SCIENCE CHINA PRESS

Recursive Narrative Alignment for Movie Narrating

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

1. School of Software, Shandong University
2. St. Lawrence College, Kingston, Canada
3. Computational Medicine Lab, Shandong University of Traditional Chinese Medicine
4. Digital Imaging Group of London, Western University

Outline

- Introduction
  - Definition
  - Significance
  - Challenge
- Related Work
- Our Method
- Experiment
- Conclusion

SCIENCE CHINA PRESS

Definition

Movie narrating aims to capture not only the subject matter but also the emotive essence of film frames into a story, which would make it possible to capture not only the subject matter but also the emotive essence of "having a good time" or "being exhausted" while relating to the narrative of "holding a birthday party".

Automated movie narrating for the Fight Club film: He finally asks, staring at me, his legs still wrapped with tears. He gives me get away from the situation with his face out. Allow repeating my question furthermore and find ways to comfort me. Lay a while, then, by the end.

Significance

- Movie narrating entails the human-like expressivity of movie shots (i.e. sets of film frames) into a cohesive and coherent narrative (i.e. sentences that tell a logical story)
- Help reaching the human-like understanding of images
- Opens up numerous new applications such as automatic emotive narrating of social media photo albums, an automatic logical summary of trips or diary generation of events
- It is an interdisciplinary field spanning computer vision, natural language processing, and philosophy, which aims to move artificial intelligence from basic understanding towards human-like understanding

Challenges

- **The latent relatedness**

  For instance, in the sentence “The sky was illuminated with a brilliance of orange hues”, it is imperative to correlate the key words “brilliance” and “orange hues” to the imagery of a picturesque sunset.

- **The weak consistency**

  Movie narrating requires film frames be described with cohesive sentences to reason about the semantic content of images.

- **The emotive state conflict**

  Movie narrating requires consistency in the emotive state between the frames and narrative. For example, frames depicting the brilliance of sunset should correspond to positive sentiment whereas images of war should be described with a negative connotation.

Related Methods

- Can generate generic descriptions of a single image or video
- Fail to convey the rich emotive content of movie shots (sets of film frames) as a cohesive and coherent story
- No movie narrating method has been proposed before

A Photo Album:



Conventional Captioning Method:

There is a beach. There is a living room. There are trees.

Our Captioning Method:

Today I went to the beach. It was very windy outside. I'd have more fun at home.

