IETE TECHNICAL REVIEW
https://doi.org/10.1080/02564602.2020.1730714

Neural Visual Social Comment on Image-Text Content

Yue Yin, Hanzhou Wu and Xinpeng Zhang

School of Communication and Information Engineering, Shanghai University, Shanghai 200444, People's Republic of China

ABSTRACT
Social bots are computer software designed for content production and interaction with humans. With the popularity of images in social networks, social bots need to have visual awareness of image content while only understanding texts is far from enough to be active in social networks. We introduce a novel task, Visual Social Comment (VSC), in which social bots should generate relevant and informative comments on social contents of both images and texts. In this task of multimodal context, our work focuses on how to extract and fuse the information of vision and text to improve the quality of generated comments, and how to deal with the problem that neural dialog models trained with maximum likelihood estimation (MLE) criteria tend to generate generic responses. In order to fuse visual and textual context features closely through the relationship between them, we adopt joint attention of multimodal context to modify the standard sequence-to-sequence (Seq2Seq) framework. We also leverage the topic information transferred from a topic classification model to build a perceptual loss function, which encourages the generative comment model to generate more informative and diverse comments with the topic corresponding to context. The experimental results of models trained with data from Sina Weibo show that comments generated by our proposed models achieve better performance in both relevance and informativeness than those generated by other baseline models.

KEYWORDS
Artificial intelligence; Context awareness; Human–computer interaction; Intelligent robots; Knowledge transfer; Natural language processing

1. INTRODUCTION

The rise of social networking services (SNS) like Facebook, Twitter, Sina Weibo, etc. has led to the rapid growth of social bots. Social bots are computer software that attempt to emulate human behaviors in social networks by automatically generating contents and interacting with human users [1].

Capabilities of social bots can bring convenience to human life. Since social bots can automate certain tasks, such as gathering information: automatically pushing news and weather news, or responding to inquiries, they can be designed to be helpful like intelligent personal assistants such as Siri, Cortana and chatbots like Microsoft's Tay [2].

On the other hand, social bots can also be designed to be harmful, performing malicious activities such as distributing spam, spreading malware, disguising, launching Sybil Attacks, and so on. These malicious features of social bots partially lead to appearance of abundant boring and meaningless information in social networks. For example, Sina Weibo [3], as the most mainstream social application in China, has a large number of repetitive and uninformative comments due to paid posters (including social bots), advertisements or users' social habits. These comments will make users bored with SNS and reduce the enthusiasm of users to participate in online social activities, therefore reducing the activity of the entire social networks.

Considering this situation, the next generation of social bots should have more advanced abilities to generate diverse and informative comments on posts containing both text and image content, which will instead attract more human users and maintain service quality and activity of SNS. The key point in the design of social bots is human–computer conversation, which has been a hot topic in the field of artificial intelligence (AI) and natural language processing (NLP). Methods for conversation agents mainly fall into three categories: the retrieval-based methods [4–6], the statistical machine translation (SMT)-based methods [7] and the neural network-based methods [8–11]. Currently, neural network-based methods represented by sequence-to-sequence (Seq2Seq) models with attention mechanism [12,13] have become the dominant paradigm, because of their scalable and end-to-end framework with capability to capture both semantic and syntactic relations between post-response pairs. However, neural conversation models trained by

maximum likelihood estimation (MLE) criteria tend to generate "safe" generic responses, such as "I don't know" or "I'm fine". In order to generate diverse and informative responses, related research improves response quality from two directions: modifying model structure to introduce more information to guide the generation [10,14,15], or modifying the objective function to adjust the generated content [16,17]. However, these works only apply to conversations grounded on textual context, without considering the existence of visual context.

Nowadays images are ubiquitous in social scenarios, so the social robot should not only have the ability to understand text but also have visual awareness. Due to tremendous progress in computer vision (CV), tasks combining vision and language have developed rapidly, especially image caption [18,19], visual question answering (VQA) [20], visual dialog (VisDial) [21], and video question answering [22,23]. There is no doubt that these tasks raise requirements for machines' abilities to see and communicate, compared with text-only dialogs. However, it is worth noting that these tasks, whether describing images or answering questions about image or video content, still remain in the understanding of their content. Since people's comments on a post containing images often express personal attitude, emotion and social style, beyond visible objects in the images, generating natural comments in social media obviously poses a challenge to current dialog systems.

In this paper, we introduce a novel task, Visual Social Comment (VSC), in which a system should generate diverse and interesting social comments on a published post containing text and image content. Different from one-to-one multiple rounds of conversation, VSC doesn't need to keep the conversation going, because any user can comment on a post, start or finish a conversation with other users under the post at any time. Therefore, VSC focuses on one round of conversation, generating diverse and informative comments on the post with textual and visual context. Moreover, while some tasks take one image of specific content as visual context, VSC doesn't constrain the number and content of images.

In the generation task with multimodal context, how to fuse features of multimodal context directly affects the quality of generated responses. Generally, visual and textual context of a post are not independent of each other, and images usually contain some extraneous information irrelevant to the post's topic. Therefore, instead of concatenation of raw visual and textual features, we adopt visual and textual joint attention to find out which part of features of multiple images are related to the textual features before concatenating them, in order to reduce the noise introduced by feature concatenation. Since generating generic responses is a common problem of generation models trained with MLE criteria, we employ a perceptual loss function based on topic knowledge transferred from a pretrained topic classification model to train the generative visual comment model, in order to generate more informative, diverse and topic-relevant responses from a limited number of human comments. The idea is inspired by our observation on human comments in social media. Although content and expression of comments on a post vary from person to person, we regard them as suitable comments because their topic corresponds to the post.

