captions in captioning datasets are mostly neutral, so the model is easier to learn this kind of knowledge from paired data.

4.6.3 Comparison with the Factual Visual Captioning

Since the factual visual captioning task trains the model using the pairs of image or video and factual caption and generates neutral captions, we only report the neutral sentiment results of our method for fair comparison. Table 5 and Table 6 show the comparison results on the COCO dataset and the MSR-VTT dataset, respectively. From the results, we observe that our Senti-Transformer works worse than the factual captioning method, but still achieves satisfactory results. The reason is that our model handles not only the factual captioning, but also the sentimental descriptions of positive and negative sentiments. We also observe that for video captioning, our Senti-Transformer achieves better or comparable results, indicating the superiority of our model on handling multimodal and sentimental information.

4.7 Ablation Study

In order to evaluate the effectiveness of the various modules and the training strategy of Senti-Transformer, we introduce several variants of it for comparison, as follows:

- w/o cls: in the fine-tuning stage, the cls reward function in Eq. 26 is removed, that is, the hyperparameter λ_{cls} is set to 0.
- w/o ppl: in the fine-tuning stage, the ppl reward function in Eq. 26 is removed, that is, the hyperparameter λ_{ppl} is set to 0.
- w/o reg: throughout the training, the regularization term $\mathcal{L}_{re}$ in Eq. 16 is removed.
- w/o pri: the prior knowledge (i.e. sentiment words in Sect. 3.3) and its encoder and attention module are removed. In other words, the textual features in decomposed multimodal attention module only contain content features, while sentiment features are removed.
- w/o mhe: the sentiment multi-head encoder in Sect. 3.4.1 only retains one head that is used to encode features of all sentiments.
- w/o dma: the decomposed multimodal attention module 3.5 is replaced to the multi-head attention module in Transformer, and all the input features are concatenated together as the key and value in Eq. 11.
- w/o audio: in the sentimental video captioning model in Sect. 3.7, the audio information and its encoder and attention module are removed.

The results of ablation studies on the COCO and MSR-VTT datasets are shown in Table 7 and Table 8, respectively. It can be observed from the results that no matter which part is removed, the performance degrades on most metrics, especially the sentiment consistency (ppl. and cls.), indicating the effectiveness of each part and our method can generate smooth and sentimental captions. It is also interesting to observe that our Senti-Transformer sometimes

Table 7 Ablation studies on the COCO dataset

Sentiment	Method	Bleu-3	CIDEr	ppl.(↓)	cls.(%)
Positive	w/o cls	**35.3**	79.0	10.8	48.0
	w/o ppl	33.7	**88.4**	12.4	98.7
	w/o reg	32.5	84.5	11.2	97.9
	w/o pri	33.4	85.2	10.4	99.0
	w/o mhe	33.2	86.6	11.7	97.6
	w/o dma	32.8	85.9	10.5	98.2
	Ours (all)	32.8	87.3	**9.9**	**99.3**
Negative	w/o cls	33.7	79.7	13.1	81.6
	w/o ppl	38.7	93.8	14.3	98.8
	w/o reg	36.7	89.7	11.5	98.3
	w/o pri	37.5	92.3	11.4	97.7
	w/o mhe	38.3	93.8	11.4	96.4
	w/o dma	38.8	94.1	11.2	98.4
	Ours (all)	**38.9**	**94.6**	**9.9**	**99.8**
Neutral	w/o cls	38.6	90.7	**6.4**	**99.9**
	w/o ppl	43.1	106.7	8.8	99.4
	w/o reg	**43.2**	106.6	6.8	99.7
	w/o pri	43.0	106.3	7.1	99.5
	w/o mhe	42.0	104.6	7.6	99.6
	w/o dma	**43.2**	106.5	7.2	**99.9**
	Ours (all)	**43.2**	**108.3**	6.8	**99.9**

