



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

Ph.D. DISSERTATION

Narrative Text Generation via Latent
Embedding from Visual Stories

잠재 임베딩을 통한 시각적 스토리로부터의 서사 텍스트
생성기 학습

BY

Min-Oh Heo

FEBRUARY 2019

DEPARTMENT OF ELECTRICAL ENGINEERING AND
COMPUTER SCIENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Ph.D. DISSERTATION

Narrative Text Generation via Latent
Embedding from Visual Stories

잠재 임베딩을 통한 시각적 스토리로부터의 서사 텍스트
생성기 학습

BY

Min-Oh Heo

FEBRUARY 2019

DEPARTMENT OF ELECTRICAL ENGINEERING AND
COMPUTER SCIENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Narrative Text Generation via Latent Embedding
from Visual Stories

잠재 임베딩을 통한 시각적 스토리로부터의 서사
텍스트 생성기 학습

지도교수 장 병 탁

이 논문을 공학박사 학위논문으로 제출함

2018 년 12 월

서울대학교 대학원

전기.컴퓨터 공학부

허 민 오

Min-Oh Heo의 공학박사 학위논문을 인준함

2018 년 12 월

위 원 장	최 진 영
부위원장	장 병 탁
위 원	정 교 민
위 원	김 건 희
위 원	하 정 우

Abstract

The ability to understand the story is essential to make humans unique from other primates as well as animals. The capability of story understanding is crucial for AI agents to live with people in everyday life and understand their context. However, most research on story AI focuses on automated story generation based on closed worlds designed manually, which are widely used for computation authoring. Machine learning techniques on story corpora face similar problems of natural language processing such as omitting details and commonsense knowledge. Since the remarkable success of deep learning on computer vision field, increasing our interest in research on bridging between vision and language, vision-grounded story data will potentially improve the performance of story understanding and narrative text generation.

Let us assume that AI agents lie in the environment in which the sensing information is input by the camera. Those agents observe the surroundings, translate them into the story in natural language, and predict the following event or multiple ones sequentially. This dissertation study on the related problems: learning stories or generating the narrative text from image streams or videos. The first problem is to generate a narrative text from a sequence of ordered images. As a solution, we introduce a GLAC Net (Global-local Attention Cascading Network). It translates from image sequences to narrative paragraphs in text as a encoder-decoder framework with sequence-to-sequence setting. It has convolutional neural networks for extracting information from images, and recurrent neural networks for text generation. We introduce visual cue encoders with stacked bidirectional LSTMs, and all of the outputs of each layer are aggregated

as contextualized image vectors to extract visual clues. The coherency of the generated text is further improved by conveying (cascading) the information of the previous sentence to the next sentence serially in the decoders. We evaluate the performance of it on the Visual storytelling (VIST) dataset. It outperforms other state-of-the-art results and shows the best scores in total score and all of 6 aspects in the visual storytelling challenge with evaluation of human judges. The second is to predict the following events or narrative texts with the former parts of stories. It should be possible to predict at any step with an arbitrary length. We propose recurrent event retrieval models as a solution. They train a context accumulation function and two embedding functions, where make close the distance between the cumulative context at current time and the next probable events on a latent space. They update the cumulative context with a new event as a input using bilinear operations, and we can find the next event candidates with the updated cumulative context. We evaluate them for Story Cloze Test, they show competitive performance and the best in open-ended generation setting. Also, it demonstrates the working examples in an interactive setting.

The third deals with the study on composite representation learning for semantics and order for video stories. We embed each episode as a trajectory-like sequence of events on the latent space, and propose a ViStoryNet to regenerate video stories with them (tasks of story completion). We convert event sentences to thought vectors, and train functions to make successive event embed close each other to form episodes as trajectories. Bi-directional LSTMs are trained as sequence models, and decoders to generate event sentences with GRUs. We test them experimentally with PororoQA dataset, and observe that most of episodes show the form of trajectories. We use them to complete the blocked part of stories, and they show not perfect but overall similar result.

Those results above can be applied to AI agents in the living area sensing with their cameras, explain the situation as stories, infer some unobserved parts, and predict the future story.

Keywords: Visual Storytelling, Narrative Text Generation, Next Event Prediction, Story Completion, Global-local Attention, Recurrent Event Retrieval Model, Successive Event Order Embedding

Student Number: 2005-21534

Contents

Abstract	i
Chapter 1 Introduction	1
1.1 Story of Everyday lives in Videos and Story Understanding . . .	1
1.2 Problems to be addressed	3
1.3 Approach and Contribution	6
1.4 Organization of Dissertation	9
Chapter 2 Background and Related Work	10
2.1 Why We Study Stories	10
2.2 Latent Embedding	12
2.3 Order Embedding and Ordinal Embedding	14
2.4 Comparison to Story Understanding	15
2.5 Story Generation	16
2.5.1 Abstract Event Representations	17
2.5.2 Seq-to-seq Attentional Models	18
2.5.3 Story Generation from Images	19

Chapter 3 Visual Storytelling via Global-local Attention Cascading Networks	21
3.1 Introduction	21
3.2 Evaluation for Visual Storytelling	26
3.3 Global-local Attention Cascading Networks (GLAC Net)	27
3.3.1 Encoder: Contextualized Image Vector Extractor	28
3.3.2 Decoder: Story Generator with Attention and Cascading Mechanism	30
3.4 Experimental Results	33
3.4.1 VIST Dataset	33
3.4.2 Experiment Settings	33
3.4.3 Network Training Details	36
3.4.4 Qualitative Analysis	38
3.4.5 Quantitative Analysis	38
3.5 Summary	40
 Chapter 4 Common Space Learning on Cumulative Contexts and the Next Events: Recurrent Event Retrieval Models	 44
4.1 Overview	44
4.2 Problems of Context Accumulation	45
4.3 Recurrent Event Retrieval Models for Next Event Prediction	46
4.4 Experimental Results	49
4.4.1 Preliminaries	51
4.4.2 Story Cloze Test	52
4.4.3 Open-ended Story Generation	53
4.5 Summary	55

Chapter 5	ViStoryNet: Order Embedding of Successive Events and the Networks for Story Regeneration	58
5.1	Introduction	58
5.2	Order Embedding with Triple Learning	60
5.2.1	Embedding Ordered Objects in Sequences	62
5.3	Problems and Contextual Events	62
5.3.1	Problem Definition	62
5.3.2	Contextual Event Vectors from Kids Videos	64
5.4	Architectures for the Story Regeneration Task	67
5.4.1	Two Sentence Generators as Decoders	68
5.4.2	Successive Event Order Embedding (SEOE)	68
5.4.3	Sequence Models of the Event Space	72
5.5	Experimental Results	73
5.5.1	Experimental setup	73
5.5.2	Quantitative Analysis	73
5.5.3	Qualitative Analysis	74
5.6	Summary	77
Chapter 6	Concluding Remarks	80
6.1	Summary of Methods and Contributions	80
6.2	Limitation and Outlook	81
6.3	Suggestions for Future Research	81
초록		101

List of Figures

Figure 1.1	Research vision and a scenario	3
Figure 1.2	Three problems to solve in this dissertation	5
Figure 1.3	Three datasets to be used in this dissertation	5
Figure 1.4	Summary of contributions	6
Figure 1.5	A comparative view of three approaches with respect to latent embedding	7
Figure 3.1	Examples of the task of visual storytelling and image captioning	22
Figure 3.2	Comparative view of sequence-to-sequence models and image-to-text models	23
Figure 3.3	Density estimate plots over automated metric scores with variants of human answers on the image captioning task	25
Figure 3.4	GLAC Net: our proposed model architecture for visual storytelling	29
Figure 3.5	A VIST dataset example: DII and SIS	33
Figure 3.6	Samples of story generation results with visual cues	37

Figure 3.7	The six score results of the human evaluations of the narrative text generation	39
Figure 3.8	The generated story examples for GLAC Net and their ground-truth annotated by humans	41
Figure 3.9	More examples of GLAC Net and their ground-truth annotated by humans.	42
Figure 3.10	Wrongly generated cases for GLAC Net and their ground-truth annotated by humans	43
Figure 4.1	Recurrent Event Retrieval Models (RERMs)	47
Figure 4.2	Component definition of RERMs	50
Figure 4.3	Performance table of RERMs for exploring the networks	52
Figure 4.4	A RERM demonstration of step-wise story generation with one episode in ROCStories dataset	56
Figure 4.5	A RERM demonstration of step-wise story generation with other free selection	57
Figure 5.1	Scenario and Video Preprocessing: building the stream of snapshots of pairs of animated gifs and dialogue texts	60
Figure 5.2	Overall encoding-decoding structures for the story completion tasks	61
Figure 5.3	A proposed architecture of ViStoryNet for Story Learning and Regeneration	66
Figure 5.4	t-SNE Visualization of events	70
Figure 5.5	Visualization of episodes of the trajectory-like form in the embedded event space	71
Figure 5.6	Plots for prediction errors (test data)	74

Figure 5.7	A comparative interpolation example of the 1-step gap with and without SEOE. In case of without SEOE, noisy events are observed.	76
Figure 5.8	An interpolation example of the 5-step gap with SEOE.	77
Figure 5.9	An example of story completion results of a pair sequence of descriptions and dialogues	78
Figure 5.10	Comparative cover map of the generated result and ground-truth	79

List of Tables

Table 3.1	Performance evaluation results with automatic metrics . .	34
Table 3.2	Human evaluation results on the VIST dataset	39
Table 4.1	Performance table of RERMs for Story Cloze Test (SCT)	54
Table 5.1	Decoder performance evaluation results. They are trained overfitted intentionally.	75

Chapter 1

Introduction

1.1 Story of Everyday lives in Videos and Story Understanding

The progress of information technology has rapidly increased the quantity of data. The immense size of video data is uploaded to the internet everyday. A great number of people use internet and social network services through personal computers, and smart devices to record their behavior, knowledge, and experience as life-logs. The logs can include a sequence of situations, actions or dialogues of people that can be told as a story.

On the other hand, recently released are socially interactive household robots such as NAO, Pepper¹, and Jibo². They have video cameras as eyes, and microphones as ears to take visual-linguistic information of their environment including a story as above. In a few years, the robots will live humans together, and they should know common knowledge of everyday lives of humans. Since

¹<https://www.aldebaran.com/en>

²<https://www.jibo.com/>

the knowledge of the family members is personal and episodic, the robots should learn via observation and interaction in the environment (Breazeal, 2004). Ideal datasets for learning the temporal knowledge of family members include both observation and interaction collected on real situated environments, but such data to include contextual stories are not available in public yet. As alternatives, we assume the robots learn by observation only, thus we will not consider tangible information in interaction but focus on visual-linguistic media such as video or some snapshots of images and texts. In this setting, we can utilize video-type datasets including stories for our research.

In short, it is desirable to study on building situation-aware AI agents such as household robots living together with humans. Ideally, learning by experience would be one of good strategies for AI agents due to similarity of humans, but currently, it is limited so far. So, learning by observation (learning by showing) may be one of the alternative strategies, which is advantageous to use video-type materials increasing everyday.³

Let us take one more step to further concretize our research from the above scenario. Temporal knowledge of family members can be called situation to be told by people. Since the term *situation* is not an explicit concept, it is not easy to define and describe. From the philosophy of end-to-end training, we can describe it using natural language, which is human-readable. Then we can adopt the narrative text generation as a surrogate task for situation explanation. Story generation tasks are one of interesting field so-called *Narrative Intelligence (NI)* (Riedl, 2016; Mani, 2013) in AI academia. We can expect to combine the research result from NI in the future.

³License issues are still remained to solve.

retrieval models as solutions. It can be used as '*situation prediction*'.

- The task of story regeneration is to generate the whole story from the partial ones. We demonstrate the methods for embedding stories as trajectory forms and ViStoryNet. It can be used as '*unseen event inference from partially observed situation*'.

Also, we utilize two visual-linguistic dataset and text story dataset as shown Figure 1.2 and 1.3. In particular, visual-linguistic datasets are relatively rare. **PororoQA dataset** (Kim et al., 2017) is one of kids video datasets. Our purpose to use Kids video datasets is to take some advantages (Heo et al., 2010; Ha et al., 2015; Kim et al., 2017): (1) omnibus style, which each episode has simple and explicit storyline in short, (2) narrative order mostly using *fabula*, which follows chronological sequencing of the events, whereas *syuzhet* is a term to designate the way a story is organized to enhance the effect of storytelling (Mani, 2013). (3) relatively small number of main characters and limited spatial environment. This is effective to reduce computational burden and data sparsity. Also, these properties are so desirable to provide as surrogate data similar to that of everyday lives in compact and explicit way.

For the task of visual storytelling, we make use of the **Visual Storytelling dataset (VIST)** (Huang et al., 2016). The **VIST dataset** is the first dataset created vision-to-language of the form of sequence-to-sequence and other story related tasks. The authors want it to be "storyable", thus they deeply use NLP techniques to filter the albums in Flickr data to be what they want. The words in the title are classified as an EVENT using WordNet3.0, the albums are allowed to include with 10 to 50 photos where all album photos are taken within a 48-hour span and CC-licensed. It consists of story-like image sequences paired with: descriptions to form a narrative over an image sequence (images/sentences

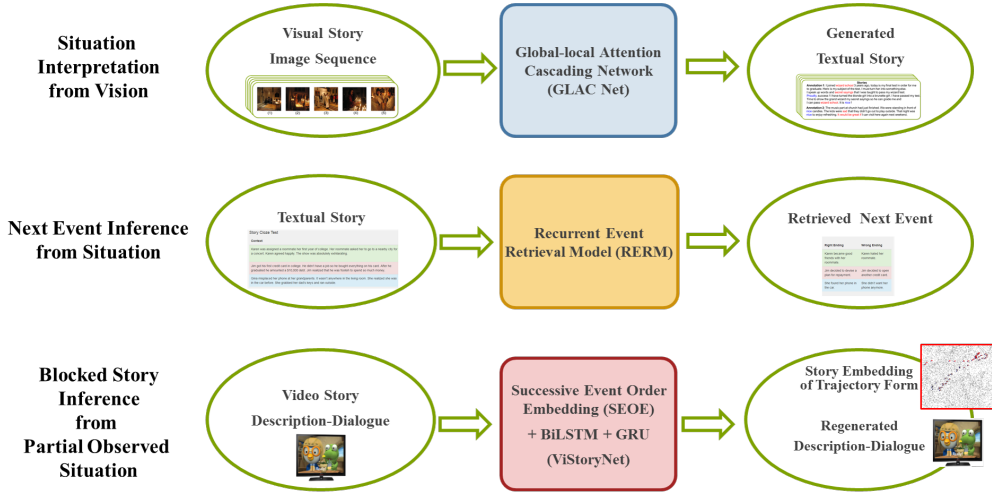


Figure 1.2 Three problems to solve in this dissertation

	# of stories	# of instances	Modality	Length of Stories	Story Type
VIST dataset [Hwang et al., '16]	Middle (~50,000)	Middle (~200,000)	Image sequence and Story (text)	Fixed (5)	Annotated stories on Event-titled Image Collection from Flickr
ROCStories [Mostafazadeh et al., '16]	High (~100,000)	High (~500,000)	Text only	Fixed (5)	Nonfictional daily events with a beginning and ending, where something happens in between
PororoQA Dataset [Kim, Heo, Choi, Zhang, '17]	Low (< 200)	Low (~20,000)	Vision-oriented Text pairs (desc-dialog)	Variable and Long (> 30)	Episodes of Kids' Videos, 'Pororo, the Little Penguin'

Figure 1.3 Three datasets to be used in this dissertation

Task, Models and Methods	Data	Main Results
Global-local Attention Cascading Network (GLAC Net) <ul style="list-style-type: none"> Task of Visual Storytelling Translation image sequences to story sentences in text Effective Information Transfer 	Visual Storytelling (VIST) dataset <ul style="list-style-type: none"> VIST Challenge Pairs of 5 images and story sentences 50K stories with 210K images 	<ul style="list-style-type: none"> Outperform other SOTA methods. 1st place in the challenge with all of 6 aspects of human evaluation
Recurrent Event Retrieval Model (RERM) <ul style="list-style-type: none"> Task of Next Event Prediction Iterative Event Retrieval 	ROCStory dataset <ul style="list-style-type: none"> Binary ending prediction 5-sentence short stories 100K episodes with 500K sentences 	<ul style="list-style-type: none"> Stepwise Story Generation Compatible with human-interactive setting State-of-the-art performance in open-length setting
ViStoryNet <ul style="list-style-type: none"> Task of Episode Regeneration Sequential Event Order Embedding (SEOE) + BiLSTM + GRU 	PororoQA dataset <ul style="list-style-type: none"> Cartoon video series 171 videos (avg. 40 steps) 16K pairs gif / description / dialogues 	<ul style="list-style-type: none"> Story Embedding of Trajectory Form Pointwise Event Interpolation Regenerated description and dialogues for Story Completion

Figure 1.4 Summary of contributions

aligned each). It consists of 50,200 sequences (stories) using 209,651 images (train: 40,155, validation: 4,990, test: 5,055), and the length is 5.