In summary, the main contributions of our work are fourfold:

- We introduce a new task, VSC, combining visual and textual context in the online social scene to generate interesting and informative comments.
- We modify the Seq2Seq framework with joint attention of visual and textual context to learn relations between the multimodal context and responses in VSC.
- We propose to build a perceptual loss function based on knowledge transferred from a topic classification model to encourage the generative visual comment model to generate more informative, diverse, and topic-relevant responses.
- We conduct an empirical study on large-scale data crawled from Sina Weibo and compare different methods by both automatic evaluation and human judgment.

The remainder of this paper is organized as follows. In Section 2, we briefly review the related work. In Section 3, we define the task of VSC and present the proposed neural visual comment model. The experimental results and analysis are presented in Section 4 followed by conclusion in Section 5.

2. RELATED WORK

2.1 Vision and Language

With precedented advances of CV, visual features combined with language modeling have received great attention in tasks at the intersection of vision and language, such as image captioning, visual question

answering (VQA), visual dialog (VisDial), and video question answering.

A VQA system is to answer questions about visual content of a given image [20]. As a step towards visual human–machine interaction, Das *et al.* extend VQA's one single round of dialog to consistent conversation with several relevant questions, which is named as VisDial [21]. VisDial requires the machine to have visual memory, understand co-reference resolution and be consistent with its answers in order to engage humans in conversation. A video question answering system is to answer a free-form natural language question about the content of a video, which requires finer understanding of video content and questions, and capability of reasoning across video frames [22].

Beyond answering questions directly from the image, Mostafazadeh *et al.* introduce the task of image-grounded conversation (IGC) [24], in which a system should hold a natural conversation around a shared image. They propose a modified Seq2Seq structure that concatenates visual features together with textual features for conversational language generation. Instead of using raw image features, Huber *et al.* extend the Seq2Seq structure with visual sentiment, facial expression and scene features [25]. The results show that their model performs most effectively on conversations grounded in images with faces, generating more informative and emotional responses to questions.

Note that there are a few subtle but significant differences between the application and approach of VSC and IGC. While IGC constrains the visual context to a single event-centric image and depends on combination of question and response generation, VSC aims to directly generate informative and meaningful social comments with multimodal context covering common topics in social media. The visual context in VSC allows to be multiple images in a post just like in social media.

Tasks at the intersection of vision and language typically involve attention mechanism attending to relevant image regions or question words. Liu *et al.* propose the history-conditioned image attentive encoder (HCIAE) [26], recognizing the region in the image helpful guided by the history and the question. We find that the final output fixed-length embedding of HCIAE is not competent enough to capture the relations between multimodal context and comments in the task of VSC, because VSC has posts and comments much longer than questions and answers in VisDial, and the number of images accompanying the post is often more than one. In this paper, we incorporate the text-conditioned image attention encoder (TCIAE) with Luong attention [13], as a joint attention in the Seq2Seq model to extract relevant parts of visual and textual context features and fuse them closely.

2.2 Neural Conversation Models

A large amount of conversation data available in social media offer a promise for data-driven conversation models. Ritter *et al.* consider response generation as a statistical machine translation (SMT) task, training models on parallel post-response pairs [7]. Neural language models in the framework of Seq2Seq have driven forward the performance of response generation based on Ritter *et al.* work. The Seq2Seq model composed of long short-term memory (LSTM) neural networks can capture semantic information and long-span dependencies in an end-to-end framework.

Confronted with the recurring problem with MLE trained neural conversation models [27], prior work has explored to increase diversity and informativeness of generated responses by optimizing the objective function [16,17], or introducing more information to encoder–decoder models [10,14,15]. Li *et al.* propose the MMI-antiLM and MMI-bidi models, penalizing high-frequency, generic responses with maximum mutual information (MMI) optimization criterion [16]. Xing *et al.* propose a topic aware sequence-to-sequence model, introducing topic information into the Seq2Seq framework [10], while Mou *et al.* put forward a sequence to backward and forward sequences model, which generates a reply containing the keyword predicted by pointwise mutual information [14]. Ghazvininejad *et al.* generalize Seq2Seq model by generating responses grounded on both conversation history and external factual information [15].

In this paper, inspired by our observation of human comments in social media, we utilize a topic classification model to learn a task-dependent topic embedding space, where comments of the same topic have similar topic feature vectors. The learned topic embedding function is transferred to the generative visual comment model to define a perceptual loss function, which encourages the distance between topic feature vectors of the generated comment and the ground truth comment to decrease. The similar perceptual loss has been adopted in style transfer and super-resolution [28], but we focus on the challenge of using the perceptual loss in social comment generation, comparing the performance of the perceptual loss against MLE.

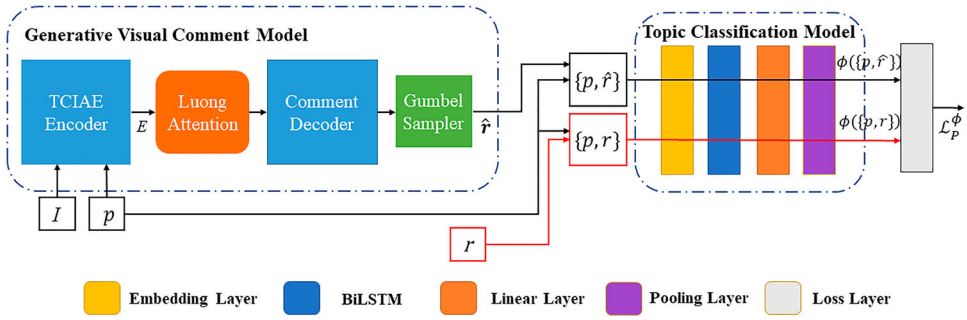

Figure 1: The structure of the neural visual comment model

3. PROPOSED METHOD

In this section, we define the task of VSC and describe the structure of our neural visual comment model. Furthermore, we describe the training procedure to transfer learned knowledge from a topic classification model (T) to the generative visual comment model (G) by minimizing a perceptual loss function.