The best results are highlighted in bold

Table 8 Ablation studies on the MSR-VTT dataset

Sentiment	Method	Bleu-3	CIDEr	ppl.(↓)	cls.(%)
Positive	w/o audio	28.9	25.0	22.9	97.8
	w/o cls	**32.0**	25.0	29.5	74.9
	w/o ppl	29.9	31.1	38.8	98.6
	w/o reg	29.0	31.4	16.3	98.2
	w/o pri	25.6	30.1	16.7	98.9
	w/o mhe	28.2	34.3	15.4	99.1
	w/o dma	27.4	30.3	14.5	99.5
	Ours (all)	29.2	**35.3**	**12.3**	**100**
Negative	w/o audio	25.9	16.7	11.9	96.2
	w/o cls	**30.1**	18.2	18.5	53.9
	w/o ppl	23.3	14.6	15.1	97.3
	w/o reg	23.5	17.7	11.9	98.8
	w/o pri	18.7	14.9	10.9	99.2
	w/o mhe	20.1	18.5	10.4	99.0
	w/o dma	22.8	17.7	10.5	99.4
	Ours (all)	23.9	**20.5**	**8.9**	**99.9**
Neutral	w/o audio	51.5	42.8	5.7	99.9
	w/o cls	50.7	39.6	7.8	98.8
	w/o ppl	53.3	43.4	6.0	**100**
	w/o reg	54.7	43.6	5.7	99.8
	w/o pri	57.4	46.0	5.4	**100**
	w/o mhe	56.5	46.4	**5.2**	99.8
	w/o dma	55.6	46.9	5.5	**100**
	Ours (all)	**58.2**	**48.1**	**5.2**	**100**

The best results are highlighted in bold

perform worse on the content relevance metric (especially the Bleu metric), but still achieves the best results on the sentiment consistency metrics, validating that it can make a good trade-off between content and sentiment. That is to say, our method trades a slight loss of content for the the significant improvement of sentiment. Compared with “w/o ppl”, our Senti-Transformer achieves better results on the sentiment metrics (i.e. ppl. and cls.), which indicates the effectiveness of the ppl reward function and further evaluates the contribution of the Senti-Transformer on improving the sentimentality and fluency of the generated captions over the preliminary InSentiCap.

4.8 Analysis of Sentiment Information Sources

To analyze the source of sentimental information in generating captions, we show the normalized fusion weights (i.e., the attention scores in Eq. 13) of different sentiment features in the decoding phase in Fig. 7. For sentimental image captioning on the COCO dataset, the sentiment features include the attended visual sentiment feature v_s extracted from image grids and the attended textual sentiment feature t_s extracted from prior knowledge. For video captioning on the MSR-VTT dataset, besides the visual and textual sentiment features, the sentiment features also include the attended audio sentiment feature. It is interesting to observe that the visual inputs and the prior knowledge (additional audio inputs for video captioning) both contribute to generating the sentimental sentences. We also observe that the prior knowledge has less effect on the sentimental video captioning than the sentimental image captioning, probably due to that the 3D visual features and audio features are more important to videos and convey more sentiment information.

4.9 Human Evaluation

We perform human evaluations to evaluate the content and sentiment of the generated captions. We compare our Senti-Transformer with the previous InSenti-Cap model. Specifically, we first randomly select 50 images for each sentiment, then generate their captions by the Senti-Transformer and InSenti-Cap models, respectively, and finally invited 52 volunteers from different majors and different grades to conduct quality assessment. The content relevance is rated from 1 (unrelated) to 3 (very related), and the sentiment consistency is rated from 1 (inconsistent) to 3 (very consistent).

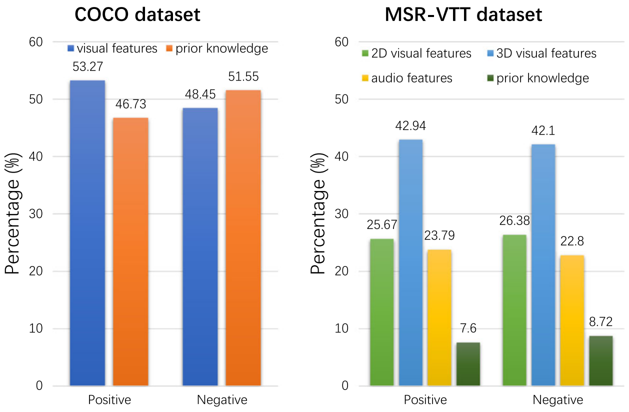


Fig. 7 Normalized fusion weights of different sentiment features in the decoding phase for sentimental image captioning on the COCO dataset and sentimental video captioning on the MSR-VTT dataset