1.3 Approach and Contribution

We propose three methods to learn *latent embeddings* to three problems in Figure 1.1. A latent embedding approach is to find a intermediate latent space and utilize it to solve the given problems. Each problem is solved with different ideas based on latent embedding as shown in Figure 1.5.

For the problem to generate a story in text from a sequence of ordered images, we introduce a GLAC Net (Global-local Attention Cascading Network). It translates from image sequences to story paragraphs in text as a encoder-decoder framework with sequence-to-sequence setting as shown in Figure 1.5 (a). It has convolutional neural networks for extracting information from images

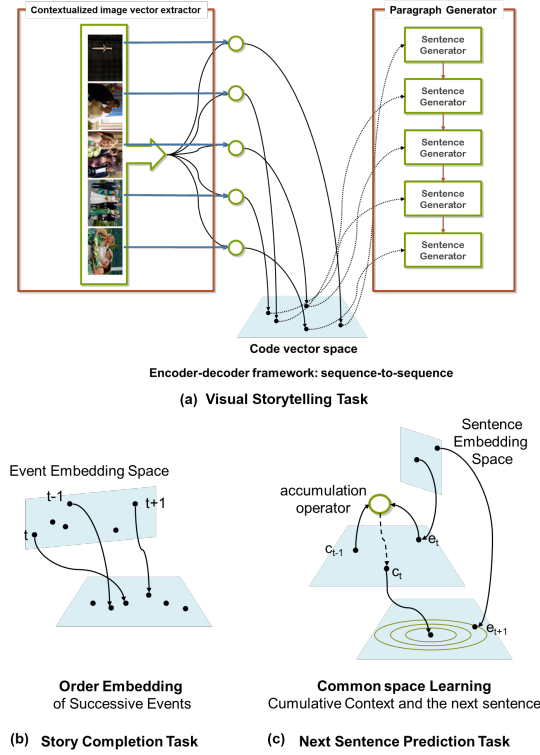


Figure 1.5 A comparative view of three approaches with respect to latent embedding

and recurrent neural networks for text generation. We introduce visual cue encoders with stacked bidirectional LSTMs, and all of the outputs of each layer are aggregated as contextualized image vectors to extract visual clues. The coherency of the generated story is further improved by conveying (cascading) the information of the previous sentence to the next sentence serially in the decoders. We evaluate the performance of it on the Visual storytelling (VIST) dataset. It outperforms other state-of-the-art results and shows the best scores in total score and all of 6 aspects in the visual storytelling challenge with evaluation of human judges.

For the problem to complete partially observed stories, we propose successive event order embedding (SEOE) for composite representation learning for semantics and order for video stories. SEOE embeds each episode as a trajectory form on the latent space, and we propose a ViStoryNet to regenerate video stories (Heo et al., 2018). We convert sentences of events to thought vectors, and train functions to make successive event embed close each other to form episodes as trajectories as shown in Figure 1.5 (b). We test them experimentally with PororoQA dataset, and observe that most of episodes show the form of trajectories. We use them to complete the blocked part of stories, and they show not perfectly but overall similar result.

For the problem to predict the following events or stories with former parts of stories with the constraint to be possible at any step within an arbitrary length, we propose recurrent event retrieval models (RERMs). Figure 1.5 (c) shows the concept of RERMs. They train a context accumulation function and two embedding functions, where make close the distance between the cumulative context at current time and the next probable events on a common latent space. They update the cumulative context with a new event as a input using bilinear operations, and we can find the next event candidates with the updated cumulative context. We evaluate them for Story Cloze Test, they show competitive performance and the best in open-ended generation setting.

Those results above can be applied to AI agents in the living area sensing with their cameras, explain the situation as stories, infer some unobserved parts, and predict the future story.

Additionally, we have done miscellaneous works for video story analysis (Heo et al., 2010, 2015a; Heo and Zhang, 2016; Heo et al., 2016), probabilistic global-local modelling (Heo et al., 2013), and probabilistic learning for human behavior from smartphone lifelogs (Heo et al., 2012, 2015b).

1.4 Organization of Dissertation

The rest of this dissertation is organized as follows:

- Chapter 2 presents a background and a survey of the related work. Firstly, we explain why we study stories. And then, we overview the works on latent embedding and discuss studies on order and ordinal embedding. After that, we show the works of story understanding. Next, we introduce story generation methods in brief.
- Chapter 3 presents 'GLAC Net' to generate a story in text from a sequence of ordered images. We show 6 human evaluation criteria, then explain global-local attention and cascading mechanisms as key elements to solve. We show various example cases and the performance of automatic score metrics and human evaluation.
- Chapter 4 proposes 'Recurrent Event Retrieval Models (RERMs)' to predict the following events or stories with former parts of stories. We explain how to train a context accumulation function and two embedding functions, where make close the distance between the cumulative context at current time and the next probable events on a latent space. Then, it is shown that experimental result for Story Cloze Test and some examples of open-ended story generation.
- Chapter 5 we describe the study on composite representation learning for semantics and order for video stories. We propose 'ViStoryNet' to regenerate (or complete) the whole stories. We explain how to build the models and experimental results.
- In Chapter 6, the dissertation is summarized and discuss the directions for future research.

Chapter 2

Background and Related Work

2.1 Why We Study Stories

When we deliver messages to others, use of visual information, such as images or graphs, is more effective and intense in attracting attention and conveying information than linguistic media such as text and voice only. Similarly, a story is very powerful to deliver what speakers want. Stories make listeners involve thinking, emotion, and imagination altogether, and engage with stories as if their body and mind are in the narrative world. Narratives provide important cognitive frameworks for the transfer and accumulation of experience. Humans have tried to use their experience as well as other people's actively to build up the necessary ability to cope with harsh environment for survival. Jerome Bruner, cognitive psychology and cognitive learning theory in educational psychology, mentioned as follows (BRUNER, 1986):

Narrative (or story) deals in human or human-like intention and action and the vicissitudes and consequences that mark their course.

It strives to put its timeless miracles into the particulars of experience
and to locate the experience in time and place

To us, we interpret that stories are important function to share experience for survival. The sharable, deliverable, and distributable nature of stories affects the building of a socio-culture, i.e., *folk psychology*, and make form part of so-called *commonsense knowledge*. Conversely, to deal with everyday stories as data is closely related to learning commonsense.

Arguably, the ability to engage in stories is the unique feature to make humans humanly. From a talk in Pittsburgh in 1997, the late evolutionary biologist Stephen J. Gould allegedly characterized humans as "the primates who tell stories." Psychologist Robyn Dawes suggested that humans are "the primates whose cognitive capacity shuts down in the absence of a story." (Dawes, 1999) Research suggests that anecdotes can be as persuasive as hard data, and that jurors are influenced by the quality of the prosecution's and defense's stories when deciding whether to find a defendant guilty. Similarly, even in science and engineering, we seek explanations, not mere descriptions; in history, we want a good narrative, not a mere sequence of events.

Stories have been studied in various forms across a range of disciplines, from literary and media studies to psychology and linguistics. In Herman (2013) book 'Storytelling and the Sciences of Mind', narratives (or stories) work as an instrument of mind, and then stories are chunking experience as source of structure. Also, gathering the concept to be spread to other people, they are extended to Folk Psychology.

But to get a handle on their potential role in human intelligence, it needs to consider how they have cropped up in AI. Researchers in AI have explored a potential role for stories since at least the 1990s. In a book 'Tell me a story: A new look at real and artificial memory' by Schank (1990), R. Schank argued for a

crucial link between narrative and intelligence, with narratives guiding learning, structuring memory and supporting generalization. Winston (2011) claimed the Strong Story Hypothesis, according to which "storytelling and understanding have a central role in human intelligence," going on to suggest an artificial system with some human-like capabilities . One reason AI systems might need to understand or produce stories is because they interact with humans. Indeed, there is an evidence that people trust robots more, and can work with them more effectively, when the robots offer more human-like explanations.

2.2 Latent Embedding

From this section, we will introduce some technical related works to ours. Firstly, the common features of our approaches is latent embedding.

Embedding methods are to convert data instances including discrete objects to continuous vectors where certain properties can be represented with distances.¹ The properties can be the distances in the lower-dimensional space (dimensionality reduction), the local distances (manifold learning)², the weights of links in graphs (graph embedding) (Goyal and Ferrara, 2018), the semantics of natural language entities such as words (word embedding) (Camacho-Collados and Pilehvar, 2018) or sentences (sentence embedding) (Perone et al., 2018), or the order of entities (order embedding) (Vendrov et al., 2016).

On the contrary, latent embedding (LE) is a generic approach to find a useful intermediate space to solve the given problems. Traditionally, the most well-known example of LE is the canonical-correlation analysis (CCA). CCA is to find a common latent embedding space via seeking linear combinations

¹Mathematically, embeddings are more abstract concept. We focus on practical use of embedding in this dissertation.

²with the assumption to follow the manifold hypothesis: the data distribution is assumed to concentrate near regions of low dimensionality. (Cayton, 2005; Ma and Fu, 2011)

of two random variables X and Y which have maximum correlation (Hotelling, 1936; Härdle and Simar, 2015). LE has been used in various methodologies and applications since it has some advantageous properties. Firstly, it may alleviate *curse of dimensionality* via embedding the raw instances of high-dimensional features onto the lower-dimensional space. Secondly, it may provide distance-measurable space on which certain information can be encoded on geometric elements such as the position or the relation of inter-instances. Mostly, we can get more efficient representation to highlight useful information retained in the data. We would divide the LE related approaches into two categories from a perspective on which role of the latent space takes: (1) common latent space, and (2) intermediate representation in the encoder-decoder frameworks.

Common space learning uses joint embedding spaces to bridge the gap between heterogeneous sources, e.g., image and label (zero-shot learning (Socher et al., 2013; Akata et al., 2016; Changpinyo et al., 2016; Xian et al., 2016; Zhang and Saligrama, 2016)), image and description (Frome et al., 2013; Kiros et al., 2014), and two sentences in different languages (Johnson et al., 2016b). Also, they can be used to consider various options in one way such as multi-class classification (Amit et al., 2007; Weinberger and Chapelle, 2009) and answer selection in question and answer (Yu et al., 2014; Wang and Nyberg, 2015; Deepak et al., 2017). In this dissertation, we introduce two works to use this approach. The first ones represents cumulative context in the stories in Chapter 4. Also, Chapter 5 shows how to embed episodes to form trajectories, then build neural networks with them for the task of story completion.

Intermediate representation in the encoder-decoder framework forwards to the decoding module from the encoding module. Mostly, the dimension of the *codes*, the output of the encoder, is lower than that of the input. Typically, encoders work as raw data converters to features, decoders generate the output

as solutions. While auto-encoders are typical examples traditionally, practically powerful cases have shown in the deep learning frameworks. The most related configuration to ours is sequence-to-sequence architecture (Sutskever et al., 2014a). It shows impressive results for language translation tasks (Cho et al., 2014; Johnson et al., 2016b) and image captioning tasks (Vinyals et al., 2015). In the subsection 2.5.2, we will review sequence-to-sequence with attention models and works on vision-to-language translation.

2.3 Order Embedding and Ordinal Embedding

Since Mikolov et al. (2013)’s word2vec became popular, various neural embedding methods have been developed from word, sentence, to structured objects such as graphs, trees, and etc. We will focus on order embedding as background of our work, and story generation for our experiments as applications.

Ordinal Embedding and Order Embedding Ordinal embedding is also called non-metric multidimensional scaling consists of finding an embedding of a set of objects based on pairwise distance comparisons (Borg and Groenen, 2005) with pioneering contributions from Shepard and Kruskal. Formally, given a set of ordinal constraints of the form $distance(i, j) < distance(k, l)$ for some quadruples (i, j, k, l) of indices, the goal is to construct a point configuration x^1, \dots, x^n in \mathbb{R}^p that preserves these constraints as well as possible. This problem is relaxed with solving a semi-definite program, generalized non-metric MDS was introduced (Agarwal et al., 2007). For embedding nearest neighbor graphs onto Euclidean space, structure preserving embedding (SPE) was researched based on similar approaches (Shaw and Jebara, 2009). Terada and von Luxburg (Terada and Luxburg, 2014) showed that if a k -nearest neighbor graph is given as local ordinal constraints, we can reconstruct the point set. Also, they showed that statistical consistency is valid. Consistency can be extended from quadruple

learning to triple learning as proven in (Arias-Castro et al., 2017).

In the machine learning community, the work of Jamieson and Nowak (2011) investigates a lower bound for the minimum number of comparison queries of the form “Is object x_k closer to x_i than x_j ?” As ranking problems, an ordinal embedding with pairwise comparison also researched (Jamieson and Nowak, 2011; Ailon, 2012; Wauthier et al., 2013).

In deep learning era, another important work is Vendrov et al. (2016). They defined learning *order embedding* as by learning a mapping which is not distance-preserving but order-preserving. They developed order embedding methods with triplet ranking loss and order violation penalty for hypernym prediction, textual entailment, and image captioning. This approach focus on the partial order structure in the semantic hierarchical relations, and train their embedding functions for partial order completion. From this work, other extended researches are introduced (Li et al., 2017; Wehrmann et al., 2018).

2.4 Comparison to Story Understanding

It is difficulty task to measure how well stories understand, even to humans. The first series of representative tasks for that are to generate description, explanation and story itself, similar to image captioning for image understanding. The second series are to answering to the questions. To quantify the performance, question and answering is relatively easier with the measurable accuracy. On the other hand, the task of story generation suffers from the lack of measuring tools. We will discuss this issue in Chapter 3 again.

Textual Story Understanding As similar works without visual cues, we can categorize into two tasks: question answering and generation. there are text comprehension tasks such as bAbI tasks (Weston et al., 2015), SQuAD (Rajpurkar et al., 2016) and Story Cloze Test (Mostafazadeh et al., 2016). They

have been used for benchmarking new algorithms on document comprehension.

Visual Story Understanding Recently, research fields that combine computer vision and natural language processing, such as image caption generation and visual question answering (VQA), are also growing fast (Antol et al., 2015). The image captioning system generates a natural language sentence describes the scene with the image as an input; VQA system generates an answer to the question by taking the natural language question and the related image. Even though there are a lot of works for VQA as shown in this survey (Wu et al., 2017), relatively less number of works for video QA (YouTube-8M (Abu-El-Haija et al., 2016), MSR-VTT (Xu et al., 2016)) are due to high complexity and the paucity of data. Still, more focused on activity recognition and pose estimation such as Sport-1M (Karpathy et al., 2014), ActivityNet (Caba Heilbron et al., 2015), and Kinetics (Kay et al., 2017; Carreira and Zisserman, 2017). To avoid high complexity, some works focus more specific like TGIF-QA (Jang et al., 2017) (the number of the repeated actions) and Mario-QA (Mun et al., 2017) (limited to specific game environment).

To deal with stories with longer time scale, most of them are used augmented information in text: aligning movies and books (Zhu et al., 2015), movie description (Rohrbach et al., 2015), movie QA on synopsis and script (Tapaswi et al., 2016a), Pororo QA, which built from kid videos with dialogues and descriptions (Kim et al., 2017). Very recently, new dataset is released on TV drama series with QA annotation, TVQA (Lei et al., 2018).

2.5 Story Generation

Most traditional story generation systems have used planning-based approaches (Lebowitz, 1985; Riedl and Young, 2010) or case-based reasoning (Gervás et al., 2005) for entertainments and educations. While they show practically impressive

results for story generation and story authoring applications (Kybartas and Bidarra, 2017; Kapadia et al., 2017), they need a large size of domain knowledge to concretize the story hypotheses: the characters involved, what their goals are, how they interact, how they make effects on the world. Recently, machine learning technologies have been focused for story generation on open domains from available story corpora playing a role of domain knowledge.

