SCIENCE CHINA
Information Sciences

• MOOP •

July 2020, Vol. 63 174101:1–174101:3
https://doi.org/10.1007/s11432-018-9908-4

Recursive narrative alignment for movie narrating

Zhongyi HAN[1,3], Hongbo WU[2], Benzheng WEI[3*], Yilong YIN[1*] & Shuo LI[4]

[1]*School of Software, Shandong University, Jinan* 250101*, China;*
[2]*School of Business, St. Lawrence College, Kingston* K7L 5A6*, Canada;*
[3]*Computational Medicine Lab, Shandong University of Traditional Chinese Medicine, Jinan* 250355*, China;*
[4]*Digital Imaging Group of London, Western University, London* N6A 4V2*, Canada*

Received 24 December 2018/Revised 6 March 2019/Accepted 6 May 2019/Published online 16 September 2019

Citation Han Z Y, Wu H B, Wei B Z, et al. Recursive narrative alignment for movie narrating. Sci China Inf Sci, 2020, 63(7): 174101, https://doi.org/10.1007/s11432-018-9908-4

Movie narrating can enable the capture of not only the subject matter but also the emotive essence of the subject. For example, "having a good time" or "feeling exhausted", which relates to the narrative of "holding a birthday party". Movie narrating converts the human-like expressivity in movie shots (i.e., sets of film frames) into a cohesive and coherent narrative (i.e., sentences that tell a logical story). Besides being essential for a human-like understanding of images, movie narrating can realize numerous new applications such as the automatic emotive narrating of photograph albums on social media, automatic logical summary of trips, or diary generation of events. Moreover, as an interdisciplinary field spanning computer vision, natural language processing, and philosophy, movie narrating can potentially elevate artificial intelligence from basic understanding toward human-like understanding.

Movie narrating methods have not been proposed previously as they must overcome the formidable challenges of latent relatedness, weak consistency, and emotive state conflicts between the frames and underlying narrative [1]. To overcome these challenges, we propose a recursive narrative alignment framework to generate movie narratives. As illustrated in Figure 1, the proposed framework adaptively aligns visual cues with keywords using a semantic-attention mechanism, thereby improving the frame-narrative coherence. Furthermore, it recursively applies the contextual expression in previous frames into the current frame to improve the narrative cohesion. Finally, the emotive conflict between frame and story is resolved by our newly designed regularizer that minimizes the style-manifold distance.

Visual cue alignment. The visual cue alignment module improves the context relatedness among descriptions of subsequent events. As an attention model, it projects the combined global and local features into a latent space. In particular, let $V_t^G(x_t)$ and $V_t^L(x_t)$ represent the high-level global and local information in the current frame x_t, respectively, among a sequence of frames X_i. The global features $V_t^G = \mathrm{VGG16}(x_t)$ are extracted using the last convolutional layer of a VGG16 [2] convolutional neural network after global pooling. The local features $V_t^L = \mathrm{FasterRCNN}(x_t)$ are computed from the region proposals detected by faster RCNN [3]. The V_t^L are the highest-dimensional outputs of the fully-connected layer of the top K region proposals after non-maximum suppression. The dimensions of the global and local features are $\mathbb{R}^{1\times D^G}$ and $\mathbb{R}^{K\times D^L}$, respectively. To ensure that each word in the current frame's description contains both global and local information, the global V_t^G and local features V_t^L are projected with dimensions of $\mathbb{R}^{1\times D}$ and $\mathbb{R}^{K\times D}$, respectively, onto a common feature space. On this space, the features are summed together with broadcasting as follows:

$$V_t^{GL} = f(W^G V_t^G + b^G) + f(W^L V_t^L + b^L), \quad (1)$$

where $f(\cdot)$ is a nonlinear activation function. W^G

[*] Corresponding author (email: wbz99@sina.com, ylyin@sdu.edu.cn)