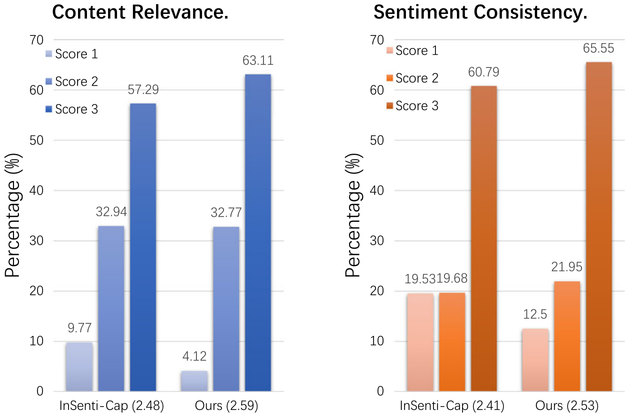
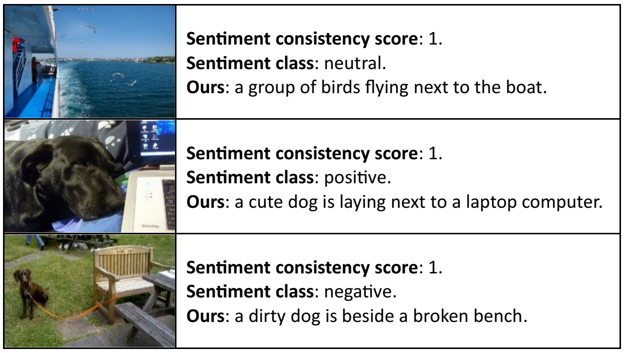


Fig. 8 Human evaluations of previous InSenti-Cap model and our Senti-Transformer. The content relevance metric is rated from 1 (unrelated) to 3 (very related), and the sentiment consistency is rated from 1 (inconsistent) to 3 (very consistent). The vertical axis represents the percentage of each score, and the average score is in parentheses

The results are reported in Fig. 8, where the vertical axis represents the percentage of each score, and the average score is in parentheses.

From the results, we can conclude that our model has a large improvement in content, with an average score increasing from 2.48 to 2.59; for sentiment, there is also a certain enhancement, with an average score increasing from 2.41 to 2.53. These evaluate the higher quality of the captions generated by our Senti-Transformer.

Figure 9 shows several failure cases which are rated "1 (bad)" for sentiment consistency in the human evaluation. For example, the image itself presents a beautiful scene and humans feel the positive sentiment, but the sentiment analyzer predicts the neutral sentiment, so the generated sentimental caption is neutral and does not fit the people's expectations.



Sentiment consistency score: 1.
Sentiment class: neutral.
Ours: a group of birds flying next to the boat.

Sentiment consistency score: 1.
Sentiment class: positive.
Ours: a cute dog is laying next to a laptop computer.

Sentiment consistency score: 1.
Sentiment class: negative.
Ours: a dirty dog is beside a broken bench.

Fig. 9 Failure cases with low scores of sentiment consistency in the human evaluation. Human evaluation scores for the sentiment consistency, sentiment class predicted by the sentiment analyzer, and sentiment caption generated by our Senti-Transformer are shown

4.10 Qualitative Results

4.10.1 Qualitative results of the ablation study

In Fig. 10, we show several examples of the sentimental sentences generated by Senti-Transformer and its variants, including "w/o cls", "w/o ppl", "w/o pri" and "w/o audio". It can be seen that our model generates captions for images or videos that are both content-related and sentiment-consistent.

As shown in the first two rows of Fig. 10(a), the captions generated by our model is more emotional than that of "w/o cls" and "w/o ppl". Taking the first row for example, "w/o cls" only generates the factual phrases of "yellow frisbee" and "small bicycle" without sentiment, while our model generates "cute dogs" and "lonely street" with significant sentiments. Compared with "w/o pri", our model describes the visual content more accurately, since the prior knowledge provides some commonly used sentiment words for describing objects and their relationships. For example, in the third row of Fig. 10(a), the sentimental descriptions of "good cake" and "dirty sky" generated by "w/o ppl" are correct but not very accurate. The sentimental phrases of "delicious cake" and "gloomy sky" generated by our model are more reasonable and lively. The audio information plays a vital role in sentimental video captioning. As shown in Fig. 10(b), if there is no audio, it is easy to be misunderstood that a girl is playing with a pet, but with the audio it is obvious that the girl is actually showing her pet.