Figure 1 shows the overview of our neural visual comment model. The neural visual comment model consists of the generative visual comment model G and the topic classification model T. The topic classification model T is pretrained to learn the topic embedding space with the textual context $\boldsymbol{p}$ and ground truth comment $\boldsymbol{r}$ as input. G is an encoder–decoder structure with visual and textual joint attention, given a list of images $\boldsymbol{I}$, textual context $\boldsymbol{p}$ as input, the encoder maps the visual and textual context into a sequence of encoder vectors $\bar{\boldsymbol{e}} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_T)$. In the decoding phase, the decoder first initializes its hidden state with the encoder's final vector $\boldsymbol{e}_T$ and produces the predictive distribution over the comment via LSTM and Luong attention. We use a Gumbel-Softmax sampler $\boldsymbol{S}$ to sample the comment token to enable end-to-end differentiability. Feeding the sampled comment $\hat{\boldsymbol{r}}$ into the topic model T, the generator G will be optimized to generate a comment which is close to $\boldsymbol{r}$ in the topic embedding space.

Next, we describe approaches for each component in detail.

3.1 Visual Social Comment (VSC)

We define the task of VSC as follows: as shown in Figure 2, given a post which contains a list of images $\boldsymbol{I} = \{\boldsymbol{i}_1, \ldots, \boldsymbol{i}_n\}$ and a textual context $\boldsymbol{p}$ as input, a visual comment model needs to generate fluent, appropriate, diverse, and informative comments on the post as output. We assume that the textual and visual context of the post are relevant.

Figure 2: The task of Visual Social Comment

3.2 Generative Visual Comment Model with Visual and Textual Joint Attention

It is obvious that the textual context and images of a post published by a human user are usually relevant. Images alone usually evoke diverse ideas from different readers, but textual context along will give readers a central idea about the topic and turns their attention to the topic-related regions in images. This means that a VSC model needs to fuse textual context with visual context closely instead of simple concatenation of individual

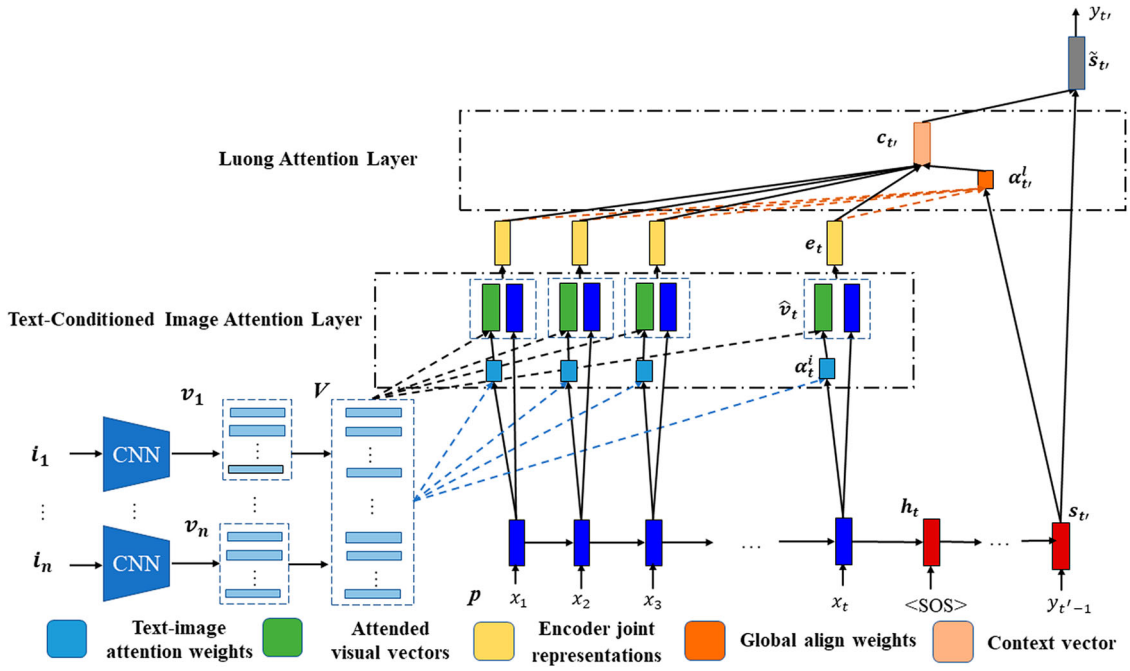

Figure 3: The structure of the generative visual comment model

representations of them. With this motivation, we adopt a similar attention mechanism at the encoder as HCIAE in [26], using the textual context to find regions of multiple images relevant. The encoder with text-conditioned image attention in our model is named TCIAE. The difference between TCIAE and HCIAE is that we encode the multimodal context into a sequence of vectors instead of a fixed-length global encoder feature vector, in order to enable the joint attention mechanism to be applied in multimodal context encoder–decoder model. The structure of the generative visual comment model with visual and textual joint attention is shown in Figure 3.