Learning Textual Story Generations: While story generation technologies as authoring tools are actively researched area (Dai et al., 2017), we focus on learning-based approaches. Mostly, text-based novels have researched on the focus of plots, e.g., folktales (Finlayson, 2012), suspenses (O’Neill and Riedl, 2014). For open-domain storytelling, some recent works focused on how to construct narrative models automatically: crowd-sourced narrative learning (Li and Riedl, 2015), Swanson and Gordon (2012) built SayAnything system using textual case-based reasoning interactively with human’s response and feedback. McIntyre and Lapata (2009) built story generation systems randomly and ranking with coherence/interest score models trained with SVMs. Recently, Martin et al. (2017) trained sequence-to-sequence models using recurrent neural networks with memory cells to represent event-to-event, and event-to-sentence relationships. And they utilized it to predict the next event and to generate narratives.

2.5.1 Abstract Event Representations

To accomplish story generation, researchers have focused on the narrative chains with NLP-based abstract representation. Chambers and Jurafsky (2008) introduce narrative cloze test and learn causal event chains that revolve around a protagonist. They developed an abstract representation that only care for the verb that occurred and the type of depend-ency that connected the event to

the protagonist. Pichotta and Mooney (2016) expanded it to a 5-tuple event representation of (verb, subject, direct object, prepositional object, preposition). Martin et al. (2017) refined the event representation with modifier term such as (s,v,o,m) where v is a verb, s is the subject of the verb, o is the object of the verb, and m is the modifier working as a wildcard. To pursue generalized semantics, they use wordnet and verbnet. And story cloze test (SCT) (Mostafazadeh et al., 2016) we will use mainly is the task of choosing the right ending between two given sentences. SCT converted story generation into a binary classification problem. 4-sentence story is given with two possible endings as the 5-th sentence.

2.5.2 Seq-to-seq Attentional Models

Sequence-to-sequence models (Sutskever et al., 2014b) were introduced for machine translation task based on source-target paired textual corpora. To improve performance of the target text generation, attentional models provide connections of all hidden information of encoders of sequence models (Bahdanau et al., 2014; Luong et al., 2015) as Figure 3.2. Recently, self-attention (or intra-attention) (Lin et al., 2017) and key-value attention (Vaswani et al., 2017) introduced intra-connections on source or target language. Especially, global-local attentional models (Luong et al., 2015) considered overall information of encoded source (global) and that of focused position (local) together. It is conceptually similar, but our proposed method are different in design level of submodules, connection to submodules each other caused from the focus on multi-modal setting.

Images to Single-text: Since AlexNet (Krizhevsky et al., 2012) as a milestone, object recognition and detection methods have grown explosively and outperformed human ability of capturing objects in accuracy aspect (Geirhos et al., 2017). While modeling the inter-relation between source-target texts for

machine translation tasks, attention models for computer vision pursue modeling salient areas in complex visual inputs such as Scene/video description tasks (Xu et al., 2015; Johnson et al., 2016a; Karpathy and Fei-Fei, 2017; Donahue et al., 2015) and image/video question answering tasks about the stories (Tapaswi et al., 2016b; Kim et al., 2017, 2018).

2.5.3 Story Generation from Images

Huang et al. (2016) introduce the VIST dataset consisting of sequential images and natural language sentences, and discussed how this data could be used for visual storytelling tasks. They show the result of basic sequence-to-sequence models as baselines.

Retrieval-based approaches The first work for multiple-frame to multi-sentence modeling is done by Park and Kim (2015). They use a coherence model in textual domain for resolving the entity transition patterns between sentences. However, they define the coherence as rigid word reappearance frequency, which is unable to address the semantic gap and therefore cannot fully express the deeply meaning. Moreover, they focus on textual coherence without acknowledging the problem of large visual variance. Liu et al. (2017) developed semantic embedding of the image features on the bi-directional recurrent architecture to generate a relevant story to the pictures. There are similar points to ours since it used bi-directional recurrent architecture for embedding image sequence context with the VIST dataset.

Adversarial training for generation Adversarial training strategies in reinforcement learning framework were tested on this task (Wang et al., 2018b,a). They have generator networks and reward models separately, the result of generators provides examples to calculate the reward with the reward models such as discriminators (Wang et al., 2018a) or regressors (Wang et al., 2018b).

The rewards give feedbacks to generators to generate more realistic fake examples. In the work (Wang et al., 2018a), the first discriminator checks whether an image and the corresponding sentence is well-matched, generated, or shuffled. The second one classifies whether a paragraph story comes from the dataset, it is generated one, or shuffled sentence in order. It has shown the state-of-the-art performance with respect to automated metric scores.

Chapter 3

Visual Storytelling via Global-local Attention Cascading Networks

3.1 Introduction

In some years, deep learning have brought about breakthroughs in processing image, video, speech and audio. The field of natural language processing (NLP) has been also interested in deep learning, e.g., sentence classification (Kim, 2014), language modeling (Bengio et al., 2003; Mikolov et al., 2013), machine translation (Sutskever et al., 2014b; Bahdanau et al., 2014; Vaswani et al., 2017), and question answering (Hermann et al., 2015).

Naturally, bridging images and texts by deep learning has been following (Belz et al., 2018) such as image captioning (Vinyals et al., 2015; Xu et al., 2015; Karpathy and Fei-Fei, 2017), visual question answering (Antol et al., 2015; Kim et al., 2016), and image generation from caption (Reed et al., 2016; Zhang et al., 2017).

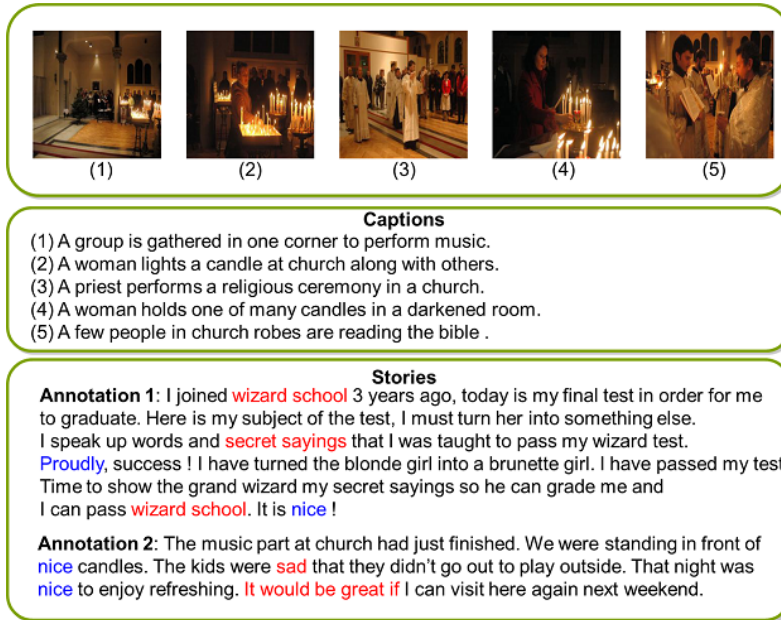


Figure 3.1 Examples of the task of visual storytelling and image captioning. Both of captions and stories are annotated by human workers. Two stories are very different. Blue emotional words and red clauses from story text are more subjective than captions. Best viewed in color.

Narratives (or stories) are fundamental parts of human intelligence as well as social intelligence (Herman, 2013; Winston, 2011; Chomsky, 2010). They serve as vehicles to share experience, information and intentions via languages. With the perspective, to generate a narrative paragraph of multiple coherent sentences from an ordered photo stream is an interesting and fundamental challenge on both computer vision and natural language processing (NLP). This task is called as 'Visual Storytelling' (Huang et al., 2016). This is challenging because of the difficulties such as detecting the visual clues spread on photo streams, understanding contexts or situations, constructing narrative structures, and generating the paragraph written in an expressive way for storytelling.

So far, most of researches have much focused on visual captioning (Vinyals

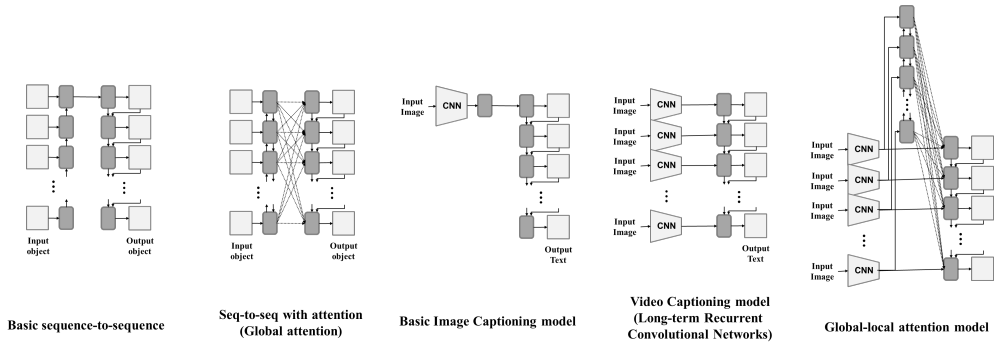


Figure 3.2 Comparative view of sequence-to-sequence models and image-to-text models. For text-to-text translation, seq-to-seq models with attention were introduced (Bahdanau et al., 2014; Luong et al., 2015). In multi-modal cases, it is important to extract features of salient parts, and deliver them to appropriate position. The rightmost figure shows the global-local configuration for our problem setting to use context information with attention and direct relation from input objects to output objects.

et al., 2015; Xu et al., 2015; Karpathy and Fei-Fei, 2017; Donahue et al., 2015; Belz et al., 2018) to show impressive results, which aim at describing the content of an image or a video in an objective expression style. Still, their capability of story generation is restrictive.

In this paper, we further investigate the capabilities in understanding more visual scenarios, composing more structured expressions, and creating better narrative paragraphs from image sequences. The main challenges of this task are as follows; At first, different from single image captioning, we should generate multiple coherent sentences to be focused on one theme, well-structured, grammatical and well-organized. All the while, they should be image-specific sentences within the context of overall images while those properties are maintained. Secondly, stories are more diverse than descriptions. If visual cues are appropriately relevant on the story text, humans may accept totally different stories as shown in Fig 3.1. This observation is valid on the VIST dataset

(Huang et al., 2016), too. Technically, it causes the severe problem of automatic evaluation using popular metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), and CIDEr (Vedantam et al., 2015), which are all compared to golden answers. As the more diverse stories are acceptable, the probabilities to match with golden answers get lower. As reported in (Kilickaya et al., 2017), the correlation coefficients to human evaluation are not so good: most correlated one is METEOR with the value ~ 0.44 on the composite dataset¹ For the VIST dataset, (Huang et al., 2016) looked for the best correlated metric: METEOR with $\rho=0.22$. Notably, one of recent works (Wang et al., 2018b) using reinforcement learning (RL) with metric scores as rewards showcases an adversarial example with average METEOR score as high as 40.2:

We had a great time to have a lot of the. They were to be a of the.
 They were to be in the. The and it were to be the. The, and it were
 to be the.

Conversely, they report to observe many relevant and coherent stories with low scores (nearly zero). To avoid this problem, we take a detour to utilize mainly human evaluation to measure their performances with criteria proposed in the Visual Storytelling challenge (Huang et al., 2018b).

To deal with the difficulties, we propose Global-local Attention Cascading Network (GLAC Net) that a sequence-to-sequence model with combination of global-local attention and context cascading mechanism. The model incorporates two simple attention: a *global* level to process overall encoding of context/narrative structure to text generator; a *local* one that chooses a certain image from a sequence of visual cues. Specifically, we introduce the visual cue

¹a mixture of Flickr8k, Flickr30k and MS-COCO

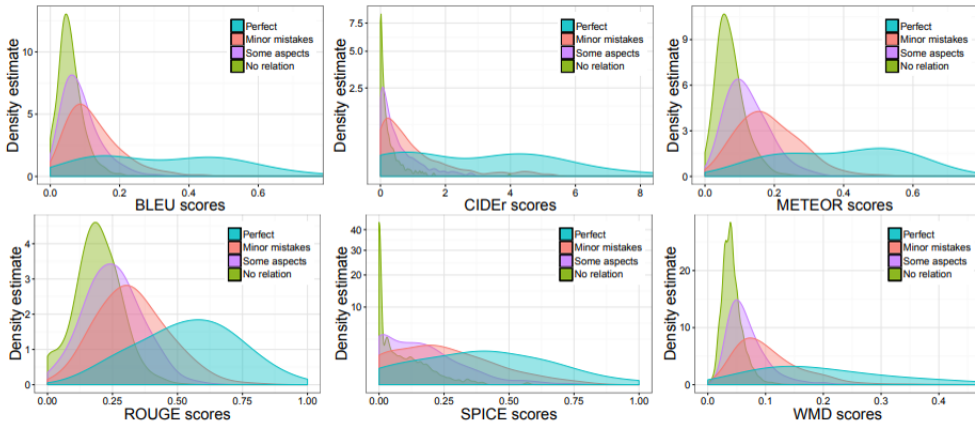


Figure 3.3 Density estimate plots over automated metric scores with variants of human answers on the image captioning task (Kilickaya et al., 2017) based on Flickr8k, Flickr30k and MS-COCO dataset. While METEOR shows the best correlation, the average value is ~ 0.44 .

encoder with stacked bi-directional LSTMs, and all of outputs of each layers are aggregated as contextualized image vectors. It can be interpreted as giving a multi-level representation, i.e., overall context encoding level (global) and image feature level (local). We give *local* attention on image features directly (Figure 3.2). Then, both of them are combined and sent to RNN-based sentence generators.

The next point is context cascading mechanisms. To improve the coherency of generated stories, we design models to convey the last hidden vector in the sentence generator to the next sentence generator as initial hidden vector while keeping the concept of global-local attention model.

Recently, the VIST dataset was released for the task of visual storytelling, which is composed of five-sentence stories, descriptions and the corresponding sequences of five images (Huang et al., 2016). We test our methods with them and outperform the state-of-the-art (SOTA) works with all 6 aspects of human evaluation criteria proposed by (Huang et al., 2018a), which was proved from

the 1st visual storytelling challenge ² by human judges.

3.2 Evaluation for Visual Storytelling

Before we introduce how to deal with the task of visual storytelling, which is to generate a story of sentence sequence from a given sequence of images, we consider evaluation criteria to establish our goals of modeling. Due to the emergence of inter-relationship of sentences in the generated stories, we need appropriate measures what good stories are. While typical automatic evaluation metrics for sequence generation tasks such as BLEU, METEOR, ROUGE, CIDEr, SPICE, and WMD are based on similarity to golden answers, they are not appropriate as objective functions (Wang et al., 2018b), and the correlations to human evaluation are not so high (Kilickaya et al., 2017). Different from single-sentence generation tasks, we should consider the properties of the generated sentences to be focused on one theme, well-structured, grammatical and well-organized. Recently, one of interesting criteria were introduced in the visual storytelling challenge (Huang et al., 2018b,a) as follows:

1. **"The story is focused"**: Each sentence of the story is relevant to the rest of the story?
2. **"The story is coherent"**: The story is well-structured, grammatical and well-organized?
3. **"I would share"**: If they were users' photos, the users have a will to share their experience with their friends?
4. **"Written by a human"**: The story sounds like it was written by a human?

²Here is the page for visual storytelling challenge: <http://www.visionandlanguage.net/workshop2018/>

5. **"Visually grounded"**: The story directly reflects concrete entities in the photos?

6. **"Detailed"**: The story provides an appropriate level of detail?

Even though they need human labors for evaluation, it is advantageous to cover several aspects such as overall properties (1,2), degree of satisfaction (3), human likeness (4), and image-specificity (5,6). Those provide a good guideline what kinds of points should be considered into models.

3.3 Global-local Attention Cascading Networks (GLAC Net)

We formulate the task of visual storytelling as a sequence-to-sequence learning problem, which the input is a sequence of images and the output is a sequence of sentences including the corresponding stories. Briefly, our methods are composed of two stages: (1) representation learning of image sequences as encoders, and (2) textual story generators as decoders. We are given a sequence of images $V = \{v_1, v_2, \dots, v_T\}$ and the corresponding sequence of sentences $S = \{s_1, s_2, \dots, s_T\}$. Note that the length of V and S is the same value T . Each $s_i = \{w_1, w_2, \dots, w_{T'}\}$ in S is a sequence of words, which is not limited rigorously in only one sentence, it can have one or two sentences. The length T' in s_i is not fixed depending on the sequence. To indicate the starting point and the end point of s_i , we add $\langle \text{START} \rangle$ and $\langle \text{END} \rangle$ symbols as special words in the word vocabulary.