4.10.2 Qualitative results compared with other methods

Fig. 11 shows several examples of the qualitative comparison with other methods, including the preliminary InSentiCap, the state-of-the-art factual captioning methods (DLCT (Luo et al., 2021) for image captioning, NACF (Yang et al., 2021) for video captioning), our Senti-Transformer and the ground-truth (GT) captions. As shown in Fig. 11, the fac-

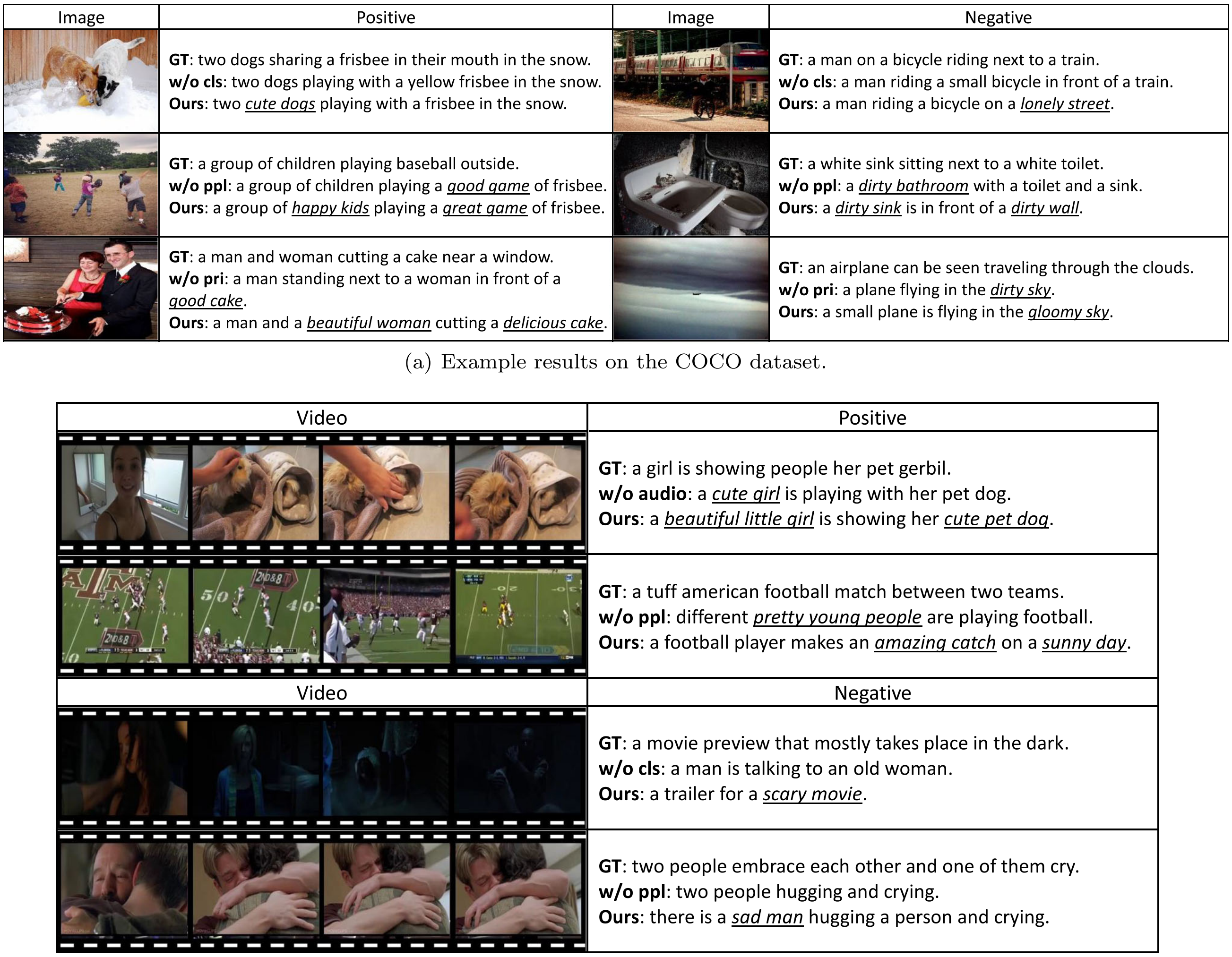


Image	Positive	Image	Negative
	GT: two dogs sharing a frisbee in their mouth in the snow. **w/o cls**: two dogs playing with a yellow frisbee in the snow. **Ours**: two *cute dogs* playing with a frisbee in the snow.		**GT**: a man on a bicycle riding next to a train. **w/o cls**: a man riding a small bicycle in front of a train. **Ours**: a man riding a bicycle on a *lonely street*.
	GT: a group of children playing baseball outside. **w/o ppl**: a group of children playing a *good game* of frisbee. **Ours**: a group of *happy kids* playing a *great game* of frisbee.		**GT**: a white sink sitting next to a white toilet. **w/o ppl**: a *dirty bathroom* with a toilet and a sink. **Ours**: a *dirty sink* is in front of a *dirty wall*.
	GT: a man and woman cutting a cake near a window. **w/o pri**: a man standing next to a woman in front of a *good cake*. **Ours**: a man and a *beautiful woman* cutting a *delicious cake*.		**GT**: an airplane can be seen traveling through the clouds. **w/o pri**: a plane flying in the *dirty sky*. **Ours**: a small plane is flying in the *gloomy sky*.

(a) Example results on the COCO dataset.

Video	Positive
	GT: a girl is showing people her pet gerbil. **w/o audio**: a *cute girl* is playing with her pet dog. **Ours**: a *beautiful little girl* is showing her *cute pet dog*.
	GT: a tuff american football match between two teams. **w/o ppl**: different *pretty young people* are playing football. **Ours**: a football player makes an *amazing catch* on a *sunny day*.