The spatial image features are extracted from pretrained convolutional neural network (CNN) with outstanding classification performance on ImageNet [29]. The image features are then transformed to d-dimensional vectors and concatenated as vector $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n) \in \mathbb{R}^{d \times n}$, where $\boldsymbol{v}_n \in \mathbb{R}^d$ denotes the visual vector of image $\boldsymbol{i}_n$. Meanwhile, the textual context $\boldsymbol{p}$ is encoded by an LSTM as all the hidden states $\boldsymbol{H} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_T)$. Conditioned on the hidden state at each time step t, the model attends to images to identify visual vectors of subregions relevant to the hidden state $\boldsymbol{s}_t$ at time step t. The attended joint representations of the text and images are concatenated to obtain the sequence of encoder vectors as follows:

$$\boldsymbol{z}_t = \boldsymbol{\omega}^T \tan h(\boldsymbol{W}_I \boldsymbol{V} + (\boldsymbol{W}_T \boldsymbol{h}_t)\boldsymbol{\gamma}^T) \tag{1}$$

$$\boldsymbol{\alpha}_t^i = \text{softmax}(\boldsymbol{z}_t) \tag{2}$$

$$\hat{\boldsymbol{v}}_t = \sum_{j=1}^{n} \alpha_{tj}^i \boldsymbol{v}_j \tag{3}$$

$$\boldsymbol{e}_t = \tan h(\boldsymbol{W}_e[\boldsymbol{h}_t, \hat{\boldsymbol{v}}_t] + \boldsymbol{b}_e) \tag{4}$$

where $\boldsymbol{\omega} \in \mathbb{R}^n$, $\boldsymbol{W}_I, \boldsymbol{W}_T \in \mathbb{R}^{n \times d}$, $\boldsymbol{W}_e \in \mathbb{R}^{d \times 2d}$, $\boldsymbol{b}_e \in \mathbb{R}^d$ are the parameters to be learned and $\boldsymbol{\gamma} \in \mathbb{R}^n$ is a vector whose elements are all 1. $\boldsymbol{\alpha}_t^i$ is the attention weight which determines the likelihood that each of the subregions is relevant to the hidden state $\boldsymbol{h}_t$ of the textual context. The attended visual vector $\hat{\boldsymbol{v}}_t$ is a weighted sum of all image features with elements of $\boldsymbol{\alpha}_t^i$ as weight. $\boldsymbol{e}_t$ is the joint representation of $\hat{\boldsymbol{v}}_t$ and $\boldsymbol{h}_t$. The final encoder output is a sequence of the joint representations as $E = \{\boldsymbol{e}_t\}_{t=1}^T$.

In the decoding phase, the final joint representation of encoder $\boldsymbol{e}_T$ is first fed into the decoder. Then the decoder begins to predict the current target word $\boldsymbol{y}_{t'}$ with Luong global attention, where the target hidden state $\boldsymbol{s}_{t'}$ and the context vector $\boldsymbol{c}_{t'}$, the weighted average over all the joint representations $\boldsymbol{E}$, are concatenated to produce an attentional hidden state $\tilde{\boldsymbol{S}}_{t'}$ as follows:

$$\alpha_{t't}^l = \frac{\text{score}(s_{t'}, h_{t'})}{\sum_{l=1}^{T} \exp(\text{score}(s_{t'}, h_l))} \tag{5}$$

$$\boldsymbol{c}_{t'} = \sum_{t=1}^{T} \alpha_{t't}^l \boldsymbol{h}_t \tag{6}$$

$$\tilde{\boldsymbol{S}}_{t'} = \tan h(\boldsymbol{W}_c[\boldsymbol{c}_{t'}, \boldsymbol{s}_{t'}] + \boldsymbol{b}_c) \tag{7}$$

where score(·) is a content-based function which computes the probability that each joint representation of textual and visual context is relevant to the current state to generate the word. $\boldsymbol{W_c} \in \mathbb{R}^{d\times 2d}$ and $\boldsymbol{b_c} \in \mathbb{R}^{d}$ are respectively weight matrix and bias vector to be learned. Unlike attention mechanism in text-only language models, our model pays attention not only to parts of the sentence, but also the attended regions in images relevant to the word to be generated. This enables the decoder to have access to information of the variable-length sequence of encoder vectors, where semantic information of the visual vector corresponds to the textual vector.

3.3 Perceptual Loss Based on Topic Knowledge

The topic knowledge of a pretrained topic classification model is utilized to build a topic-sensitive perceptual loss function, encouraging the generative visual comment model to generate informative and diverse comments with a topic similar to that of human comments.

We select the recurrent convolutional network (RCNN) for text classification [30] as our topic classification model T.

Figure 4 shows the architecture of RCNN model, which comprises three layers: convolutional layer and pooling layer and output layer.

The input of RCNN model is a sequence of words of the post-comment pair. At the convolutional layer, the model captures the context around each word with bidirectional LSTM (BiLSTM) and then concatenates the output of BiLSTM $\left(\overrightarrow{\boldsymbol{h}}_{\boldsymbol{i}}^{\boldsymbol{T}}, \overleftarrow{\boldsymbol{h}}_{\boldsymbol{i}}^{\boldsymbol{T}}\right) \in \mathbb{R}^{2m}$ to the word embedding as the contextual representation of the word. Then the latent semantic vector $\tilde{\boldsymbol{h}}_{\boldsymbol{i}}$ is obtained by applying a linear transformation together with the tan h activation function to the representation of each word. The max-pooling layer determines the most significant factor for representing the topic throughout all latent semantic vectors of the text. Finally, the output layer linearly transforms the output vector of pooling layer and apply softmax function to convert the output numbers into probabilities $\boldsymbol{p_T}$.