Similar to sequence-to-sequence model setting (Sutskever et al., 2014b), we define the objective of training as to estimate the conditional probability $p(S|V)$ with LSTM language model to decode textual stories.

$$\begin{aligned} p(S|V) &= p(s_1, s_2, \dots, s_T | v_1, v_2, \dots, v_T) \\ &= \prod_{t=1}^T p(s_t | s_1, \dots, s_{t-1}, v_1, \dots, v_T) \end{aligned} \tag{3.1}$$

Note that the formulation becomes the same to the image captioning framework (Vinyals et al., 2015) if $T = 1$.

Different from the text, the variance of values in V representing at the pixel-level is very high even though the context of V is the same. To get better estimated probabilities, we can encode images using pretrained Convolutional Neural Networks (CNNs) such as VGGNet (Simonyan and Zisserman, 2015) or ResNet (He et al., 2015), which have been used for various computer vision tasks, and are currently state-of-the-art for object recognition and detection. These features represent single images as real-valued vectors with smaller number of dimension than the one of image pixels.

$$X_{\text{CNN}} = \{x_1, x_2, \dots, x_T\} = \text{CNN}(V) \quad (3.2)$$

Then, we can rewrite more effective formula $p(S|X_{\text{CNN}})$ instead of $p(S|V)$. X_{CNN} is the features of each image separately.

3.3.1 Encoder: Contextualized Image Vector Extractor

Similar to (Bahdanau et al., 2014), we use bi-directional LSTMs (BiLSTM) as main components of context encoders. BiLSTM consists of forward and backward LSTM's. The forward LSTM \vec{f} reads the input sequence as it is ordered, and computes a sequence of forward hidden states $(\vec{h}_1, \dots, \vec{h}_T)$. The backward LSTM \overleftarrow{f} reads the input sequence in the reverse order, and generates a sequence of backward hidden states $(\overleftarrow{h}_1, \dots, \overleftarrow{h}_T)$. Typically, the layer output h_t at time t is the concatenation of two directional outputs $[\vec{h}_t; \overleftarrow{h}_t]$.

Recently, context word embeddings, e.g., CoVe (McCann et al., 2017) and ELMo (Peters et al., 2018) were developed to provide transferrable pre-trained encoder for a variety of NLP tasks similar to CNNs trained on ImageNet for computer vision. They embed words as additional real vectors to include the

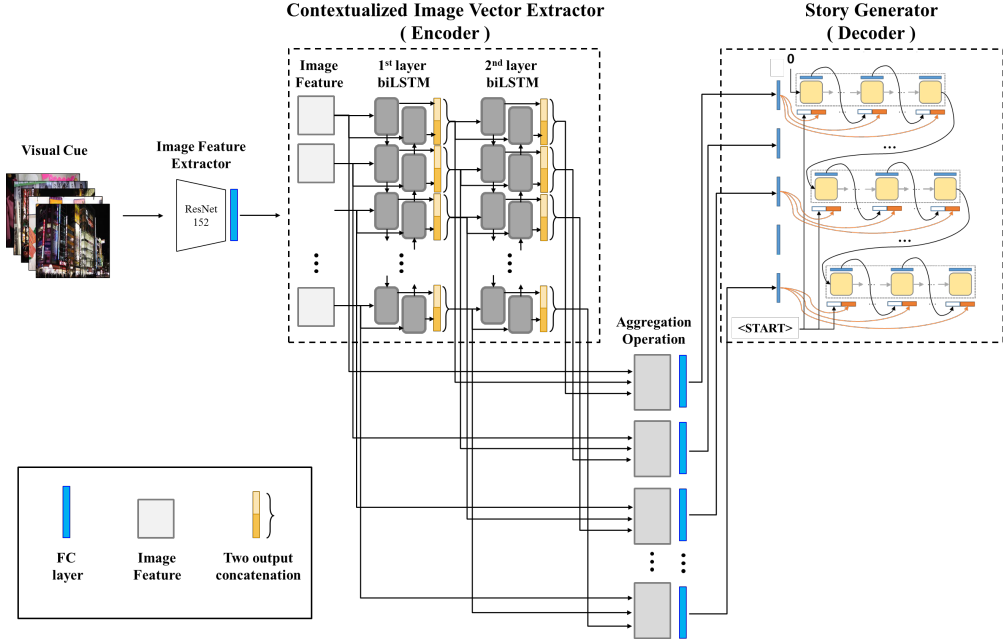


Figure 3.4 Our proposed model architecture for visual storytelling. Basic encoder-decoder structure. Note that activation function (ReLU), dropout, batch normalization, and softmax layer are omitted for readability. Best view in color.

context information within the sentence using the outputs of each layers of stacked BiLSTM. Inspired from the contextualized word embeddings, we design an encoder module of neural networks to convert sequences of image vectors onto contextualized ones within the given image sequence. Specifically, we introduce the visual cue encoder with stacked BiLSTM with residual connections, and all of each layer’s output are aggregated as contextualized image vectors for one image in the visual cue as shown Figure 3.4. We check the quality of generated result depending on the number of layers, stacked models show slightly better than single-layer one.

For each image feature x_t , a L -layer BiLSTMs computes a set of $2L$ representations: $h_{t,j} = [\overrightarrow{h_{t,j}}; \overleftarrow{h_{t,j}}]$ is the concatenated output for each BiLSTM layer

where j is the layer index ($j = 1, \dots, L$).

For the aggregation functions of all output of BiLSTM layers, we consider concatenation of all layers (all), concatenation of the input and the last layers (tb), element-wise summation (ews), element-wise product (ewp), and mixed operations.

$$\begin{aligned}
f_t^{all} &= [x_t; h_{t,1}; \dots; h_{t,L}] \\
f_t^{tb} &= [x_t; h_{t,L}] \\
f_t^{ews} &= x_t + h_{t,1} + \dots + h_{t,L} \\
f_t^{ewp} &= x_t \odot h_{t,1} \odot \dots \odot h_{t,L}
\end{aligned} \tag{3.3}$$

Note that the dimension of all layer concatenate aggregation is increasing depending on L , and for the operations of ews or ewp, it needs to match the dimensions of x_t and $h_{t,j}$.

3.3.2 Decoder: Story Generator with Attention and Cascading Mechanism

Importance of visual storytelling is how to generate image-specific sentences within the context of overall images. To achieve the above goal, we design our models to use both of context information from the output of encoders and raw image features together.

In the typical global attention (Bahdanau et al., 2014; Luong et al., 2015) in Figure 3.2, they can refer all hidden information. We give hard constraints on global attention and local attention since our input object is not a word but an image.

Let us define sentence generator G , which can generate one or two sentences with arbitrary length.

$$(S_i, l_i) = G(o_{1:T,L}, l_{i-1}) \tag{3.4}$$

where $o_{1:T,L}$ is the $d \times T$ matrix of output of the encoder (if d is the value of dimension), l_i is the final hidden state of generator G of i_{th} iteration and needs to define l_0 vector before use. G can be implemented with arbitrary RNNs, we use LSTMs. S_i is the i_{th} generated result.

From our experiments, we observe that the constraint to give better performance (Figure 3.7) as following:

$$(S_i, l_i) = G(o_{i,L}, l_{i-1}) \quad (3.5)$$

From the result, different from standard attention models, we do not need to use activation functions to induce probabilistic distributions such as softmax.³ Also, the GLAC Net implements them in a very simple way via hard connections from the aggregated outputs of encoders or each image feature onto each corresponding sentence generator while standard attentional configuration may need a large number of parameters. The coherency of the generated story is further improved by conveying (cascading) the information of the previous sentence to the next sentence serially in the decoders.

The outputs of encoders include overall information of the sequence (global). On the other hand, the image-specific features are constrained only on the image (local). The aggregated vector of them (global+local vector) is obtained from the global-local attention containing the story flow and the information of each image. They can be represented as 'hard' attention each on the specific inputs or the encoding vectors.

Cascading mechanism

In GLAC Net, we need to design our decoders to generate several sentences sequentially. To implement this, we introduce the cascading mechanism to use

³That is because it is equivalent to choose one among one candidate.

hidden states as conveyer channels. It needs to initialize the hidden values of the first sentence generator as zeros. It is different from the standard image-sentence connection setting shown in image captioning papers (Vinyals et al., 2015; Karpathy and Fei-Fei, 2017; Xu et al., 2015), which the outputs of CNN encoders are connected to the initial hidden states of sentence generators. Instead, we connect the outputs of encoders to the every input layers of sentence generators in all steps as Figure 3.4. Then, the hidden values in the final step are used as the initial hidden values of the next sentence generator. When we need to remove the mechanism for ablation study, we disconnect the cascading information flow from the previous sentence generator to the next one.

Following the standard sequence-to-sequence problem setting, we use each s_i is produced one word at a time. We use cross-entropy loss over the training data.

Avoiding Duplicates

As simple heuristics to avoid duplicates in the decoders, we sample words one hundred times from the word probability distribution of the LSTM output, and choose the most frequent word from the sampled pool. This reduces the number of repetitive expressions and improve the diversity of the generated sentences. On the process of generating sentences of the story, We also count the selected words. The selection probabilities of the words are decreased according to the frequency of each word as Equation 3.6, and normalized.

$$\hat{p}(word) = p(word) \times \frac{1}{1 + k \cdot count_{word}} \quad (3.6)$$

where k is a constant for sensitivity. We use $k=5$.

To build grammatically correct sentences, the probabilities of some function words such as prepositions and pronouns are not changed regardless of the frequency of occurrence.




					
DII	A black frisbee is sitting on top of a roof.	A man playing soccer outside of a white house with a red door.	The boy is throwing a soccer ball by the red door.	A soccer ball is over a roof by a frisbee in a rain gutter.	Two balls and a frisbee are on top of a roof.
SIS	A discus got stuck up on the roof.	Why not try getting it down with a soccer ball?	Up the soccer ball goes.	It didn't work so we tried a volley ball.	Now the discus, soccer ball, and volleyball are all stuck on the roof.

Figure 3.5 A VIST dataset example. DII: Descriptions of images in isolation. SIS: Stories of images in sequence.

3.4 Experimental Results

3.4.1 VIST Dataset

VIST dataset is the first dataset particularly created for sequential vision-to-language and other story related tasks (Huang et al., 2016). It consists of story-like image sequences paired with: (1) descriptions for each image in isolation (DII) ($\sim 80\%$ only), (2) descriptions to form a narrative over an image sequence (images/sentences aligned each) (SIS) as shown in Figure 3.5. It consists of 50,200 sequences (stories) using 209,651 images (train: 40,155, validation: 4,990, test: 5,055), and the length is 5.

3.4.2 Experiment Settings

We follow the split of VIST dataset and adopted both the automatic metrics (BLEU, METEOR and ROUGE-L) and the six human evaluation criteria. Every scores are evaluated on the test dataset. We utilize the open source evaluation

Configuration	Validation Perplexity	Test Perplexity	B-1	B-2	B-3	B-4	ROUGE-L	METEOR
Seq2Seq (Beam=10)	-	-	-	-	-	-	-	0.231
Seq2Seq (Greedy)	-	-	-	-	-	-	-	0.278
Seq2Seq (-Dups)	-	-	-	-	-	-	-	0.301
Seq2Seq (+Grounded)	-	-	-	-	-	-	-	0.314
Sentence-Concat	30.58	31.02	0.411	0.198	0.096	0.050	0.257	0.301
Story-Flat	28.35	28.32	0.271	0.139	0.070	0.037	0.204	0.232
CRCN	-	-	0.367	0.132	0.042	0.016	0.195	0.231
SRT	-	-	0.409	0.202	0.099	0.052	0.259	0.303
Ours (-Cascading)	20.24	20.54	0.440	0.219	0.104	0.053	0.259	0.301
Ours (-Global)	18.32	18.47	0.369	0.173	0.082	0.041	0.250	0.291
Ours (-Local)	18.21	18.33	0.373	0.181	0.091	0.049	0.251	0.294
Ours (-Count)	18.13	18.28	0.334	0.156	0.077	0.041	0.245	0.276
Ours	18.13	18.28	0.406	0.194	0.091	0.045	0.246	0.296
Ours (4 stacked BiLSTM+all)	18.27	18.32	0.385	0.191	0.097	0.052	0.255	0.301
Ours (4 stacked BiLSTM+sum)	18.33	18.35	0.370	0.183	0.092	0.049	0.252	0.300
Ours (4 stacked BiLSTM+product)	20.14	20.07	0.325	0.139	0.066	0.034	0.232	0.282
Ours (3 stacked BiLSTM+all)	18.31	18.36	0.378	0.188	0.096	0.052	0.251	0.302
Ours (3 stacked BiLSTM+sum)	18.28	18.37	0.374	0.186	0.095	0.051	0.252	0.302
Ours (3 stacked BiLSTM+product)	18.43	18.51	0.379	0.187	0.094	0.05	0.256	0.299
Ours (2 stacked BiLSTM+all)	18.29	18.34	0.376	0.187	0.096	0.053	0.255	0.303
Ours (2 stacked BiLSTM+sum)	18.28	18.37	0.374	0.186	0.095	0.051	0.252	0.301
Ours (2 stacked BiLSTM+product)	18.37	18.43	0.372	0.182	0.092	0.049	0.253	0.300
Ours (1 stacked BiLSTM+all)	18.30	18.37	0.366	0.182	0.093	0.050	0.250	0.298
Ours (1 stacked BiLSTM+sum)	18.27	18.34	0.379	0.187	0.095	0.051	0.255	0.299
Ours (1 stacked BiLSTM+product)	18.25	18.34	0.372	0.182	0.049	0.253	0.253	0.302

Table 3.1 Performance evaluation results with automatic metrics. Baselines are reported in (Huang et al., 2016). B-1~4 designate BLEU-1~4. Compared with the performance of baselines (Huang et al., 2016), the GLAC Net is very competitive without beam search methods. From the results of 'GLAC Net (-Count)' and 'Baselines (-Dups)', the heuristics are helpful to reduce redundant sentences and improve the scores. Compared to LSTM Seq2Seq models, GLAC Net-based model shows better performance in general. While GLAC Net (-Cascading) looks like the best, the human evaluation demonstrates that the GLAC Net shows the best in total score and their 4 aspects out of 6 ones.

code⁴ used in (Yu et al., 2017).

The generated result for the challenge and their demo systems with your images can be accessed at <http://glac.droppages.com/>.

Compared methods for both of automatic and human evaluation are as follows:

- Story-Concat (Vinyals et al., 2015): Concatenation of popular image captioning models with CNN-RNN framework to generate captions for single images.
- Story-Flat (Huang et al., 2016): Basic sequence-to-sequence model as a translation task. two unidirectional GRUs are used for image sequence encoding and sentence generation each.
- CRCN (Park and Kim, 2015): CRCN combines CNN, RNN and an entity-based local coherence model to learn the semantic relations from streams of images and texts. It is a retrieval-based approach to be less performed in case of the large number of instances.
- SRT (Wang et al., 2018a): Generative adversarial training on the CNN, RNN and two discriminators to generate adversarial signals as rewards. The first discriminator checks whether an image and the corresponding sentence is well-matched, generated, or shuffled. The second one classifies whether a paragraph story comes from the dataset, it is generated one, or shuffled sentence in order. It has shown the state-of-the-art performance with respect to automated metric scores.

⁴https://github.com/lichengunc/vist_eval. For the Visual storytelling challenge, the official evaluation code to calculate METEOR is offered: <https://github.com/windx0303/VIST-Challenge-NAACL-2018>. This tool scores more higher around 0.001 due to consideration of sets of golden answers.

As ablation study, we evaluate the effects of the GLAC Net, we perform ablation experiments as shown in Table 3.1 (automatic metric), Table 3.2 and Figure 3.7 (human evaluation). We consider various models: simple LSTM Seq2Seq network, GLAC Net without context cascading, GLAC Net without global information, GLAC Net without local information, GLAC Net without heuristics of duplicate avoidance, and complete GLAC Net.

For human evaluation, we recruit 316 human judges on Amazon Mechanical Turks. Workers were asked to rate 200 randomly selected stories⁵ on the six aspects, using a 5-point Likert scale from 'Strongly Disagree (1)', 'Disagree (2)', 'Neither Disagree nor Agree (3)', 'Agree (4)', to 'Strongly Agree (5)'. We take 3.51 evaluations per story averagely after removing wrong submissions.

3.4.3 Network Training Details

The training image data is resized to 256×256 in advance. At the training stage, each image is augmented by 224×224 random cropping and horizontal flip process, and the value of each pixel is normalized to $[0,1]$. All parameters are trained with the Adam optimizer. The learning rate and weight decay values are 0.001 and $1e-5$, respectively. Each word is embedded into a vector of 256 dimensions, and the LSTM is trained using teacher forcing. We also use batch normalization and dropout techniques to prevent overfitting and improve performance in training. We use 64 batch size and the training data reshuffled at every epoch.⁶

⁵Strictly, we used 200 stories in the opened sample pages in <https://github.com/windx0303/VIST-Challenge-NAACL-2018>. We asked organizers how to pick them, we got the answer 'randomly'.