Video	**Negative**
	GT: a movie preview that mostly takes place in the dark. **w/o cls**: a man is talking to an old woman. **Ours**: a trailer for a *scary movie*.
	GT: two people embrace each other and one of them cry. **w/o ppl**: two people hugging and crying. **Ours**: there is a *sad man* hugging a person and crying.

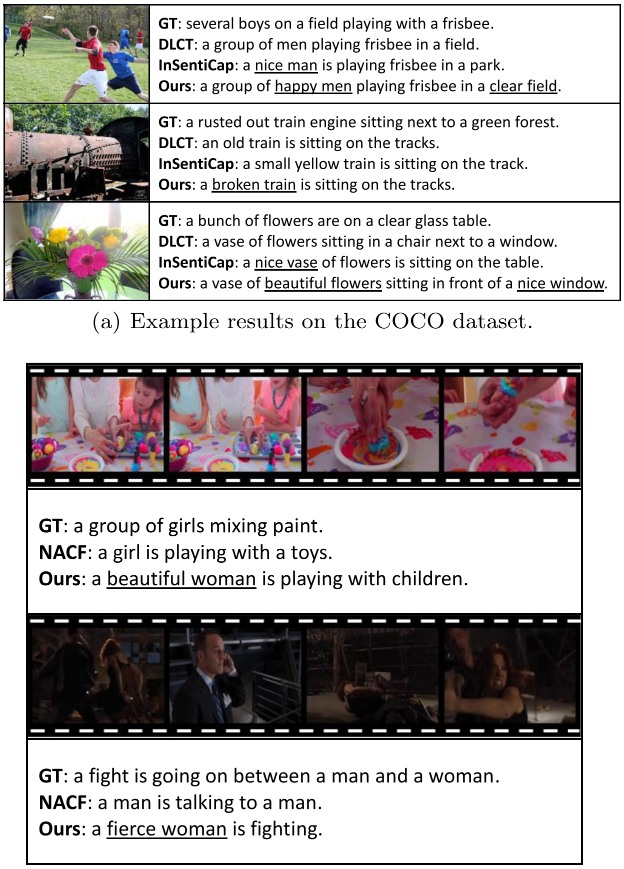
(b) Example results on the MSR-VTT dataset.

Fig. 10 Example results of the ablation study. Each item includes an image (or a video) and the corresponding ground-truth caption, the result of ablation study (i.e. "w/o cls", "w/o ppl", "w/o pri" or "w/o audio"), and the sentimental caption generated by our Senti-Transformer

tual captioning methods (DLCT and NACF) only focus on the visual content, and our method successfully takes into account both content and sentiment. For video captioning, in some cases, our method even achieves better results than NACF on generating factual captions, which validates its good ability to encode multimodal information. Taking the second video case in Fig. 11(b) for example, our Senti-Transformer succeeds in describing the factual content as "fighting", while NACF wrongly generates the caption of "talking". Compared with the preliminary InSentiCap, our Senti-Transformer performs better in terms of both content and sentiment, clearly validating the superiority of multi-modal Transformer on sentimental visual captioning. For example, as shown in Fig. 11(a), for the first image case, our Senti-Trnasformer accurately describes "a group of men playing", while InSentiCap only identifies a man. For the second image case, Senti-Trnasformer embellishes the train with a correct sentiment word "broken", while InSentiCap generates a neutral caption that does not match the image sentiment.