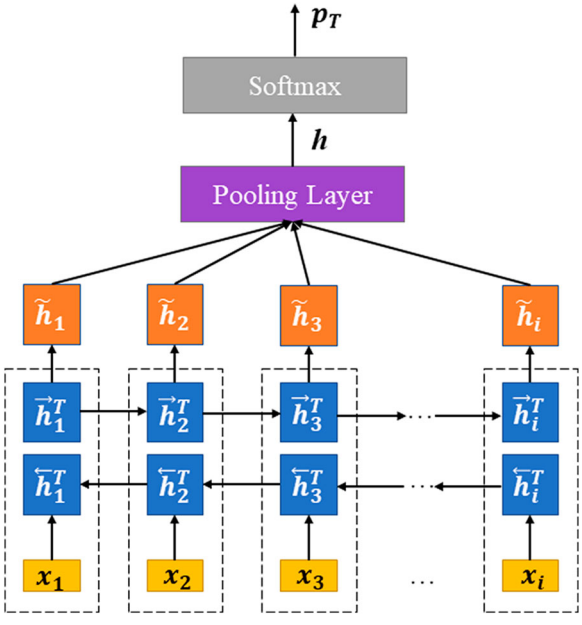

Figure 4: The structure of recurrent convolutional network for topic classification

The RCNN topic classification model is trained with cross-entropy loss function to learn a topic embedding space where the topic feature vectors of post-comment pairs with the same topic are close to each other. Since Euclidean distance between topic feature vectors is adopted to measure the similarity between topics of post-comment pairs, we expect topic feature vectors to reflect general features of comments of a specific topic but keep some degree of distinction. Although the probability vector indicates probability distribution over predefined topics, building our perceptual loss function based on the low dimensional probability vector probably ignores many differences between comments of similar topics. That will result in reducing the diversity of generated comments. By contrast, the output vector $\boldsymbol{h}$ of pooling layer is high dimensional feature vector containing specific topics' general feature information but still keeping some post-comment pair's own distinction. Therefore, we select the high dimensional output vector $\boldsymbol{h}$ of the pooling layer as topic feature vector instead of output probabilities of the softmax function.

To transfer knowledge from the pretrained T to G, we follow [26] to adopt a Gumbel-Softmax [31,32] sampler to face the challenge that the discrete output symbols of G disable the differentiability in the backward pass from T to G. At each step to generate the comment, we extend the decoder LSTM with a Gumbel-Softmax sampler to sample the comment word from the conditional distribution.

When we connect G with pretrained T, we repeatedly feed the textual context $\boldsymbol{p}$ with the sampled comment $\hat{\boldsymbol{r}}$ generated by the generator G into T to update G's parameters until the distance between $\{\boldsymbol{p}, \hat{\boldsymbol{r}}\}$ and $\{\boldsymbol{p}, \boldsymbol{r}\}$ in the topic embedding space is short enough, where $\boldsymbol{r}$ is the human comment. The learned topic embedding of the pretrained topic model T is transferred to the perceptual loss function which G aims to optimize. The perceptual

loss function based on topic features is formulated as follows:

$$\mathcal{L}_P^{\phi}(\boldsymbol{p}, \hat{\boldsymbol{r}}, \boldsymbol{r}) = \frac{1}{m}\phi(\{\boldsymbol{p}, \widehat{\boldsymbol{r}}\}) - \phi(\{\boldsymbol{p}, \boldsymbol{r}\})_2^2, \tag{8}$$

where ϕ is the topic embedding function learned by T. To be exact, $\phi(\cdot)$ is the function which outputs the topic feature vector of the max-pooling layer of T, and the perceptual loss is the normalized and squared Euclidean distance between topic feature vectors. Under the setting that comments of similar topics have topic vectors close in the topic embedding space, updating G's parameters to minimize $\mathcal{L}_P^{\phi}$ can be explained as a process of learning to generate comments consistent with the topic of the post and human comments, but of diverse content and expression. The perceptual loss function aims to encourage the model to generate interesting comments around the central topic of the post and human comments, straying away from the MLE training process of generating comments as much like human comments as possible.

4. EXPERIMENTS AND ANALYSIS

4.1 Dataset

We train our model on a dataset of post-comment pairs crawled from Sina Weibo, a prevalent Twitter-like microblogging service in China which allows users to post or comment on posts published. First, we crawl 100,000 post-comment pairs from Sina Weibo covering 8 common topics, including pet, makeup, star, food, photography, home, fashion, and traveling. After preprocessing, there are 16,200 post-comment pairs containing 1058 distinct posts left. From these pairs, we randomly sample 850 posts with 13,153 corresponding comments as training data and 208 posts with 3047 corresponding comments as test data. We use a Chinese word processing toolkit FoolNLTK [33] to tokenize the post-comment pairs, and then build a vocabulary of words that occur at least 3 times in training data, resulting in 6401 words. Words outside the vocabulary are all replaced by a special token "UNK".

4.2 Training Details

In our experiments, the 2 LSTMs in G are 2 layers with 512-dimensional hidden states, while the LSTM in T is a single layer bidirectional LSTM with 512-dimensional hidden states. We apply Tencent AI Lab Chinese embeddings [34] to initialize our word embeddings. For the better feature representation ability of ResNet [29], we use ResNet to get representations of images. In tasks of image-text generation, the activation values at the fully connected layer in the CNN are usually extracted as the global feature vector, representing the semantic content of the overall images. However, in our attention-based approach, we need spatial visual features containing a set of visual vectors for subregions in the image, which can enable joint attention mechanism to determine the likelihood that each of subregions is relevant to the post or comment. Therefore, activation values at the fully connected layer in ResNet does not apply to our approach. Since the output feature maps of the last convolutional layer in ResNet are spatial representations of the image, it is taken as image features in our approach.

In order to ensure the fluency of generated comments, we take a similar training method as [26]. We first pretrain T and G with cross-entropy loss for 20 and 30 epochs respectively. Afterward, we train G with $\mathcal{L}_P^{\phi} + \alpha\mathcal{L}_{CE}$, which is the combination of the perceptual loss $\mathcal{L}_P^{\phi}$ and the cross-entropy loss $\mathcal{L}_{CE}$. The latter is helpful for the fluency and grammaticality of comments generated. Empirically we set α as 0.5.