⁶As the used hyperparameters of GLAC Net, input dimension of image features is 1024, hidden dimensions are 1024, aggregation function is 'tb' and the number of layers of stacked BiLSTM is 2.

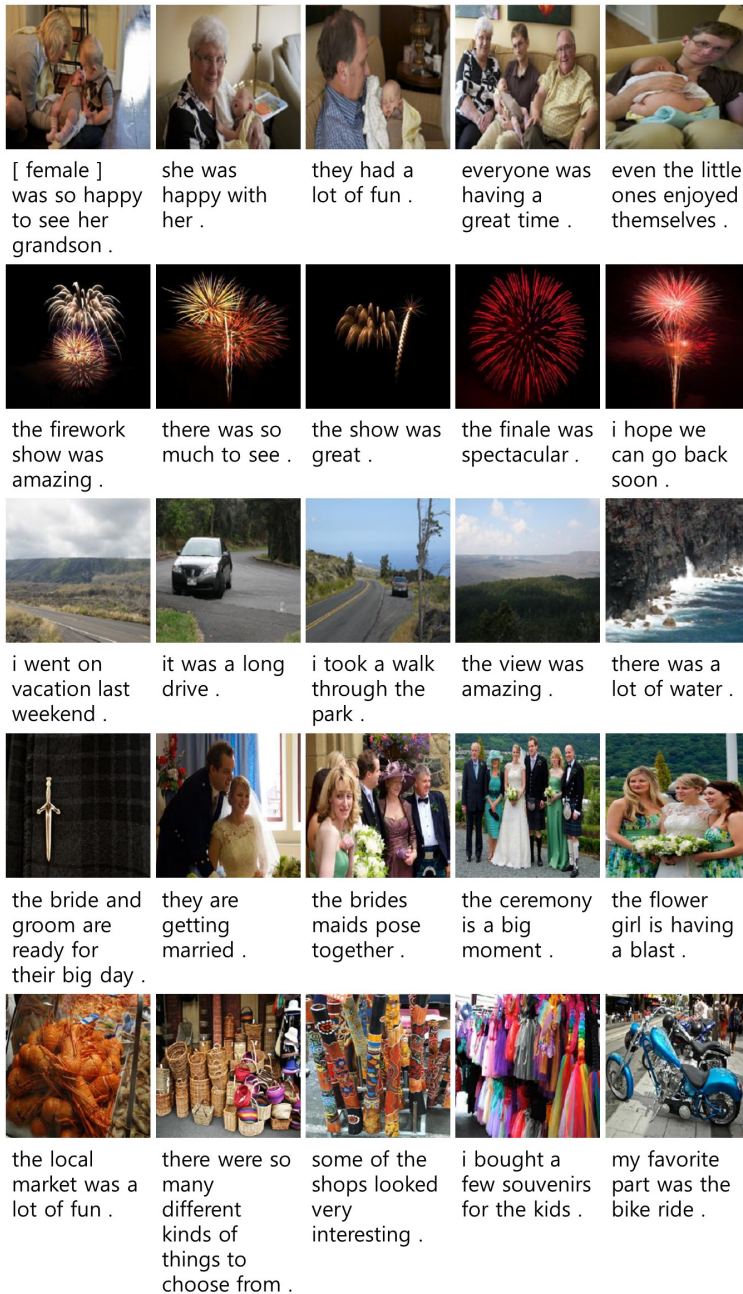


Figure 3.6 Samples of story generation results with visual cues

3.4.4 Qualitative Analysis

As shown in Figure 3.6, 3.8, 3.9 and 3.10, the exemplar generated stories with GLAC Net from test dataset are presented. The context of successive images is well reflected, and the content of each image is properly described.

In the first story in Figure 3.6, every picture shows very similar scenes that person is hugging a baby. The generated stories shows that our models are good at gender/age recognition, pronoun usage. In the second story, the words in the generated sentences and objects from the images are well-matched and visually-grounded: car and drive (2nd image), sea and water (5th image). In the third story, there is no clue that the story will be related to wedding. Catching overall context from visual cues, the sentence in the first part can be generated properly. Generally, all stories in Figure 3.6 show the structure of stories. Especially, the second one in Figure 3.6 is generated with only firework images.

Figure S1 shows that the generated stories depending on decoding methods (showing usefulness of proposed heuristics) and training epoch with greedy generation (showing stability of our method). Figure S2 presents the generated stories with diverse ablation settings.

Figure S3 shows more generated cases of good, acceptable and wrong quality. In the wrong cases, overall structure is still maintained.

3.4.5 Quantitative Analysis

The automatic evaluation results are shown in Table 3.1. Compared with the performance of baselines (Huang et al., 2016), the GLAC Net is very competitive without beam search methods. From the results of 'GLAC Net (-Count)' and 'Baselines (-Dups)' in Table 3.1, the heuristics are helpful to reduce redundant sentences and to improve the scores. Compared to GLAC Net-based model

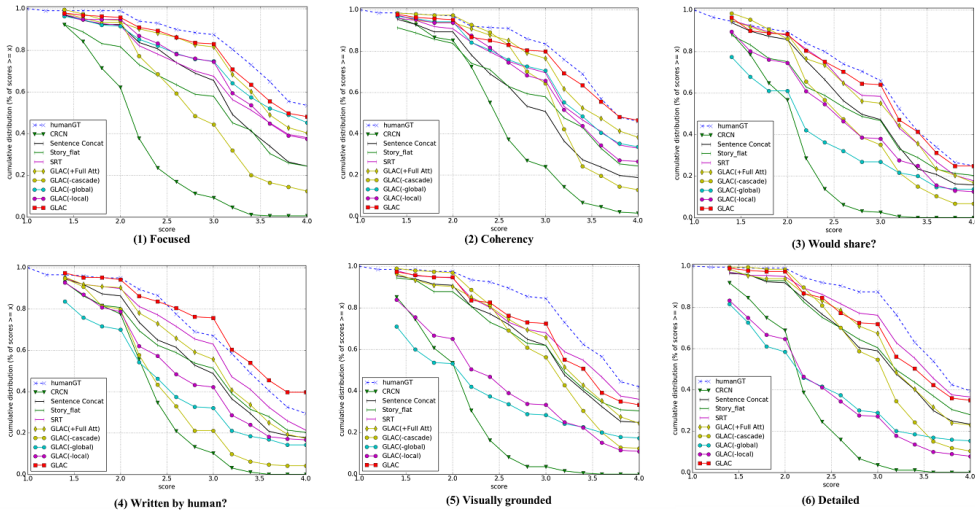


Figure 3.7 The six score results of the human evaluations of the narrative text generation by GLAC Net, ablation models and compared methods. GLAC Net shows the best scores in 1~4 criteria. Though SRT still shows the best in 5~6 ones, GLAC Net is very competitive. Best view in color.

Model	Focused	Coherent	I would Share	Written by human	Visually-grounded	Detailed	Total score
GLAC Net (ours)	3.548	3.524	3.075	3.589	3.236	3.323	20.295
DG-DLMX	3.347	3.278	2.871	3.222	2.886	2.893	18.498
NLP5A501	3.111	2.870	2.769	2.870	3.072	2.881	17.574
AREL	3.236	3.065	2.767	3.029	3.032	2.867	17.995
Human (Public Test set)	4.025	3.975	3.772	4.003	3.965	3.857	23.596
GLAC Net (ours)	3.588	3.547	3.061	3.382	3.282	3.306	20.167
GLAC Net (-Cascading)	2.803	3.047	2.547	2.340	2.952	2.918	16.606
GLAC Net (-Global)	3.405	3.198	2.191	2.341	2.162	2.254	15.551
GLAC Net (-Local)	3.359	3.186	2.600	2.671	2.466	2.344	16.627
Sentence Concat	2.955	2.988	2.692	2.816	3.101	3.127	17.680
Story-Flat	3.118	2.888	2.746	2.826	3.056	3.060	17.693
CRCN	2.092	2.502	1.970	2.239	1.952	2.106	12.861
SRT	3.322	3.257	3.019	3.122	3.340	3.411	19.472

Table 3.2 Human evaluation results on the VIST dataset. The upper part shows the announced results of 4 teams of the 1st Visual Storytelling Challenge (Huang et al., 2018a). Our model outperforms all other teams on all of 6 aspects. AREL means Adversarial REward Learning. The lower part presents the results of ablation study with GLAC Nets and those of other previous methods (Vinyals et al., 2015; Huang et al., 2016; Park and Kim, 2015; Wang et al., 2018a) or their variants. GLAC Net shows the best performance in total score as well as 4 aspects out of 6 ones.

shows better performance in general. While GLAC Net (-Cascading) looks like the best, the human evaluation shows the GLAC Net is the best one in every aspects in Table 3.2 and Figure 3.7. As we mentioned in Section 1, automatic metric scores partially effective on human evaluation.





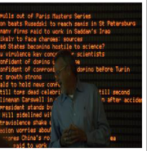
In Figure 3.7, the 6 score results of the human evaluations of the story generation by GLAC Net, ablation models and compared methods. GLACNet shows the best scores in 1~4 criteria. SRT still shows the best in 5~6 ones.

The lower part of Table 3.1 shows the result of automatic metric scores (BLEU, ROUGE-L, METEOR and perplexity). Due to budget problem, we don't perform human experiment on the effect of number of stacks and the aggregation functions. By our observation with self-evaluation of 25 randomly selected cases, they show little significant difference. GLAC Net is the model with 2 stacked BiLSTM and tb aggregation function in the encoder. It shows moderately better scores in overall, and better qualitative results.

3.5 Summary

We propose the GLAC Net that uses global-local attention and context cascading mechanisms to generate stories from a sequence of images. The model is designed to maintain the overall context of the story from the image sequence and to generate context-aware sentences for each image. In the experiment using the VIST dataset, the proposed model proves to be effective and outperforming with total score and 4 aspects of human evaluation criteria out of 6. It shows our method is more focused on overall structure than detailed.

Although the experimental results are promising, the task of visual story-telling remains a challenge. We plan to extend and refine the GLAC architecture to further improve its performance considering local information. In addition, a subject to be studied in the future is how to generate various stories based on

					
Human's	a presentation is taking place in the convention center .	there is tons of wording on the screen .	the man stands next to a projector trying to show evidence of the presentation .	the man wraps up his presentation .	at the end he asks if anyone would like questions answered .
Generated	the man was giving a speech .	he was very happy to be there .	he was very excited .	he was very happy .	he was a great speaker .





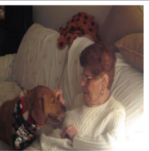
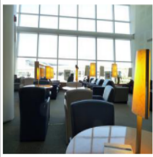

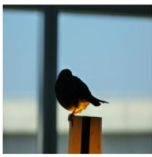


					
Human's 1	going over to see my grandma for christmas this year .	she lives in a very beautiful house outside of the city .	it is always so beautiful out there this time of year .	i was so happy to see my grandma and her great aunt .	of course her dog is also part of the family ! we had a great time .
Human's 2	we started the holidays by <unk> up a tiny christmas tree .	the streets where calm outside . everyone was celebrating the holidays .	i went for an afternoon walk to my favorite spot before heading home .	the family sat together and celebrated christmas .	the night came to an end and everyone slowly went home .
Human's 3	we visited family for christmas .	they live out in the country far from the city .	the trees lost their leaves because it is so cold outside .	they were so happy that we had arrived .	even the dog had a marry christmas .
Generated	the family was having a party .	they went outside to play .	they saw a tree that was very pretty .	they also took photos with the dog .	they had a great time .

Figure 3.8 The generated story examples for GLAC Net and their ground-truth annotated by humans. Even though they are different stories with the same image sequences, the generated ones are acceptable. We can also observe that (1) the length of the generated sentences relatively short. (2) the expression style is concise and abstract, (3) Mostly written in the past tense (more than 98% in the test set)

the purpose and theme from the same image sequence.

					
Human's 1	the school was full of information .	the men we re dressed in safety suits .	i did n't know how well organized the event was .	all the men were coordinated .	i drove my car home .
Human's 2	the action day event had plenty of information available for its attendees .	the staff worked very hard to maintain safety and security throughout the event .	people from all walks of life came to participate .	the march wrapped up and a great time was had by all .	we were having so much fun leaving was the hardest part of the day .
Human's 3	a tourist is viewing the attractions .	the guards are clearing out the area .	the people are protesting .	the guards are stopping the protest .	a man is handing a taxi driver money .
Generated	the protesters were very excited .	they were all protesting .	they were giving a speech .	they were very happy to see their cause .	they were very glad that they were there .

Figure 3.9 More examples of GLAC Net and their ground-truth annotated by humans.

					
Human's 1	i love the view from my office .	so often when i look out the window, there are birds .	this little guy was perched here fro quite awhile this morning .	this one stopped by only for a few minutes .	this afternoon i was able to watch this one for awhile . being i love birds, it s no wonder i love my view .
Human's 2	i saw the bird as soon as i walked in and froze .	i did n't know if i should photograph it quietly or <unk> it away thru and open window .	i was enamored by the lighting upon the bird .	i tried to photograph it , but of course it moved around when it recognized that i was there .	i ended up getting an okay photo , just not the one i wanted .
Generated	the family was excited to go on a trip .	they saw a dog and the dog .	they saw a cat .	they saw a koala .	they also saw a dog .

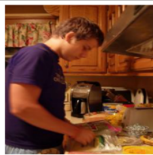




					
Human's 1	this guy got up in the morning and made breakfast .	he drove to school and sat through traffic .	gas was pretty average that day , so he stopped and filled up his tank .	he got to school and got his books for his classes .	he studied throughout the day , making sure to finish his homework .
Human's 2	a typical day for my son . he helps with breakfast .	i let him drive because he just got his license and needs the practice .	i think we 'll wait on filling up and see if the gas price goes down .	the halls are a busy place before classes .	the kids get a break to read during the day before heading home to do homework .
Generated	the man was happy to be at the restaurant .	he was excited about the new car .	the family was there to celebrate .	they were having a great time .	they were very happy .

Figure 3.10 Wrongly generated cases for GLAC Net and their ground-truth annotated by humans. Overall stories do not deviated on one theme, and some of the visual clues are visually-grounded on the images.

Chapter 4

Common Space Learning on Cumulative Contexts and the Next Events: Recurrent Event Retrieval Models

4.1 Overview

This chapter introduces methods to predict the next sentences from the former parts of documents at any step within an arbitrary length, which is an open-ended next sentence prediction problem. The problem can be seen to choose automatically a sequence of events, situations, actions or dialogues that can be told as a story (Martin et al., 2017). If the procedure run automatically and sequentially, our methods can apply to story generation problem (Mostafazadeh et al., 2016; Martin et al., 2017; Huang et al., 2016).

Even though recurrent neural networks (RNNs) shows surprisingly successful cases, it is still difficult to catch the context through several sentences due to an enormously large number of possible scenarios without any constraints (Bowman

et al., 2016)¹

In this chapter, we focus on retrieval approaches for sentences of events. If the number of possible sentences are abundant, it is useful to solve the problem. We propose recurrent event retrieval models (RERMs) to predict the following events or stories with former parts of stories. RERMs are composed of a context accumulation function and two embedding functions, where make close the distance between the cumulative context at current time and the next probable events on a latent space. They update the cumulative context with a new event as a input using bilinear operations on common latent space, and we can find the next event candidates with the updated cumulative context. While it can limit the representational power depending on the number of possible candidates, it is advantageous to focus on the coherence of stories avoiding the difficulty of surface realizing narrative generation. Fortunately, released was ROCStories dataset (Mostafazadeh et al., 2016) to be composed of approximately 100K textual five-sentence commonsense stories for Story Cloze Test (SCT). As a result, the number of possible sentences are around 500K. As RERM evaluation, they show competitive performance for SCT, and the state-of-the-art results in open-ended sequence generation setting. Also, they can be applied to generating stories with humans feedback interactively.