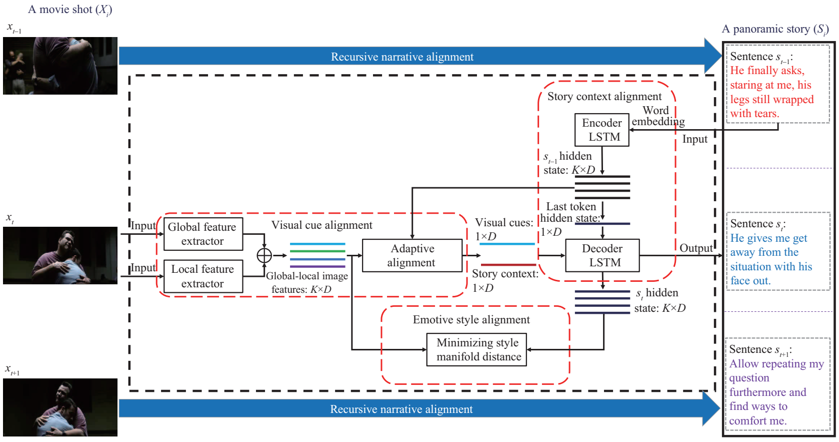
A movie shot (X_i)
x_{t-1}
x_t
x_{t+1}
Recursive narrative alignment
Recursive narrative alignment
A panoramic story (S_i)
Sentence s_{t-1}:
He finally asks, staring at me, his legs still wrapped with tears.
Sentence s_t:
He gives me get away from the situation with his face out.
Sentence s_{t+1}:
Allow repeating my question furthermore and find ways to comfort me.
Story context alignment
Word embedding
Encoder LSTM
Input
s_{t-1} hidden state: $K\times D$
Last token hidden state: $1\times D$
Decoder LSTM
Output
s_t hidden state: $K\times D$
Input
Input
Global feature extractor
Local feature extractor
Visual cue alignment
Global-local image features: $K\times D$
Adaptive alignment
Visual cues: $1\times D$
Story context: $1\times D$
Emotive style alignment
Minimizing style manifold distance

Figure 1 (Color online) Our recursive narrative alignment framework comprises three modules: visual cue alignment, story context alignment, and emotive style alignment. The framework recursively generates sentences to span the movie narrative in a film shot of unlimited size.

and W^L denote the weights of the global and local features, respectively. b^G and b^L are the biases of the global and local features, respectively. After the projection, the dimension of V^{GL} is $\mathbb{R}^{K\times D}$.

The next task aligns the global-local features V^{GL} with the word-embedding space of the previous frame description. Therefore, an adaptive alignment function is designed as the following semantic process:

$$C_t^V = \sum_{k=1}^{K} \alpha_t V_{t_k}^{GL}, \quad C_t^T = \sum_{k=1}^{K} \alpha_t H_{(t-1)_k}, \tag{2}$$

where C^V represents the visual cues of dimension $\mathbb{R}^{1\times D}$, and C^T is the story context of dimension $\mathbb{R}^{1\times D}$. The adaptive attention coefficient α for aligning the visual cues with story context is semantically computed as follows:

$$\alpha_t = \frac{\exp(z_{t_k})}{\sum_{k=1}^{K} \exp(z_{t_k})}, \quad \forall t \in (1, \ldots, T). \tag{3}$$

In this expression,

$$z_{t_k} = W^z \tanh(W^V V_{t_k}^{GL} + W^H H_{(t-1)_k}), \tag{4}$$

where K is the number of words in the previous frame description and H_{t-1} is the hidden feature of the word embedding in the previous frame description. The visual features and text descriptions are weighted by W^V and W^H, respectively.

Story context alignment. To model the contextual correlation of a movie narrative, the story context alignment module uses two long short-term memory (LSTM)-based recurrent neural networks (RNNs): an encoder LSTM for the previous story context and a decoder LSTM for inferring the current story context. The encoder LSTM transforms a source sentence into a sequence of vectors, each representing one token in the source sentence. Considering a list of hidden vectors and global-local frame features and after the visual cue alignment, the decoder LSTM uses the story and frame contexts to produce a coherent description of the current frame, one token at a time.

Encoder LSTM. As illustrated in Figure 1, the encoder LSTM captures the context of the past story events to provide continuity between the frame descriptions. The encoder inputs a sequence of token embeddings $s_{t-1} = w_1, \ldots, w_K$, $\forall w \in \mathbb{R}^E$ and transforms them into a sequence of hidden representations $H_{t-1} = (h_1, \ldots, h_K)$, $\forall h_k \in \mathbb{R}^D$. The transformation formula is

$$h_k = \mathrm{LSTM}^{\mathrm{enc}}(w_1, \ldots, w_{k-1}). \tag{5}$$

Decoder LSTM. To encourage a consistent narrative in the storytelling, we generate our frame descriptions recursively from the contextual features in the narrative and frames. The current frame description $s_t = y_1, \ldots, y_K$ is built token-by-token using a modified LSTM unit, which inputs the story context C^T, visual cues C^V, and previously generated tokens of the current frame:

$$p(y_k) = \mathrm{LSTM}^{\mathrm{dec}}(y_0, y_1, \ldots, y_{k-1}; C^T; C^V). \tag{6}$$

Here, y_0 is a special START token and $p(y_k)$ is the probability of the k-th word after Softmax.