4.3 Evaluation Metrics

How to evaluate a response generation model automatically has been an open problem without final conclusion. Following [16,35], we adopt perplexity and distinct-1 and distinct-2 as automatic metrics to evaluate our model's performance. Noted that we don't use BLEU [36] as our evaluation metrics, because BLEU is known to correlate poorly with human judgment in evaluating responses [37].

In information theory, perplexity measures how well a probability distribution or probability model predicts the response. Low perplexity generally indicates that the probability model predicts the comment sample well. The perplexity of a probability model is defined as

$$PPL = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log(p(c_i))\right) \tag{9}$$

where N is the number of comment samples and c_i is a comment sample. Notice that the perplexity shown in Table 1 is a language model perplexity of per word in comment samples.

Table 1: The results of human annotations

Models	+2	+1	0
T-Gen	34.8%	20.9%	44.1%
V&T-Gen	24.3%	32.4%	48.6%
V&T-E-Gen	32.9%	27.6%	39.5%
V&T-JAM-CE	56.4%	25.3%	18.2%
V&T-JAM-PL	77.6%	15.7%	6.7%

We respectively count numbers of distinct unigrams and bigrams in the generated comments, and follow [10] to denote the numbers and the ratios of unigrams and bigrams as distinct-1 and distinct-2 respectively. These two metrics measure the informativeness and diversity of the generated comments. The higher distinct-1 and distinct-2 are, the more content the generated comments will have.

In addition to these automatic metrics, we also adopt human annotation to compare the performance of different response generation models. Following the scoring criteria established in [8], we invite 10 labelers with more than 3-year Sina Weibo experience to judge the quality of comments generated by different models with scores from 0 to 2 according to the naturalness, relevance, and diversity of responses. We select generated comments by greedy search and randomly shuffle them for each labeler. The agreement among labelers is measured by Fleiss' kappa [38].

4.4 Baselines

We compare the performance of our model V&T-JAM-PL with the following baselines to evaluate the individual contributions of joint attention of visual and textual features (JAM) and our training procedure of the perceptual loss (PL).

T-Gen: The Seq2Seq model only taking textual context as input for generation in [24]. Comparing this baseline to other models with both textual and visual contexts establishes the improvement due to the visual context.

V&T-Gen: The best response generation model in [24] for the task of IGC. Comparing this model to our proposed V&T.JAM-CE indicates the improvement due to joint attention of visual and textual features in our model.

V&T-E-Gen: The best image-grounded dialogue generation model that incorporates image scene and sentiment understanding into Seq2Seq model in [25]. Comparing this model to our proposed V&T.JAM-CE indicates the improvement due to joint attention of visual and textual features in our model.

V&T-JAM-CE: Our proposed generative model with joint attention of visual and textual features trained with cross-entropy loss function. Comparing this variant to V&T-JAM-PL establishes the improvement due to the knowledge transferring from T to G.

4.5 Evaluation Results

Table 1 shows the results of human annotations. From scores for comments generated by different models, we can find that models with TCIAE (V&T-JAM-CE and V&T-JAM-PL) significantly outperform the baseline models (T-Gen, V&T-Gen and V&T-E-Gen) in terms of relevance and informativeness. Especially, V&T-JAM-PL gets the best performance among all models. Compared with V&T-JAM-CE, V&T-JAM-PL improves the proportion of "+2" comments by 21.2% and decreases the proportion of "0" comments by 11.5%, which confirms the benefits of knowledge transfer from topic embedding space to the generative comment model. The perceptual loss based on topic knowledge gives the model more space to generate more interesting and diverse comments within the topic relevant to the context. Similar to V&T-JAM-PL, V&T-JAM-CE outperforms V&T-E-Gen, V&T-Gen, and T-Gen. Compared to T-Gen, it generates more "+2" and "+1" comments, suggesting that visual context provides more useful information to generate relevant and content-rich comments when the textual context is ambiguous, or images contain content relevant to the textual context but is not mentioned explicitly. It is worth noting that V&T-Gen performs significantly worse than V&T-JAM-CE, even worse than T-Gen. The reason hiding behind is concatenation of textual and visual features without joint attention introduces much noise into generation, especially when visual context contains more than one image. When the textual vector is directly concatenated to several image feature vectors, it is difficult for decoder to differentiate useful information from noise. Similar to V&T-Gen, although V&T-E-Gen chooses image scene and sentiment as high-level image features, it still performs much worse than V&T-JAM-CE.

Except T-Gen, the Fleiss' kappa scores of all the other models range from 0.3 to 0.4, which indicates fair agreement. The kappa score of T-Gen is relatively higher as 0.435, which indicates moderate agreement, since comments only depending on textual context probably seem uncorrelated with the context, leading to more agreement in such cases. We also conduct sign test on all labelers' annotations to compare the statistic significance between V&T-JAM-PL and other baseline models. The difference is statistically significant ($p < 0.01$).