4.2 Problems of Context Accumulation

In this section, let us define our problem. Consider the dataset $D = \{S_1, S_2, \dots, S_N\}$ has N sequences consisting of arbitrary objects $e \in E$, where E is the set of possible objects. The sequence is a variable length of the objects. And there is context c_t at step t . We define sequential retrieval process and the context

¹It is similar to dialogue generation tasks. Currently, researches are focused on task-oriented bot and chit-chat bot (Chen et al., 2017).

integration function f_{integ} to accumulate the contexts and reflect new events together. The process can iterate the procedure with updating the integrated context:

$$c_{t+1} = f_{integ}(c_t, e) \quad (4.1)$$

Also, we define score function s with the integrated context vector c_t at time step t and certain object e , and we can infer the next object e^* via choosing the one of the highest score (or top-n selection) s with the integrated context vector c_t :

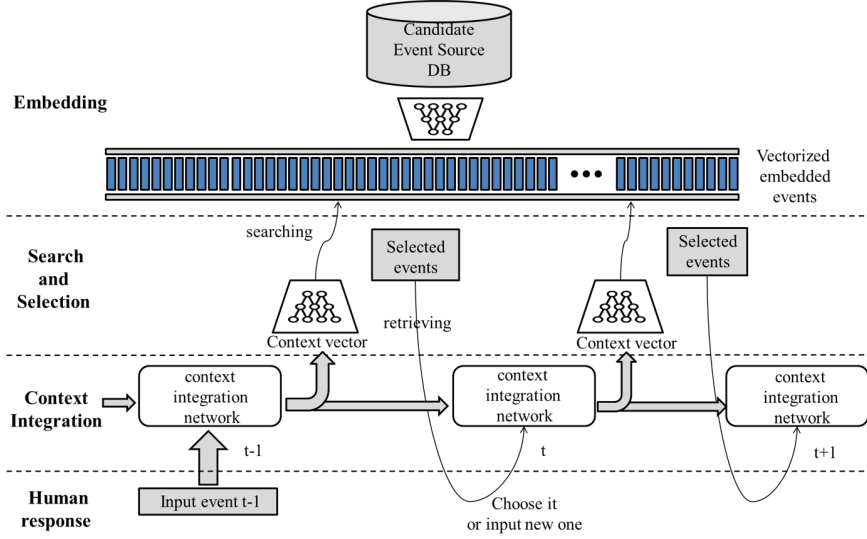
$$e^* = inference(c_t, E) = \operatorname{argmax}_{e \in E} s(c_t, e) \quad (4.2)$$

where E is a set of reference objects such as sentences, images, or videos in the database. If we generate several steps, then it iterate from cue object e to c_{t+1} recurrently.

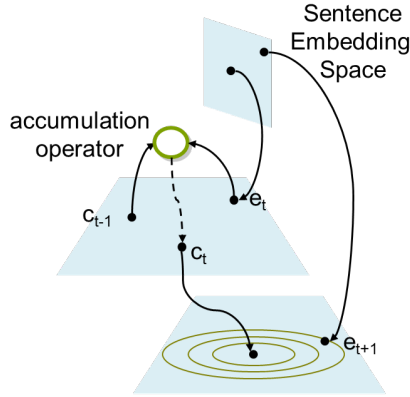
Our goal is to find models capable to learn the context integration functions in order to get good accuracy of inference.

4.3 Recurrent Event Retrieval Models for Next Event Prediction

In this section, we propose novel models to achieve the goal: Recurrent Event Retrieval Models (RERMs). We design RERMs to match contexts and the next events on the common low-dimensional embedding space with each mapping functions, and to integrate them recurrently as shown in Figure 4.1. So, we introduce two embedding functions to deliver objects to the spaces. To represent an event object in the dataset D , we have preprocessing procedure onto one vector. For example, we can use skip-thought vectors as vector representation for sentences of events. They have two embedding functions from sentence embedding space, and have one accumulation operator to accumulate the previous context



(a) Recurrent Event Retrieval Models (RERMs)



(b) Functions and spaces for Recurrent Event Retrieval Models

Figure 4.1 (a) Concept view of Recurrent Event Retrieval Models (RERMs). Every time step, the context vector is updated with certain input object e , and we can infer the next object e^* via choosing the one of the highest score (or top- n selection) s with the integrated context vector c_t . (b) Functions and spaces of RERMs. They have two embedding functions from sentence embedding space, and one accumulation operator for the previous context c_{t-1} and the current input object e_t . The result c_t is embedded onto the other space to compare the similarity scores. We search the closest answers on this space.

c_{t-1} and the current input object e_t . The result c_t is embedded onto the other embedding space to compare the similarity scores. We search the closest answers on this space.

To train all of the functions introduces above, we define the objective functions. In our setting, we modify the error function for sequence setting:

$$Err : \sum_c \sum_k \max\{0, \delta - s(c_t, c_{t+1}) + s(c_t, c_k)\} \quad (4.3)$$

where δ is margin, s is the score function to measure how similar from each other. This objective function is very similar to *triplet ranking loss* with anchor, positive example and negative examples popular used in deep metric learning.

We can use cosine similarity function for training phase, and adding penalty terms for diverse candidates considering the length of story-like sequences for open-ended generation. Practically, we can choose k as the negative examples randomly in the dataset.

Since our representation includes recurrent connections, the computational graphs are unfolded to the length of sequences to be deep structure. To find the appropriate structure for RERMs, we define the general form of embedding nets, context integration nets as shown in Figure 4.2. It is similar to multi-hop end-to-end memory network structure (Sukhbaatar et al., 2015), our models should choose one at every iteration with the labels for coherency consideration. The general form of embedding nets is standard 1 or 2 layered networks of inner product layer with tanh activation (MLP module).

$$f(e) = \alpha_f \cdot f_{MLP}^1(e) + \beta_f \cdot f_{MLP}^2(f_{feats}(e)) \quad (4.4)$$

where the embedding function f is a linear combination of MLP of the input and MLP of the features of the input.

$$h_1(c_t, e) = \alpha_h 1 \cdot c_t \odot e + \beta_h 1 \cdot c_t \otimes e + \gamma_h \cdot CBP(c_t, e) \quad (4.5)$$

$$h_2(c_t) = \alpha_h 2 \cdot f_{MLP}^3(c_t) + \beta_h 2 \cdot c_t \quad (4.6)$$

$$f_{integ}(c_t, e) = h_1(h_2(c_t), e) \quad (4.7)$$

where the integration function f_{integ} is a linear combination of MLP of the previous context, residual connection of it, and the new event input.

Fortunately, linear matrices show good representational power in the case of cross-modal learning or zero-shot learning. The general form of context integration nets has several options: MLP modules, residual connection, and integration operators. Dotted line in the middle in 4.2 (b) can be replaced with element-wise sum, element-wise product or CBP operation. MLP module is not used when we choose CBP as an integration operator.

4.4 Experimental Results

We test on the Story Cloze Test (SCT) as textual stories. Even though the length of stories is fixed, we can test at an arbitrary position.

We formulate a story generation problem as choosing sequentially the proper next events coherently. Naturally, sequential behaviors of humans are not deterministic. So, we should take probabilistic approaches or several candidates together like beam search strategies to consider several possible ways to do as the next one. Additionally, if human’s feedback is available, it is valuable to utilize it. It can induce that the system works interactively. Also, it can update the database on the fly if necessary. Taking all these good points, we build generalized matching modules from the current context to the next events using neural networks with ranking losses to drive semantically consistency, retrieving top-n results easily. When we consider predicting the very next events from the current input, it means that Markov assumption is assumed, which only care the 1-step (or more fixed step) previous input. So, we attach context integration

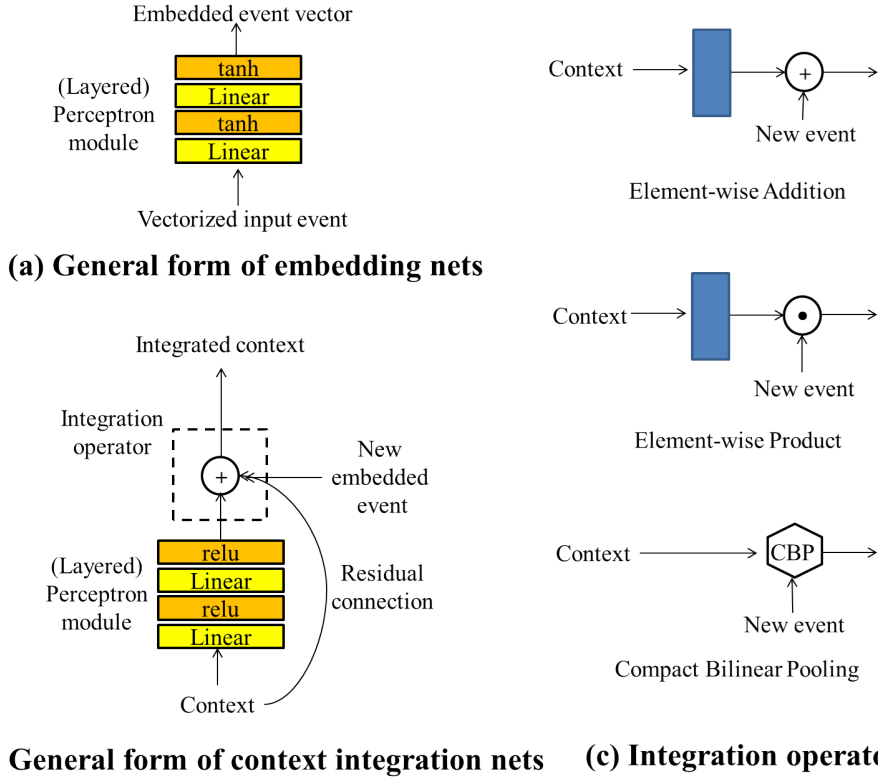


Figure 4.2 Component definition of RERMs: (a) A general form of embedding nets. 1 or 2 layered networks of inner product layer with tanh activation (MLP) module. (b) A general form of context integration nets with MLP module, residual connection and integration operators. Dotted line in the middle can be replaced in other options in (c). The box in (b) between input and operator designate MLP modules.

modules to pursue maintaining all the previous situations. The followings explain the modules respectively.

4.4.1 Preliminaries

Skip-thought Vector: Event representation with multimodal sources is very complicated to process as information. Most popular method for natural language sentence-to-vector conversion is skip-thought vector (STVec) (Kiros et al., 2015). STVec converts one textual sentence to one multi-dimensional real-valued vector (a.k.a thought vector) for general event in text, scene description, or dialogues. It is inspired by the skip-gram structure in popular word2vec (Mikolov et al., 2013). The key concept of word2vec is that the word meaning is determined by the surrounding words. Similarly, the STVec model is trained to reconstruct the surrounding sentences to map the sentences that have semantic meaning onto similar vectors using RNN adding GRU memory cells as language models. Kiros et al. trained 11,038 books with 74,004,228 sentences for STVec, frequently used to represent semantic closeness of sentences

Compact Bilinear Pooling: To deal with multi-modal sources, one of possible way is to use bilinear features from video and text. The bilinear features are very high dimensional, typically on the order of hundreds of thousands to a few million, which drives them impractical. Compact bilinear pooling (CBP) (Gao et al., 2016) is to make two compact bilinear representations with the same discriminative power as the full bilinear representation but with only a few thousand dimensions. CBP is easily expanded to multimodal fusion, it has been applied to visual QA and movie QA problems successfully (Fukui et al., 2016; Na et al., 2017; Tapaswi et al., 2016a).

Performance Metrics: In our experiments, we measure SCT performance for reference, and use 2 metrics: Perplexity and BLEU score, which are based on

Configuration	SCT		Perplexity				Bleu			
	Validation	Test	Step 2	Step 3	Step 4	Step 5	Step 2	Step 3	Step 4	Step 5
Separate	0.642	0.639	173.98	175.86	181.20	183.06	0.139	0.140	0.123	0.128
MLP + residual	0.636	0.657	176.57	174.03	179.42	181.37	0.115	0.108	0.096	0.093
sum	0.625	0.631	200.07	206.31	212.70	215.43	0.139	0.129	0.108	0.104
MLP + sum	0.623	0.636	179.48	185.23	189.37	192.97	0.126	0.123	0.105	0.100
CBP	0.658	0.661	191.49	177.55	190.70	189.72	0.099	0.097	0.087	0.081
MLP+product	0.624	0.641	164.20	176.80	192.42	194.57	0.070	0.073	0.063	0.069
MLP + gate+prod	0.640	0.651	199.01	205.48	205.65	208.18	0.104	0.097	0.085	0.084

Figure 4.3 Performance table of RERMs for exploring the networks. The best SCT performance is shown in CBP configuration for recurrent context integration without MLP module. BLEU scores are higher in separate cases, which means no recurrent connection. Perplexity values are the configuration of preferring sparse information: MLP + product, CBP.

language models or on the golden answers. We can use the next event information as golden answers for BLEU. Perplexity is the measure of how “surprised” a model is by a training set. We use it to check a sense of how well the probabilistic model we trained can predict the data. We built the model using an n-gram length of 1. And BLEU score compares the similarity between the generated output and the “ground truth” with respect to n -gram precision. Those are not perfect measure how well generated stories are acceptable for humans, but other researches use them actively (Martin et al., 2017) in NLP researches. For our experiment, we use adam optimizer and step scheduling for controlling learning rate (initial value: 0.0005, $\times 0.5$ per 50 epoch). Negative examples per one positives: 1499.

4.4.2 Story Cloze Test

Story cloze test (SCT) (Mostafazadeh et al., 2016) is ‘fill-in-the-blank’ task considering the context of 4-sentence story. The candidate sentences for the

blank in the 5-th position are only 2, as to be cast to binary classification problems. Dataset is composed of the Train (98,161 episodes, 490805 sentences), Validate and Test gold data sets (1871 episode each). We use the separation to train our systems, and all of sentences (490,805) in the Train as reference event.

A lot of researchers published their results as shown Table 4.1. However, their approaches are mostly focused on classification, only some of works only possible to generate new candidates or to apply many candidates (in other words, ‘generative’). And it is not scalable to apply to arbitrary length of story (‘open-length’). Our system is approximately generative (not purely generative, it is close to generative considering the number of candidates) and open-length without classification-based learning. Figure 4 and 5 show the performance table of SCT. The best SCT performance is shown in CBP configuration for recurrent context integration without MLP module. BLEU scores are higher in separate cases, which means no recurrent connection. Perplexity values are the configuration of preferring sparse information: MLP + product, CBP.

4.4.3 Open-ended Story Generation

For story generation test, we can control the diversity of candidates controlling the score as follows:

$$score_{overall}(c_q, e_i) = score_1(c_q, e_i) - \lambda \cdot similarity(c_q, e_i) \quad (4.8)$$

The $score_1$ is the same formula with the objective function. We use cosine similarity as the similarity function to same with objective function.

The duplicated sentences are blocked when finding the close candidates. Since named entity recognition can be applied, subjects and objects are changed to the pronoun and check the duplication.

Figure 4.4 shows an example of generated case as following one of the instance

Table 4.1 Performance table of RERMs for Story Cloze Test (SCT). Even though our systems trained for generation, they show relatively good performance (Mostafazadeh et al., 2017; Wang et al., 2017; Chaturvedi et al., 2017).

METHOD	GENERATIVE	OPEN-LENGTH	VALIDATION	TEST
HUMAN	-	-	1.000	1.000
RANDOM	Y	Y	0.514	0.513
FREQUENCY	Y	Y	0.506	0.520
N-GRAM-OVERLAP	Y	Y	0.477	0.494
GENSIM	Y	Y	0.545	0.539
SENTIMENT-FULL	Y	Y	0.489	0.492
SENTIMENT-LAST	Y	Y	0.514	0.522
SKIP-THOUGHTS	Y	Y	0.536	0.552
NARRATIVE-CHAINS-AP	Y	Y	0.472	0.478
NARRATIVE-CHAINS-STORIES	Y	Y	0.510	0.494
DSSM	Y	Y	0.604	0.585
GRU	N	Y	0.573	0.561
CONDITIONAL GAN (WANG ET AL., 2017)	N	Y	0.625	0.609
DAVAR LEXICON (FLOR)	N	Y	0.654	0.620
(ROEMMELE)	N	N	-	0.672
(BUGERT)	N	N	-	0.700
STYLISTIC FEAT (SCHWARTZ)	N	N	-	0.752
FRAME+EMOTION+TOPIC (CHATURVEDI)	N	N	-	0.776
Ours (CBP)	Y	Y	0.661	0.658
Ours (SEPARATE)	Y	Y	0.647	0.647
Ours (MLP+SUM)	Y	Y	0.660	0.641
Ours (MLP+SUM+RESIDUAL)	Y	Y	0.651	0.637

story in the dataset. The figure shows sentences in the cue box are input to RERMs. Then, RERMs outputs top-9 or 10 candidates sentences as the result of top-k closest to the cumulative context point. From (1) to (5), we can see the plausible candidates. Sometimes the right answers are not included in the candidate box. Also, not all of sentences are correct semantically and logically. However, most of them are very plausible with respect to the context.