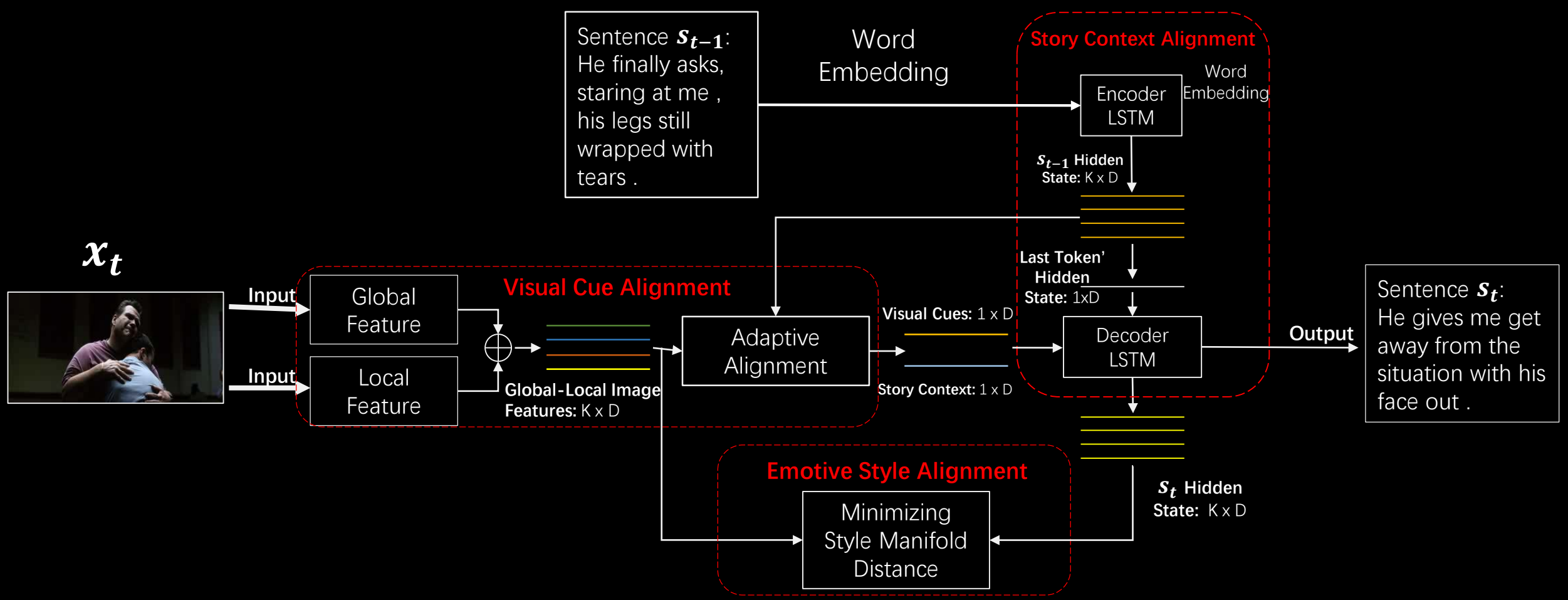
Our Method

We propose the **Recursive Narrative Alignment** method for overcoming aforementioned challenges.

Contributions:

- It is the first movie narrating system
- **It** comprises three novel alignment models
- It successfully achieves movie narrating after 11 films containing about 20,000 shots aligned to their respective novels.
- It is able to not only solve basic computer vision challenges such as describing what a person in the image is doing but also use high-level abstractions

Recursive Narrative Alignment Framework



A movie shot (X_i)

x_{t-1}

Recursive Narrative Alignment

A panoramic story (S_i)

Sentence s_{t-1}: He finally asks, staring at me , his legs still wrapped with tears .

Recursive Narrative Alignment

x_t

Input

Input

Global Feature

Local Feature

Visual Cue Alignment

Global-Local Image Features: K x D

Adaptive Alignment

Visual Cues: 1 x D

Story Context: 1 x D

Story Context Alignment

Word Embedding

Encoder LSTM

Input

s_{t-1} Hidden State: K x D

Last Token' Hidden State: 1xD

Decoder LSTM

Output

Sentence s_t: He gives me get away from the situation with his face out .

Emotive Style Alignment

Minimizing Style Manifold Distance

s_t Hidden State: K x D

x_{t+1}

Recursive Narrative Alignment

Sentence s_{t+1}: Allow repeating my question furthermore and find ways to comfort me.

Recursive Narrative Alignment

- Visual Cue Alignment
  - It uses a semantic attention mechanism to adaptively align visual cues with keywords for a better frame-narrative coherence.
  - The attention mechanism comprises global and local information from each frame. It is adaptively combined into objective keywords and subjective emotive words.
- Story Context Alignment
  - The framework then recursively applies the contextual expression of previous frames into the current frame to improve the cohesion of the narrative.
- Emotive Style Alignment
  - The emotive conflict between frame and story is resolved by our newly-designed style regularize, which minimizes the style manifold distance between the file frame and story.

Visual Cue Alignment

Let $V_t^G(x_t)$ be global feature from VGG16 neural network, while $V_t^L(x_t)$ is local feature that is computed from region proposals detected by Faster-RCNN. The global and local feature are projected into V_t^{GL}:

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

where $f(\cdot)$ is a non-linear activation function, W^G and W^L are the weights of global and local features, respectively. b^G and b^L are the bias of the global and local features respectively. The task is then to align the global-local features V_t^{GL} with the word embedding space of the previous frame description. To achieve that, an adaptive alignment function is designed as the following semantic process:

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

where C_t^V represents the visual cues, while C_t^T is the story context with dimension. α_t is an attention coefficient of the joint alignment between current frame feature and previous frame description. The adaptive coefficient α applied to align visual cues and story context is semantically computed as follows:

$$\alpha_t = \frac{exp(z_{t_k})}{\sum_{k=1}^{K} exp(z_{t_k})}, \forall t \in (1, \cdots, T), \quad \text{in which } z_{t_k} = W^z \tan\mathrm{h}(W^V V_{t_k}^{GL} + W^H H_{(t-1))\ \ k}).$$

Story Context Alignment

The Story Context Alignment is capable of modeling the contextual correlation of a movie narrative using two LSTM based RNNs: an encoder LSTM for previous story context and decoder LSTM for inferring current story context.