The results of automatic metrics are summarized in Table 2 consisting of the perplexity and numbers and ratios of distinct unigrams and bigrams on test data. The difference in perplexity of different models doesn't reach

Table 2: The results of automatic metrics

Models	PPL	Distinct-1	Distinct-2
T-Gen	186.65	1143/0.098	2250/0.311
V&T-Gen	351.96	965/0.119	2158/0.341
V&T-E-Gen	425.48	1051/0.132	2431/0.363
V&T-JAM-CE	458.88	1349/0.108	3848/0.308
V&T-JAM-PL	487.77	1565/0.171	4532/0.415

an order of magnitude. V&T-Gen achieves the lowest perplexity. However, that doesn't indicate the generation ability of T-Gen gets the best performance, because the length of comments generated by T-Gen is generally shorter than those generated by other models. Since we introduce visual and textual joint attention and a perceptual loss based on topic embeddings respectively into V&T-JAM-CE and V&T-JAM-PL, they tend to deviate from the training paradigm of MLE to enrich contents of outputs, resulting in the slight growth of perplexity. Our proposed models achieve higher numbers of distinct unigrams and bigrams than the other baseline models. That supports our proposal that the introduction of visual information is beneficial to improve the informativeness of comments. Among them, V&T-JAM-PL achieves the best performance both on numbers and ratios of distinct unigrams and bigrams, improving the performance of V&T-JAM-CE by 0.063 on distinct-1 ratio and 0.107 on distinct-2 ratio. These comparisons confirm the modification to encoder–decoder framework by visual and textual joint attention is not sufficient to diversify outputs, while the knowledge transferred from the pre-trained topic embedding space greatly improves *G*. Note that ratios of distinct unigrams and bigrams of V&T-JAM-CE and T-Gen are worse than those of V&T-Gen, apparently because comments generated by V&T-Gen are usually shorter than those generated by V&T-Gen and T-Gen. Our interpretation of this phenomenon is that the noise introduced by concatenation of visual and textual feature vectors weakens their ability to reply to complex contexts.

4.6 Qualitative Comparison

In Table 3, we present a couple of qualitative examples to compare comments generated by our proposed models and the other baseline models. We can find that T-Gen tends to generate generic and conservative comments such as "So beautiful", and sometimes even exists problems of semantic relevance. Confronted with complex multimodal context of multiple images and text, V&T-Gen and V&T-E-Gen are more likely to generate comments less relevant to the context, *e.g.* comments with mismatched entity name and logical problems. In contrast, V&T-JAM-CE and V&T-JAM-PL achieve better performance in semantic relevance. Furthermore,

Table 3: Qualitative comparison

Visual Context			
Textual Context	张艺兴《红秀grazia》第397期封面大片 Zhang Yixing's cover of the 397th issue of Grazia	柴犬的清凉夏日写真 Shiba inu's cool summer photos	走过北海道的雪，然后再一起去看鸭川的樱花 Walk through the snow of Hokkaido, then go to see cherry blossoms of Kamogawa
T-Gen	啊，好漂亮啊 Ah, so beautiful.	这是什么神仙地方？ What fairy place is this?	现在这个季节正是看樱花的季节 This is the season to see cherry blossoms
V&T-Gen	感谢vogueme分享，范丞丞带来的简直是惊喜。 Thanks to the sharing of vogueme, Fan Chengcheng brings a surprise.	美人都可以这么美，真好看 Beautiful people can be so beautiful, really beautiful	好美！找不到其他词汇形容它的美丽 How beautiful! There are no other words to describe its beauty
V&T-E-Gen	期待我们黄明昊的新作品，一起走花路吧 Look forward to new works of Huang Minghao. Hope everything goes well for him.	柴犬这笑容超治愈，治愈各种烦恼 Shiba inu's smile is super healing, healing all kinds of troubles	这组也可以做头像 This set of photos can also be used as an avatar.
V&T-JAM-CE	很喜欢，拍得很好 I love it. It was well taken.	夏天到了！ Summer comes	好美的风景，好美的以轩姐。 What a beautiful scenery. What a beautiful lady.
V&T-JAM-PL	这扑面而来的少年感，心动 His juvenile sense blowing on my face is exciting.	柴犬治愈般的笑容，好像小天使 Shiba inu has a healing smile, like an angel	好美啊，什么时候能感受到这样的雪啊 So beautiful. When can I feel this kind of snow?

compared to V&T-JAM-PL, the content of comments generated by V&T-JAM-CE is not rich enough, without stepping outside the box of safe and generic responses. V&T-JAM-PL reduces a lot of such situations and tend to generate far more informative and interesting comments.

5. CONCLUSION

We introduce a new task of visual social comment, which needs to generate informative and interesting comments on published posts containing multiple images and a text. In this multimodal context task, we propose to fuse visual and textual information by joint attention instead of concatenation to improve informativeness in output comments. Since Seq2Seq models trained with MLE training paradigm tend to generate generic responses, we also introduce a topic classification model to transfer topic embedding knowledge to the generative model. Experimental results of automatic evaluation metrics and human annotations both show that our proposed model can generate more informative and diverse comments than the state-of-the-art generation models.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author(s).

FUNDING

This work was supported in part by the National Natural Science Foundation of China [U1636206], [61525203], [U1936214], [61902235]. It was also supported by "Chen Guang" project co-funded by the Shanghai Municipal Education Commission and Shanghai Education Development Foundation.

ORCID

Yue Yin http://orcid.org/0000-0002-9256-9506

REFERENCES

1. E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Commun. ACM.*, Vol. 59, pp. 96–104, Jul. 2016.

2. A. Karataş, and S. A. Şahin. "A review on Social Bot Detection techniques and research directions," in *Proceedings of the International Conference on Information Security and Cryptology*, Ankara, Oct. 20–21, 2017, pp.156–161.

3. S. Weibo. Available: https://www.weibo.com/.

4. Y. Wu, W. Wu, C. Xing, M. Zhou, and Z. Li. "Sequential matching network: a new architecture for multi-turn response selection in retrieval-based chatbots," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Jul. 30-Aug. 4, 2017, pp. 496–505.

5. X. Zhou, L. Li, D. Dong, Y. Liu, Y. Chen, W. X. Zhao, D. Yu, and H. Wu. "Multi-turn response selection for chatbots with deep attention matching network," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Jul. 15–20, 2018, pp. 1118–1127.