Figure 4.5 shows an other example of step-wise story generation not following the original episode, but choosing other storyline. Still, RERMs show plausible candidates.

4.5 Summary

This chapter introduces recurrent event retrieval models (RERMs) for open-ended story generation. It is important to be opened at any time step and endlessly via retrieving the related objects considering cumulative context. It explores the appropriate embedding functions and the accumulation operator for RERMs. Additionally, it can be used for interactive setting with humans. Potentially, it can be used for situation inference and easily updated on the conversation with humans.

Cue

'Carrie had just learned how to ride a bike.'

1. 'Carrie had just learned how to ride a bike.'

2. 'She didn't have a bike of her own.'

3. 'Carrie would sneak rides on her sister's bike.'

4. 'She got nervous on a hill and crashed into a wall.'

5. 'The bike frame bent and Carrie got a deep gash on her leg.'

Top-10 candidates

1. 'She wanted to learn how to ride a bike.'
2. 'She had to learn how to ride without training wheels first.'
3. 'She recently decided to take horseback riding lessons.'
4. 'Her dad took her to a ranch so she could ride one for the first time.'
5. 'It would be her first time riding a horse.'
6. 'She didn't know how to ride it without training wheels.'
7. 'She wanted to learn to ride a motorcycle.'
8. 'Finally Juanita decided to learn how to ride a bike.'
9. 'She learned how to ride on an older horse named Beau.'

(1)

Cue

'Carrie had just learned how to ride a bike.'

'She didn't have a bike of her own.'

1. 'Carrie had just learned how to ride a bike.'

2. 'She didn't have a bike of her own.'

3. 'Carrie would sneak rides on her sister's bike.'

4. 'She got nervous on a hill and crashed into a wall.'

5. 'The bike frame bent and Carrie got a deep gash on her leg.'

Top-10 candidates

1. Finally she decided to learn how to ride a bike.
2. She was afraid to learn to ride a bike.
3. She wanted to learn how to ride a bike.
4. So she saved up money for a bike.
5. Carrie had just learned how to ride a bike.
6. She did not know how to ride a bike.
7. She had to learn how to ride without training wheels first.
8. She wanted to learn how to bike.
9. She decided to help her young son Frank learn to ride a bike.
10. She decided to practice riding her bike to school.

(2)

Cue

'Carrie had just learned how to ride a bike.'

'She didn't have a bike of her own.'

'Carrie would sneak rides on her sister's bike.'

1. 'Carrie had just learned how to ride a bike.'

2. 'She didn't have a bike of her own.'

3. 'Carrie would sneak rides on her sister's bike.'

4. 'She got nervous on a hill and crashed into a wall.'

5. 'The bike frame bent and Carrie got a deep gash on her leg.'

Top-10 candidates

1. She was afraid to learn to ride a bike.
2. She had just learned how to ride a bike.
3. Finally she decided to learn how to ride a bike.
4. She was going to learn to ride her bike.
5. She and her friend went riding when Amy got a new bike.
6. She had to learn how to ride without training wheels first.
7. She ended up walking the bike up the hill instead of riding.
8. She did not know how to ride a bike.
9. She wanted to learn how to ride a bike.
10. She was excited for her mother to teach her how to ride a bike.

(3)

Cue

'Carrie had just learned how to ride a bike.'

'She didn't have a bike of her own.'

'Carrie would sneak rides on her sister's bike.'

'She got nervous on a hill and crashed into a wall.'

1. 'Carrie had just learned how to ride a bike.'

2. 'She didn't have a bike of her own.'

3. 'Carrie would sneak rides on her sister's bike.'

4. 'She got nervous on a hill and crashed into a wall.'

5. 'The bike frame bent and Carrie got a deep gash on her leg.'

Top-10 candidates

1. She fell from her bike and bruised her knee.
2. She ended up walking the bike up the hill instead of riding.
3. Suddenly, She fell off the bike and scraped hurt her knee.
4. She had an accident at work and hurt her leg badly.
5. She was going to learn to ride her bike.
6. She and her friend went riding when Amy got a new bike.
7. She had to get a cast and wasn't able to ride the bike for a while.
8. She was afraid to learn to ride a bike.
9. She had just learned how to ride a bike.
10. She was used to riding the bike with training wheels on.

(4)

Cue

'Carrie had just learned how to ride a bike.'

'She didn't have a bike of her own.'

'Carrie would sneak rides on her sister's bike.'

'She got nervous on a hill and crashed into a wall.'

The bike frame bent and Carrie got a deep gash on her leg.

1. 'Carrie had just learned how to ride a bike.'

2. 'She didn't have a bike of her own.'

3. 'Carrie would sneak rides on her sister's bike.'

4. 'She got nervous on a hill and crashed into a wall.'

5. 'The bike frame bent and Carrie got a deep gash on her leg.'

Top-10 candidates

1. She had an accident at work and hurt her leg badly.
2. She shed a tear at the pain and her leg was sore for days.
3. She fell from her bike and bruised her knee.
4. She was surprised the pain did not hurt at all.
5. She ended up walking the bike up the hill instead of riding.
6. She will never forget the day she broke her ankle.
7. She was sad that she would have to postpone her sewing lessons.
8. Suddenly, she fell off the bike and scraped hurt her knee.
9. She is taken to the hospital for her injuries.
10. She needed to beat her coffee crash before volleyball practice.

(5)

Figure 4.4 A RERM demonstration of step-wise story generation with one episode in ROCStories dataset. Sentences in the cue box are input to RERMs. Then, RERMs outputs top-9 or 10 candidates sentences as the result of top-k closest to the cumulative context point. From (1) to (5), we can see the plausible candidates. Sometimes the right answers are not included in the candidate box. Also, not all of sentences are correct semantically and logically. However, most of them are very plausible with respect to the context.

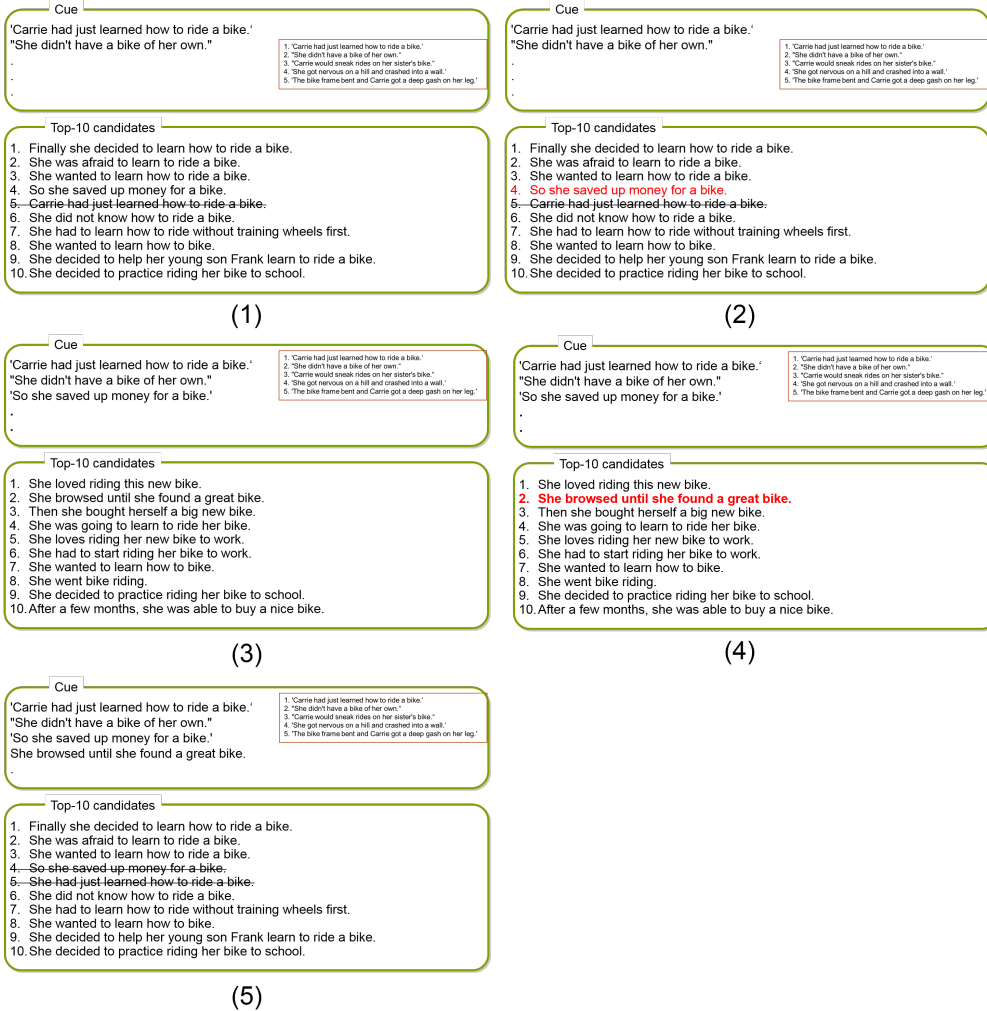


Figure 4.5 An RERM demonstration of step-wise story generation not following the original episode, but choosing other storyline. Still, RERMs show plausible candidates.

Chapter 5

ViStoryNet: Order Embedding of Successive Events and the Networks for Story Regeneration

5.1 Introduction

This chapter introduces the frameworks to regenerate episodes to complete the partially blocked ones on time axis. We train 5~7 minute-long videos including stories with the successive event order supervision for contextual coherence. We explore the question of the successive event order embedding (SEOE) to provide the scaffolds to construct composite representation of ordering and semantics for story generation.

To improve the effectiveness of SEOE on the proposed frameworks, so we give some constraints to reduce the problem complexity. Firstly, we use kids videos as training dataset due to some advantages: (1) omnibus style, which each episode has simple and explicit storyline in short, (2) narrative order mostly using fabula, which follows chronological sequencing of the events, whereas syuzhet is a term

to designate the way a story is organized to enhance the effect of storytelling. (3) relatively small number of main characters and limited spatial environment. This is effective to reduce computational burden and data sparsity. Potentially, these properties are so desirable to provide as surrogate data similar to that of everyday lives in compact and explicit way. Secondly, instead of attaching directly video understanding technologies, we define a contextual event using both of the description sentence including scene context and the dialogue sentence spoken by character, and represent an episode as a sequence of contextual events. And we build the encoder-decoder structure as shown in Figure 5.2, using skip-thought vectors (Kiros et al., 2015) as encoders and sentence generators with standard RNNs as decoders. On the latent space, we learn bi-directional Long Short-term Memory (BiLSTM) with the successive event order embedding (SEOE). To generate multi-step sequences, on the learning process, we control the mixing rate with training data and the generated stories depending on the epoch, which follows scheduled sampling methods (Bengio et al., 2015).

We use ‘PororoQA dataset’, which is the dataset from 3D animation videos for kids, entitled ‘Pororo the Little Penguin’, consisting of 16,066 scene-dialogue pairs created from the video of 20.5 hours in total length, 27,328 fine-grained descriptive sentences for scene descriptions (Kim et al., 2017). With the models to train them, we demonstrate the performance and the generated episodes. We give empirical results for the effectiveness of SEOE. Also, each episode shows a trajectory-like shape on the latent space of the model, which gives the opportunities to interpolate and extrapolate with the geometric information for the sequence models.

5.2 Order Embedding with Triple Learning

In this section, we introduce the background of our works. Starting with classical ordinal embedding, early work only addressed the continuous case, where the x 's span a whole convex subset $U \subset \mathbb{R}^d$. In that setting, the goal of learning embedding functions becomes to characterize *isotonic functions* on U , that is, functions $f: U \mapsto \mathbb{R}^d$ satisfying

$$\|x - y\| < \|x' - y'\| \implies \|f(x) - f(y)\| < \|f(x') - f(y')\|, \forall x, y, x', y' \in U \quad (5.1)$$

Also, we can say that a function $f: U \subset \mathbb{R}^d$ is *weakly isotonic* if

$$\|x - y\| < \|x - z\| \implies \|f(x) - f(y)\| < \|f(x) - f(z)\|, \forall x, y, z \in U \quad (5.2)$$

It is also known that any locally weakly isotonic function on an open U is also locally isotonic on U (Kleindessner and Luxburg, 2014).

Even though we have finite sample only, it is shown that bounded and converged are the difference between isotonic functions and locally weakly

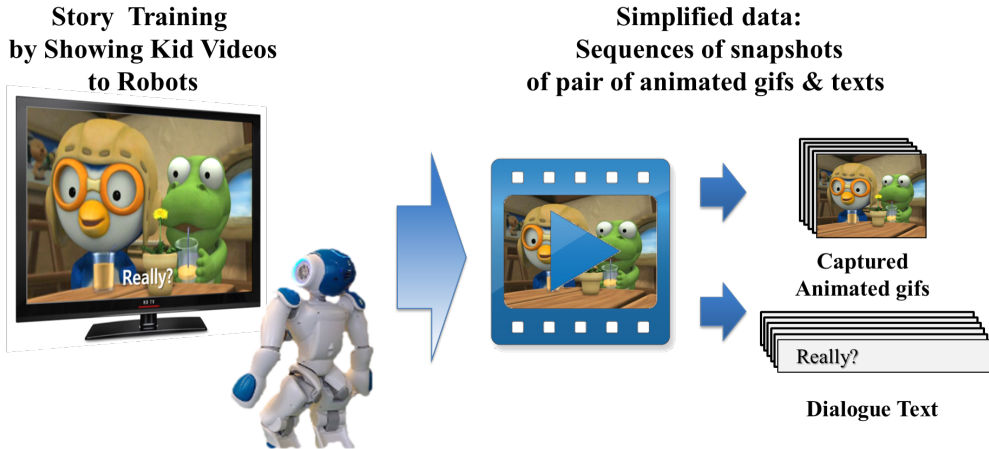


Figure 5.1 Scenario: Robot Training by showing video series. As simplified data, a video stream converted to the stream of snapshots of pairs of animated gifs and dialogue texts.

isotonic function $\phi: \Omega \mapsto \mathbb{R}^d$ with similarity transformation S coincides with ϕ on Ω under some assumptions (Arias-Castro et al., 2017). It is not difficult in finite \mathbb{R}^d space, we can change ordinal learning problem to learning with triples including anchor x in the equation 5.2.

The locality property of ordinal embedding is that if a k -nearest neighbor graph is given as local ordinal constraints, we can reconstruct the point set, which is shown in (Terada and Luxburg, 2014). Its statistical consistency of the embedding method is valid. The consistency can be extended from quadruple learning to triple learning as proven in (Arias-Castro et al., 2017).

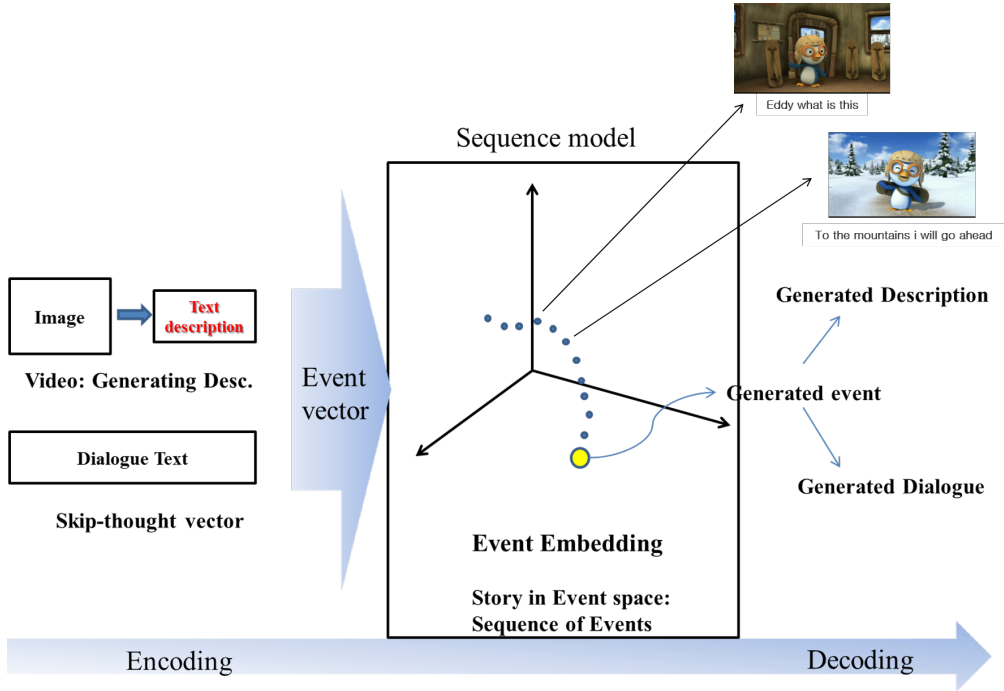


Figure 5.2 Overall encoding-decoding structures for the story completion tasks. A episode is mapped as a trajectory on the embedded event space.