- Encoder LSTM

  It captures the context of past story events in order to provide continuity between frame descriptions. The encoder takes as input a sequence of token embeddings $s_{t-1} = w_1, \cdots, w_K, \forall w \in \mathbb{R}^E$ and transforms them into a sequence of hidden representation $H_{t-1} = (h_1, \cdots, h_K)$, using the following formula:

$$h_k = LSTM^{enc}(w_1, \cdots, w_{k-1}).$$

- Decoder LSTM

  In order to encourage the storytelling to have a consistent narrative, we generate our frame descriptions recursively by relying on the narrative and frame context features. We generate the current frame description $s_t = y_1, \cdots, y_K$ one token at a time by using a modified LSTM unit, which accepts the story context C^T, visual cues C^V and the previously generated tokens of current frame as input:

$$p(y_k) = LSTM^{dec}(y_0, y_1, \cdots, y_{k-1}; C^T; C^V),$$

  where y_0 is a special START token, and $p(y_k)$is the probability of the k word after SoftMax.

Emotive Style Alignment

In order to avoid the discrepancy of describing a “sad” scene using “happy” language, we introduce a regularize that minimizes the difference in emotive style between frame and text. Since the frame and language share commonplace in emotive representation, we define the emotive style manifold distance between scene and description as follows:

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

where $\mathcal{G}_V$ is the emotive style of the scene and $\mathcal{G}_S$ is the emotive style of the corresponding frame description.
The emotive style of sentence is represented by the hidden state of the decoder LSTM. The emotive style consistency is thus ensured by minimizing the difference between the Gram matrix of artistic style $V^{GL}$and linguistic style H:

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

Memory Efficient Hybrid Training Strategy

The final objective function is a hybrid loss function which combines text loss and emotive style difference loss. Specifically, the framework is optimized recursively as follows:

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

$$\theta^* = \text{argmin}_\theta - \sum_{t=1}^{T} \log p(s_t | LSTM(C_t^V(x_t), C_t^T(s_{t-1})); \theta) + ||\mathcal{G}_V - \mathcal{G}_S||^2,$$

for each frame $x_t$in the movie shot X_t Since the first frame x_1 in the sequence does not have a previous sentence, we initialize it to a zero vector $S_0 = 0$ In order to avoid the model being likely to generating zero padding tokens, we also add a conditioner to reduce the weight of zero padding tokens when training.

Advantage: Since our framework is recursively optimized, we do not require a fixed movie shot size during training. This implies that not only is our model more generalizable since shots with an unlimited number of frames can be used during training, it also provides a T-fold (where T is the number frames in a movie shot) reduction in memory overhead since we don't need to load the whole shot into memory at once.

Experimental Settings

- Movie-Book datasets
  - 11 annotated movies
  - Each movie shot (about six-second frames) in the movie was manually annotated with sentences from its corresponding book
  - 80% of shots for training, 10% for validation and 10% for testing
  - Each movie has on average 1,800 shots annotated with an average of 7,750 sentences
- Settings
  - Beam search = $\{1,5,10\}$
  - Word embedding size is 256
  - Faster-RCNN pretrained on MS-COCO
  - Mini-batch stochastic gradient descent for 1,000 iterations