6. C. Tao, W. Wu, C. Xu, W. Hu, D. Zhao, and R. Yan. "Multi-representation fusion network for multi-turn response selection in retrieval-based chatbots," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, Melbourne, Feb. 11–15, 2019, pp. 267–275.

7. A. Ritter, C. Cherry, and W. B. Dolan. "Data-driven response generation in social media," in *Proceedings of the conference on empirical methods in natural language processing*, Edinburgh, Jul. 27–31, 2011, pp. 583–593.

8. L. Shang, Z. Lu, and H. Li. "Neural responding machine for short-text conversation," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, Jul. 26–31, 2015, pp. 1577–1586.

9. I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau. "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Thirtieth AAAI Conference on Artificial Intelligence*, Phoenix, Feb. 12–17, 2016, pp. 3776–3783.

10. C. Xing, W. Wu, Y. Wu, J. Liu, Y. Huang, M. Zhou, and W. Y. Ma. "Topic aware neural response generation," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, Feb. 4–9, 2017, pp. 3351–3357.

11. C. Xing, Y. Wu, W. Wu, Y. Huang, and M. Zhou. "Hierarchical recurrent attention network for response generation," in *Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Feb. 2–7, 2018, pp. 5610–5617.

12. D. Bahdanau, K. Cho, and Y. Bengio. "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.

13. M. T. Luong, H. Pham, and C. D. Manning. "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Sep. 17–21, 2015, pp. 1412–1421.

14. L. Mou, Y. Song, R. Yan, G. Li, L. Zhang, and Z. Jin. "Sequence to backward and forward sequences: a content-introducing approach to generative short-text conversation," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, Osaka, Dec. 11–16, 2016, pp. 3349–3358.

15. M. Ghazvininejad, C. Brockett, M. W. Chang, B. Dolan, J. Gao, W. T. Yih, and M. Galley. "A knowledge-grounded neural conversation model," in *Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Feb. 2–7, 2018, pp. 5110–5117.

16. J. Li, M. Galley, C. Brockett, and B. Dolan. "A diversity-promoting objective function for neural conversation models," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, Jun. 12–17, 2016, pp. 110–119.

17. J. Li, and D. Jurafsky. "Mutual information and diverse decoding improve neural machine translation," arXiv preprint arXiv:1601.00372, 2016.

18. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Boston. Jun. 7–12, 2015, pp. 3156–3164.

19. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. S. Zemel, and Y. Bengio. "Show, attend and tell: neural image caption generation with visual attention," in *International conference on machine learning*, Lille, Jul. 6–11, 2015, pp. 2048–2057.

20. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. "Vqa: visual question answering," in *Proceedings of the IEEE international conference on computer vision*, Santiago, Dec. 7–15, 2015, pp. 2425–2433.

21. A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. "Visual dialog," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Jul. 21–26, 2017, pp. 326–335.

22. L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann, "Uncovering the temporal context for video question answering," *Int. J. Comput. Vis.*, Vol. 124, pp. 409–421, Jul. 2017.

23. K. H. Zeng, T. H. Chen, C. Y. Chuang, Y. H. Liao, J. C. Niebles, and M. Sun. "Leveraging video descriptions to learn video question answering," in *Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, Feb. 4–9, 2017, pp. 4334–4340.

24. N. Mostafazadeh, C. Brockett, B. Dolan, M. Galley, J. Gao, G. Spithouraki, and L. Vanderwende. "Image-grounded conversations: multimodal context for natural question and response generation," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, Taipei, Nov. 27-Dec.1, 2017, pp. 462–472.

25. B. Huber, D. McDuff, C. Brockett, M. Galley, and B. Dolan. "Emotional dialogue generation using image-grounded language models," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal, Apr. 21–26, 2018, pp. 277.

26. J. Lu, A. Kannan, J. Yang, D. Parikh, and D. Batra. "Best of both worlds: transferring knowledge from discriminative learning to a generative visual dialog model," in *Advances in Neural Information Processing Systems*, Long Beach, Dec. 4–9, 2017, pp. 314–324.

27. J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky. "Adversarial learning for neural dialogue generation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Sep. 7–11, 2017, pp. 2157–2169.

28. J. Johnson, A. Alahi, and L. Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*, Amsterdam, Oct. 8–16, 2016, pp. 694–711.

29. K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, Jun. 26- Jul. 1, 2016, pp. 770–778.

30. S. Lai, L. Xu, K. Liu, and J. Zhao. "Recurrent convolutional neural networks for text classification," in *Twenty-ninth AAAI conference on artificial intelligence*, Austin, Jan. 25–30, 2015, pp. 2267–2273.

31. C. J. Maddison, A. Mnih, and Y. W. Teh. "The concrete distribution: A continuous relaxation of discrete random variables," arXiv preprint arXiv:1611.00712, 2016.

32. E. Jang, S. Gu, and B. Poole. "Categorical reparameterization with gumbel-softmax," arXiv preprint arXiv:1611.01144, 2016.

33. FoolNLTK, Available: https://github.com/rockyzhengwu/FoolNLTK

34. Y. Song, S. Shi, J. Li, and H. Zhang. "Directional skip-gram: Explicitly distinguishing left and right context for word embeddings," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Jun. 1–6, 2018, pp. 175–180.

35. O. Vinyals, and Q. V. Le. "A Neural Conversational Model," arXiv preprint arXiv:1506.05869, 2015.

36. K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*, Philadelphia, Jul. 7–12, 2002, pp. 311–318.

37. C. W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau. "How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Nov. 1–5, 2016, pp. 2122–2132.

38. J. L. Fleiss, and J. Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability," *Educ. Psychol. Meas.*, Vol. 33, pp. 613–619, Oct. 1973.