5 Conclusion

We have proposed a novel sentimental visual captioning task that can generate captions in line with the intrinsic sentiments of images or videos. To address this task, we have presented a multimodal Senti-Transformer model and designed a two-stage training strategy using the pairs of images or videos and factual captions as well as the extra sentimental corpus. Our Senti-Transformer is capable of understanding the content and sentiment of the image or video and describing the visual content with a linguistic style, simultaneously. Extensive experiments demonstrate the superiority of our method.



GT: several boys on a field playing with a frisbee.
DLCT: a group of men playing frisbee in a field.
InSentiCap: a nice man is playing frisbee in a park.
Ours: a group of happy men playing frisbee in a clear field.

GT: a rusted out train engine sitting next to a green forest.
DLCT: an old train is sitting on the tracks.
InSentiCap: a small yellow train is sitting on the track.
Ours: a broken train is sitting on the tracks.

GT: a bunch of flowers are on a clear glass table.
DLCT: a vase of flowers sitting in a chair next to a window.
InSentiCap: a nice vase of flowers is sitting on the table.
Ours: a vase of beautiful flowers sitting in front of a nice window.

(a) Example results on the COCO dataset.

GT: a group of girls mixing paint.
NACF: a girl is playing with a toys.
Ours: a beautiful woman is playing with children.

GT: a fight is going on between a man and a woman.
NACF: a man is talking to a man.
Ours: a fierce woman is fighting.

(b) Example results on the MSR-VTT dataset.

Fig. 11 Example results compared with other methods. Each item includes an image (or a video), the corresponding ground-truth caption, the result of factual captioning method (DLCT or NACF), the sentimental caption generated by InSentiCap (for image captioning) and the sentimental caption generated by our Senti-Transformer

In future, we are going to implement the pipeline of sentimental visual captioning task through unsupervised learning to alleviate the dependency on the paired training data of images or videos and factual captions. We will also intend to extract richer information to assist the generation of captions, such as scenes, actions, and text obtained through OCR and ASR.

Acknowledgements This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant No. 62072041.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp 6077–6086).

Bargal, S. A., Barsoum, E., Ferrer, C. C., & Zhang, C. (2016). Emotion recognition in the wild from videos using images. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (pp 433–436).

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Borth, D., Ji, R., Chen, T., Breuel, T., & Chang, S. F. (2013). Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM* (pp 223–232).

Campos, V., Jou, B., & Giro-i Nieto, X. (2017). From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction. *Image and Vision Computing, 65*, 15–22.

Chen, C. K., Pan, Z., Liu, M. Y., & Sun, M. (2019). Unsupervised stylish image description generation via domain layer norm. In *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (pp 8151–8158).

Chen, Y., Wang, S., Zhang, W., & Huang, Q. (2018). Less is more: Picking informative frames for video captioning. In *Proceedings of the European conference on computer vision (ECCV)* (pp 358–373).

Cornia, M., Stefanini, M., Baraldi, L., & Cucchiara, R. (2020). Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp 10578–10587).

Denkowski, M., & Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *proceedings of the ninth workshop on statistical machine translation* (pp 376–380).

Fang, K., Zhou, L., Jin, C., Zhang, Y., Weng, K., Zhang, T., & Fan, W. (2019). Fully convolutional video captioning with coarse-to-fine and inherited attention. In *Proceedings of the AAAI Conference on Artificial Intelligence 33* (pp 8271–8278).

Gan, C., Gan, Z., He, X., Gao, J., & Deng, L. (2017). Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp 3137–3146).

Guo, L., Liu, J., Yao, P., Li, J., & Lu, H. (2019). Mscap: Multi-style image captioning with unpaired stylized text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp 4204–4213).

Guo, L., Liu, J., Zhu, X., Yao, P., Lu, S., & Lu, H. (2020). Normalized and geometry-aware self-attention network for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp 10327–10336).

Hara, K., Kataoka, H., & Satoh, Y. (2018). Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp 6546–6555).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp 770–778).