Results

Style Consistency	Beam Width	BLEU1	BLEU2	BLEU3	BLEU4	CIDEr	METEOR	ROUGE-L
RNA	B=1	8.22	2.97	1.34	0.80	1.34	6.33	10.52
	B=5	**26.19**	10.69	4.93	2.90	7.21	12.28	14.16
	B=10	25.82	**10.77**	**5.30**	**3.21**	**7.63**	**12.54**	**14.38**
RNA-noalignment	B=1	17.63	5.54	1.12	0.40	1.10	7.13	12.68
	B=5	17.85	5.51	1.26	0.40	1.45	9.46	12.21
	B=10	19.97	6.14	1.12	0.38	2.22	9.76	12.66
RNA-nocontext	B=1	18.53	5.93	1.92	0.82	2.92	8.81	12.53
	B=5	24.41	8.74	3.07	1.19	3.43	12.03	12.97
	B=10	24.08	8.77	3.22	1.27	3.54	11.94	12.64
RNA-nostyle	B=1	0.00	0.00	0.00	0.00	0.03	1.33	1.41
	B=5	20.50	7.34	2.43	1.03	1.60	9.10	13.20
	B=10	21.01	6.95	1.99	0.78	1.35	9.38	13.54

The Recursive Narrative Alignment framework achieved the highest performance based on four Machine Translation metrics. The references of generated movie narrating are from movie books. The baseline model without Visual Cue Alignment as RNA-noattention, without story context as RNA-nocontext, and without emotive style as RNA-nostyle. The four metrics can produces high correlation with human judgment. The higher the metric values are, the better the model is. All values are reported as precentage (%)}.

Results

Movie narratives generated using the Fight Club film of the Movie-Book Data

RNA: He finally asks , staring at me , her legs still wrapped in tears. Doing the riverbank in a faded and lay a while , then , by the end. she gives me get away from the situation with his face out. Allowed to rephrase my question and found ways to spend out of her face.

RNA-noattention: He finally asks , after every guy. He asks tiredly. He asks , vaguely concerned. Carbohydrates that penetrate the hair shafts for improved strength and shine.

RNA-nocontext: The family was excited to go to the school. The kids were able to go to the beach. The kids were very excited to see. The kids were very excited to be there. The kids were very excited to see the kids.

RNA-nostyle: The family was waiting for the family to the family. The band was a little boy. The family was a great day. The family was very happy. The crowd was very happy.

Results

Movie narratives generated using the Gone Girl film of the Movie-Book Data

RNA: He finally asks , as if he'd been putting off the whole week. He had some hard to be any of this way to pay the other. He looks up in a low , young , while we don't see. The door was pressed against the wall , he looked at the room. “She has a good job ”, he said uplifting his arms.

RNA-noattention: And she'd got it off a man's his *** stare , either in. To his moment going right with him pulled at all with laughter? Nobody bang bang used over 22 about - or slow down wall just chuckled. “Warn counter he will never strikes at the explanation room so'' he said. Swings please , huh ? open light beautiful *** by this other cup yet.

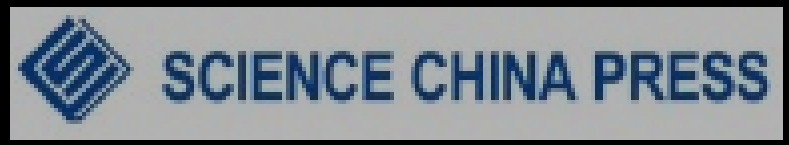
RNA-nocontext: Them months he says why wanted stay over , sat grinning how followed *****. Not very enough given quickly opened the money card back in and other eye. `` annoyed in hell shook roof here or ? Stock Her standing an sort of ice else to know both of here who here. People good lot in on an side was about more to keep years ?

RNA-nostyle: Of the magazines be road ?'' upward a good fifth white upward. Manager Feet , and he looked at the boy said , a good fathern’t. `` , he looked up at the boy , and then he said. The road and then , and then looked at the road the boy said. And stood looking at the road , and then he said the road road.

Reference: He asks the girl if he should wait another whole week. He is so depressed that he want to say anymore. He looks up a young girl, while we cannot see. He looked at the room through the door. “She has a good job and please to find her”, he said throwing up his arms in despair.

Conclusion

- For the first time, our proposed movie narrative enables the retelling of human-like coherent and cohesive narrative from film frames of unlimited size.
- The proposed Recursive Narrative Alignment framework is the first vision-language model capable of generating movie narratives by aligning visual cues, story contexts, and emotive style.
- The proposed Emotive Style Regularize is the first regularization method to resolve the problem of emotive state conflict between frame and narrative by aligning the sentiment of the text to the atmospheric mood of the frame.



Thanks