Emotive style alignment. To avoid the discrepancy of describing a "sad" scene with "happy" language, we introduce a regularizer that minimizes the emotive style differences between the frame and text (Figure 1). Because the frame and language share common features in an emotive representation, we define the emotive style manifold distance between scene x_t and description s_t:

$$\mathcal{L}_{\text{style}} = ||\mathcal{G}_V - \mathcal{G}_S||^2, \tag{7}$$

where $\mathcal{G}_V$ and $\mathcal{G}_S$ denote the emotive style of the scene and corresponding frame description, respectively. The emotive style of frame x_t is captured by extracting the global-local feature V_t^{GL} using Eq. (1).

The emotive style of a sentence s_t is represented by the hidden state H_t of the decoder LSTM. To ensure consistency in the emotive style, we minimize the difference between the Gram matrices of the artistic style V^{GL} and linguistic style H:

$$\mathcal{G}_V(V^{GL}) = V^{GL}(V^{GL})^{\mathrm{T}}, \quad \mathcal{G}_S(H) = HH^{\mathrm{T}}. \tag{8}$$

Minimizing the high-level manifold distance improves the consistency between the emotive styles in the frame and text.

Memory efficient hybrid training strategy. The final objective function is a hybrid loss function that combines the text loss and emotive style difference loss. In particular, the framework is optimized recursively as follows:

$$\mathcal{L}(x_t; s_t, s_{t-1}) = \mathcal{L}_{\text{text}}(x_t; s_{t-1}) + \phi\mathcal{L}_{\text{style}}(x_t; s_t). \tag{9}$$

The recursive optimization negates the requirement of a fixed movie shot size during training. This implies the high generalizability of our framework because shots with an unlimited number of frames can be used during training. Moreover, because the whole shot does not need to be loaded into memory, our framework reduces the memory overhead by T-fold (where T is the number frames in a movie shot).

Experiments. We trained and validated our approach on the Movie-Book dataset [4], which contains 11 annotated movies. Each movie shot (approximately six-second frames) in the movie was manually annotated with sentences from its corresponding book. We randomly selected 80% of the shots for training, 10% for validation, and 10% for testing. A movie narrative was generated via processing the whole film frame-by-frame. Following the common protocol of sequence-to-sequence decoding [5], we heuristically selected the best frame description by a beam search of varying width ($B = \{1, 5, 10\}$). Each frame description was generated by assigning an initial START token to the decoder LSTM and allowing it to sample successive tokens until it reached a STOP token, or until $K = 32$ tokens have been sampled. Our model tokenizes each word by a word embedding of size $E = 256$, and represents the frame and text contexts in a latent space of $D = 512$. To understand the contribution of each module to the quality of the generated story, we systematically evaluated the effects of visual cue alignment, story context alignment, and emotive style alignment in isolation. After global pooling, the global features V_t^G (obtained by VGG16 pretrained on ImageNet) were reduced to $\mathbb{R}^{D=512}$ and duplicated 32 times to match the number of local features. The local features V_t^L of the region proposals were extracted by Faster-RCNN pretrained on MS-COCO. The empirical value $\phi = 100$ yielded the most accurate results; therefore, it was used in all tested models. All models were trained via the mini-batch stochastic gradient descent method over 1000 iterations using the Adam optimizer implemented in Python using Keras and Tensorflow.

We computed the widely adopted automated machine translation metrics BLEU, ROUGE-L, METEOR, and CIDEr [6]. Recursive narrative alignment achieved a BLEU1 of 25.82%, a BLEU2 of 10.77%, a BLEU3 of 5.30%, a BLEU4 of 3.21%, a CIDEr of 7.63%, a METEOR of 12.54%, and a ROUGE-L of 14.38%. These results demonstrate that our method achieves the highest overall performance among baselines.

Acknowledgements This work was partly funded by National Natural Science Foundation of China (Grant Nos. 61872225, 61876098, 61573219), Natural Science Foundation of Shandong Province (Grant No. ZR2015FM010), Project of Science and Technology Plan of Shandong Higher Education Institutions Program (Grant No. J15LN20), and Project of Shandong Province Medical and Health Technology Development Program (Grant No. 2016WS0577).

Supporting information Videos and other supplemental documents. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Johnson J, Hariharan B, van der Maaten L, et al. Inferring and executing programs for visual reasoning. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2017

2 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv:1409.1556

3 Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 1137–1149

4 Zhu Y K, Kiros R, Zemel R, et al. Aligning books and movies: towards story-like visual explanations by watching movies and reading books. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2015

5 Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks. 2014. ArXiv:1409.3215

6 Sharma S, Asri L E, Schulz H, et al. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. 2017. ArXiv:1706.09799