Hershey, S., Chaudhuri, S., Ellis, DP., Gemmeke, JF., Jansen, A., Moore, RC., Plakal, M., Platt, D., Saurous, RA., Seybold, B., et al. (2017). Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, IEEE (pp 131–135).

Huang, L., Wang, W., Chen, J., & Wei, XY. (2019). Attention on attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision* (pp 4634–4643).

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *CVPR* (pp 3128–3137).

Kingma, D. P., Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR*.

Lei, J., Wang, L., Shen, Y., Yu, D., Berg, TL., & Bansal, M. (2020). Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. In *ACL*.

Li, G., Zhai, Y., Lin, Z., & Zhang, Y. (2021a). Similar scenes arouse similar emotions: Parallel data augmentation for stylized image captioning. In *Proceedings of the 29th ACM International Conference on Multimedia* (pp 5363–5372).

Li, T., Hu, Y., & Wu, X. (2021b). Image captioning with inherent sentiment. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE.

Lin, C., Zhao, S., Meng, L., & Chua, TS. (2020). Multi-source domain adaptation for visual sentiment classification. arXiv preprint arXiv:2001.03886.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp 740–755). Springer.

Luo, Y., Ji, J., Sun, X., Cao, L., Wu, Y., Huang, F., Lin, CW., Ji, R. (2021). Dual-level collaborative transformer for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp 2286–2293).

Machajdik, J., & Hanbury, A. (2010). Affective image classification using features inspired by psychology and art theory. In *ACM MM* (pp 83–92).

Mathews, AP., Xie, L., & He, X. (2016). Senticap: Generating image descriptions with sentiments. In *Thirtieth AAAI conference on artificial intelligence*.

Nguyen, D., Nguyen, K., Sridharan, S., Dean, D., & Fookes, C. (2018). Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition. *Computer Vision and Image Understanding, 174*, 33–42.

Pan, B., Cai, H., Huang, D. A., Lee, K. H., Gaidon, A., Adeli, E., & Niebles, J. C. (2020). Spatio-temporal graph for video captioning with knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp 10870–10879).

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp 311–318).

Peng, K. C., Sadovnik, A., Gallagher, A., & Chen, T. (2016). Where do emotions come from? predicting the emotion stimuli map. In *ICIP* (pp 614–618).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*(8), 9.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017). Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp 7008–7024).

Stolcke, A., (2002) Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Suin, M., & Rajagopalan, A. (2020). An efficient framework for dense video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (pp 12039–12046).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp 5998–6008).

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp 4566–4575).

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp 3156–3164).

Wang, W., Chen, Z., & Hu, H. (2019). Hierarchical attention network for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence* 33, (pp 8957–8964).

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning, 8*(3), 229–256.

Wu, X., Zhao, W., & Luo, J. (2022). Learning cooperative neural modules for stylized image captioning. *International Journal of Computer Vision, 130*(9), 2305–2320.

Xu, J., Mei, T., Yao, T., & Rui, Y. (2016). Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp 5288–5296).

Yang, B., Zou, Y., Liu, F., & Zhang, C. (2021). Non-autoregressive coarse-to-fine video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence* 35, (pp 3119–3127).

Yang, J., She, D., Lai, Y. K., Rosin, P. L., & Yang, M. H. (2018a). Weakly supervised coupled networks for visual sentiment analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp 7584–7592).

Yang, J., She, D., Sun, M., Cheng, M. M., Rosin, P. L., & Wang, L. (2018). Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Transactions on Multimedia, 20*(9), 2513–2525.

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., & Courville, A. (2015). Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision* (pp 4507–4515).

You, Q., Luo, J., Jin, H., & Yang, J. (2015). Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Twenty-ninth AAAI conference on artificial intelligence*.

You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp 4651–4659).

You, Q., Jin, H., & Luo, J. (2017). Visual sentiment analysis by attending on local image regions. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Yu, H., Wang, J., Huang, Z., Yang, Y., & Xu, W. (2016). Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp 4584–4593).

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J. (2021). Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp 5579–5588).

Zhao, S., Ma, Y., Gu, Y., Yang, J., Xing, T., Xu, P., Hu, R., Chai, H., & Keutzer, K. (2020). An end-to-end visual-audio attention network for emotion recognition in user-generated videos. In *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (pp 303–311).

Zhao, W., Wu, X., & Zhang, X. (2020). Memcap: Memorizing style knowledge for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (pp 12984–12992).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.