5.2.1 Embedding Ordered Objects in Sequences

We define the problem of sequence generation of order-preserving embedding. In this task, we are given a set of positive examples $P = (u, v)$ of ordered pairs drawn from a sequence (X, \preceq) . And a set of negative examples N that is not the next object. Since every object except the starting position or the last position is in certain sequences, they have the previous object and the next object. That means that all objects in the sequence is connected of the form of chain. Considering the locality property, we can ignore the relationship from the objects in the several steps later. So, we can only consider 3 objects - certain object as an anchor, the next object as the positive example, an not-next object as the negative example.

5.3 Problems and Contextual Events

5.3.1 Problem Definition

We define a story here since the definition of a story is so various (Mateas and Sengers, 2003) depending applications. Since our problems highly depend on videos, we should consider that.

- **Contextual event** is defined as a vector representing the context including situations, actions and dialogues at some moment in videos. (we call it short for “event” in the rest of the paper)
- **Story** (or episode) is defined as a discrete sequence of contextual events aligned in the chronological order including one independent story.

Comparatively, a plot is distinguished from the story. In plots, the sequence of events can be rearranged or some parts of them can be skipped for narrative interestingness. We formulate a story generation task as regenerating the whole story from the partial cues.

Formally, the video dataset $D = \{S_1, S_2, \dots, S_N\}$ has N episodes consisting of a set of video scenes $V = \{v_i\}_{1, \dots, |V|}$, and a set of dialogues $L = \{l_i\}_{1, \dots, |L|}$, where v_i is a sequence of image frames (3-second-long animated GIF images), is a natural language sentence of a dialogue. Each episode $S_n = ((v_{n_1}, l_{n_1}), (v_{n_2}, l_{n_2}), \dots, (v_{n_{|S_n|}}, l_{n_{|S_n|}}))$ is a n -th discrete sequence of pairs of video scene v_{n_i} and dialogue sentence l_{n_i} . The sequence length $n_{|S_n|}$ can be different depending on each episode. For every pair of video scene and dialogue sentence, we encode it into one contextual event vector: $e_{n_i} = \text{encoder}((v_{n_i}, l_{n_i}))$. We assume the encoder can convert the contextual information into a vector.¹ Eventually, each episode is represented as $S_n^e = (e_{n_1}, e_{n_2}, \dots, e_{n_{|S_n|}})$, which is the sequence of variable number of chronologically ordered event vectors e_t at time index t . To get the partial cues, we define a mask $M_n = \{m_i\}_{1, \dots, |S_n|}$ as a binary sequence. Where the value of mask at time t is 1, the event vector value will be $\langle \text{None} \rangle$. The partial cues can be built from M_n and S_n , and we can define error function E between an original story S_n and the generated story \hat{S}_n . Our learning procedure searches for the story S to minimize the error function as follows:

$$\hat{S}^* = \underset{S}{\operatorname{argmin}} E(M_n, S_n, \hat{S}_n; \theta, D) \quad (5.3)$$

We can define several tasks depending on the masking part: former part generation, mid-part generation, and latter part generation. Note that we pursue not only masking part, but also regenerate the whole story. Also, we will consider the problem to capture the structure of the stories. And the problem to generate new whole story based on the structure. Those problems are important to computational narrative intelligence fields (Riedl, 2016), and the context-aware

¹Human’s arbitrary thought can be represented in one vector or not? It is still controversial. But it is not problematic in our work since our sentences are descriptive and relatively simple.

oriented applications such as smart devices and household robots.

5.3.2 Contextual Event Vectors from Kids Videos

Video scene (Animated GIF) to sentence

Event detection and extraction from videos are not easy task, which can be so variously defined depending on the problems. A few seconds of data are focused for most works on video learning such as action recognition task (Simonyan and Zisserman, 2015), upcoming behavior prediction task (Vondrick et al., 2016). These works use spatio-temporal 3D convolution since the coherence between frames very important for those problems. To deal with longer time scale, mostly is used augmented information in text: aligning movies and books (Zhu et al., 2015), movie QA on synopsis and script (Tapaswi et al., 2016b). Similar to the latter, we try to annotate scene description sentences on visual animated scenes to represent semantic information of scenes. It is desirable that each event should have 5W1H information. But, not all scenes have every information explicitly. So, we use scene descriptions augmented by humans on watching the corresponding animated GIFs and dialogue texts (Kim et al., 2017). By this approach, we can reduce the problem complexity and makes us focus on the SEOE. Additionally, recent image captioning tools (Vinyals et al., 2015; Karpathy and Fei-Fei, 2017) also available, we can fine-tune them with the description dataset.

As in 5.1, the snapshots of animated GIFs and texts pairs are used when subtitles appear on the screen. As the formulated above, each snapshot has one event. Humans do not constantly observe every sequential event in a real-world situation to catch the context, but they observe only partly in a temporally aperiodic manner. Sometimes they keep their eyes on carefully, but often they not. Storytelling in a video also constructs narratives in a similar way, sometimes authors intentionally use it. So, we assume that observers can perceive not

seamlessly all of events but some parts of them. In other words, the observer can miss some events in the environments. That is, we should consider that a story we defined above can have intervening events between arbitrary two events, or some events are skipped. We will consider that to design SEOE objectives.

Encoding Events with Skip-thought Vector

The encoder converts to one multi-dimensional real-valued vector (a.k.a thought vector) from the scene description and the dialogue text. Most popular method for this is skip-thought vector (STVec) (Kiros et al., 2015). It is a natural language sentence-to-vector converter inspired by the skip-gram structure in popular word2vec (Mikolov et al., 2013). The key concept of word2vec is that the word meaning is determined by the surrounding words. Similarly, the STVec model is trained to reconstruct the surrounding sentences to map the sentences that have semantic meaning onto similar vectors using RNN adding GRU memory cells (Chung et al., 2014) as language models. Kiros et al. (2015) trained 11,038 books with 74,004,228 sentences for STVec, frequently used to represent semantic closeness of sentences. So, we combine the information of a scene description sentence and a dialogue text together by concatenating the output of STVecs with scene description sentence and that of dialogue text as an event vector:

$$e_t = STVec(desc(v_t)) || STVec(l_t) \quad (5.4)$$

Note that a short-term video clip v_t is converted to scene description sentence with $desc(v_t)$. Following an original setting, STVec converts one sentence into 4800 dimensional real-valued vector. Our event vector has 9600 dimensions eventually.

Naturally, natural language sentences have a lot of variance. Converting the sentences to STVec makes data less sparse? The answer is yes, but very little.

The ratio of unique sentences in the dialogues is 74.11%, but that of STVec is 71.15%. In PororoQA dataset, one character named ‘Crong’ plays a role of a baby dinosaur, he says mostly ‘crong’ instead of concrete answers. Overall, it takes 3.70% of all dialogues. And the ratio of unique sentences in the descriptions is 97.06%, and that of STVec is 96.99%. The ratio of unique event vector defined above shows 99.94%, almost all of event vectors is not same each other. If we use the frequencies of them, it would be problematic, we take the embedding approach introduced in next section instead.

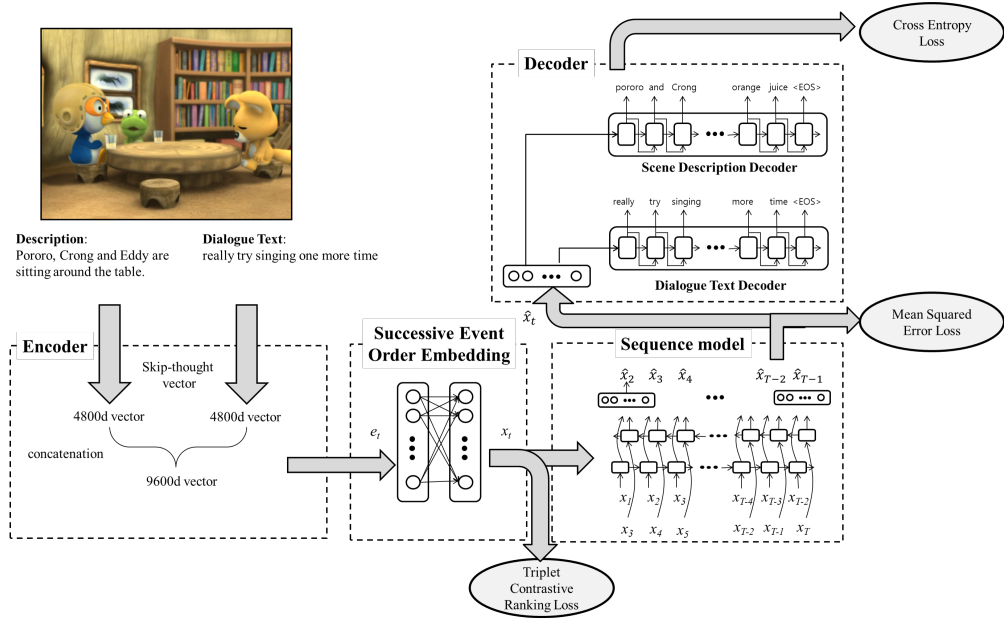


Figure 5.3 Neural architecture of ViStoryNet for story learning and regeneration. Encoder takes two sentences – scene description sentence and dialogue text with skip-thought vectors. Next, Successive event order embedding with triplet ranking loss maps the consecutive events onto ordinaly close points. Those results as inputs are given to sequence models – bidirectional LSTMs to predict previous/next time step vectors. This model can predict missing parts of the sequence. The vectors can be decoded with the sentence generator module as a decoder.

5.4 Architectures for the Story Regeneration Task

The overall structure of the proposed networks is composed of 4 parts: encoder, decoder, sequence model and SEOE module shown in Figure 5.3.

The encoder and the decoder work as the interface for linguistic expression to vectors of semantic information. On the other hand, SEOE induces composite representation of order and semantics. On the space of the representation, sequence models learn the inter-relation among the events of the composite representation (embedded events). To discriminate the vector of the embedded event space from an encoded vector e_{n_i} introduced earlier, we use x_{n_i} as a vector in the embedded event space.

The followings are a summary of 4 components.

- 1) Encoder: the conversion function from an input video (a sequence of pairs of animated GIFs and dialogue texts) to the sequence of events $S_n^e = (e_{n_1}, e_{n_2}, \dots, e_{n_{|S_n|}})$ with the concatenation of skip-thought vectors of scene descriptions and dialogues
- 2) Order Embedding: the embedded vectors $x_t = f_{SEOE}(e_t)$ with the embedding function to map the consecutive events onto ordinally close points in the common latent space. We will explain it in Chapter 5.4.2.
- 3) Sequence Model: BiLSTMs (the forward LSTMs and the backward LSTMs) and the combining function of the outputs of BiLSTMs. \vec{h}_t is the hidden state of the forward LSTMs at time t and \overleftarrow{h}_t is the hidden state of the backward LSTMs at time t .

$$\begin{aligned}
 \vec{h}_t &= LSTM_f(x_{t-1}, \vec{h}_{t-1}) \\
 \overleftarrow{h}_t &= LSTM_b(x_{t+1}, \overleftarrow{h}_{t+1}) \\
 \hat{x}_t &= f_{combine}(\vec{h}_t, \overleftarrow{h}_t)
 \end{aligned} \tag{5.5}$$

- 4) Decoders: two functions from \hat{x}_t to the generated scene description $\widehat{desc}(v_t)$, and the generated dialogue text \hat{l}_t each.

$$\begin{aligned}\widehat{desc}(v_t) &= f_{desc}(\hat{x}_t) \\ \hat{l}_t &= f_{diag}(\hat{x}_t)\end{aligned}\tag{5.6}$$

As a result, the output of sequence models $\hat{S}^e = (\hat{x}_1, \hat{x}_2, \dots, \hat{x}_T)$ with the length T and the surface realized story $\hat{S} = ((\widehat{desc}(v_1), \hat{l}_1), (\widehat{desc}(v_2), \hat{l}_2), \dots, (\widehat{desc}(v_T), \hat{l}_T))$ is generated with two decoders.

5.4.1 Two Sentence Generators as Decoders

From the points in the embedded event space, decoders recover the event vector to the corresponding sentences: scene description and dialogue text. The decoders generate sentences from vectors in the latent space (as eq. 5.6). The output vectors from sequence models \hat{S}^e are given as the inputs to the decoders. The decoders are implemented with RNNs with GRU cells. The decoders work as open-loop mode, they have no additional other input without initial hidden vectors. The output vectors use as the input for the next time step. In this setting, cross entropy loss is widely used for one-hot representation for words in the dictionary. The output value $\hat{y}_{w,n}$ and label $y_{w,n}$ with the w -th word, t -th time step of RNNs have their probabilities given by softmax function, the cross entropy loss as follows:

$$-\sum_n \sum_w [y_{w,n} \log \hat{y}_{w,n}]\tag{5.7}$$

5.4.2 Successive Event Order Embedding (SEOE)

SEOE is the core module of this paper, which build for the structure for contextual coherence. Based on the facts that every story has implicitly shared situation in the short term, and narratives in videos are composed of selective

shots of observations (not seamlessly continuous showing), we consider that contextual events should be embedded separately and not evenly distant on consecutive events. Also, it is enough to check neighbors of each nodes in the chain graphs as shown in the previous section. Considering the assumptions and properties, we adopt triplet ranking loss as objective function to learn SEOE. By using this objective function, we focus on the ordinal information in the sequences ignoring the inter-distance information.

x_t is an embedded event vector at time t . The embedding function f_{SEOE} maps an event vector onto the embedded space. The goal of function $f_{SEOE}(e_t)$ is to make event vectors be reorganized on the latent space so that each episode constructs to show trajectory-like relationship in the space by embedding consecutive events onto ordinal close points:

$$x_t = f_{SEOE}(e_t) \quad (5.8)$$

This can be achieved with triplet ranking loss as follows:

$$\min_{\theta} \sum_x \sum_k \max\{0, \alpha - s(x_t, x_{t+1}) + s(x_t, x_k)\} \quad (5.9)$$

where α is margin, s is score function to measure how far from each other. We use cosine distance for it. k is any other indices except for $t + 1$.

Function Forms of SEOE

As the function form of SEOEs, we test 1-layered or 2-layered fully connected neural networks, which is necessary to represent more than isomorphic relationship. These modules can easily integrate other neural networks. To use triplet loss, we need one positive example and one negative example per one case. For x_t , we can choose x_{t+1} as the positive example deterministically. The negative examples are chosen randomly from the training set and resampled every epoch.

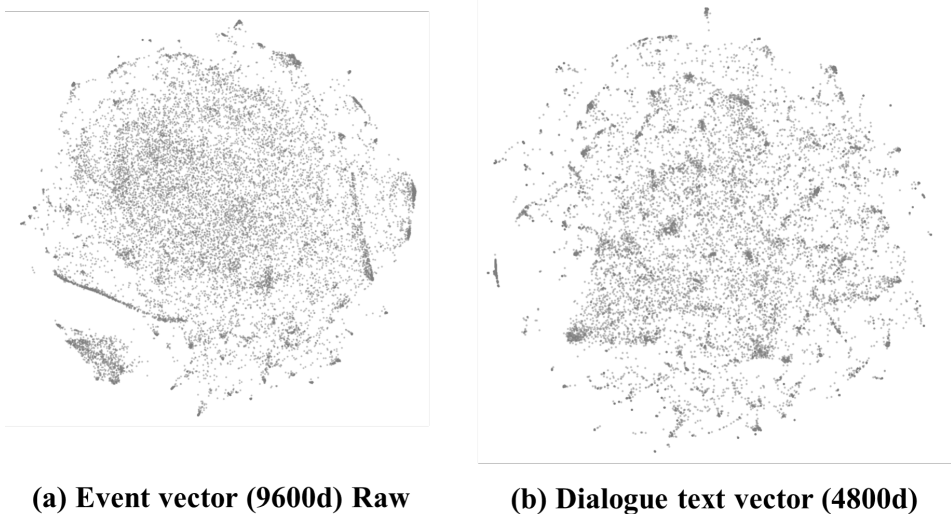


Figure 5.4 t-SNE Visualization of (a) all of event vectors and (b) dialogue text vectors before the SEOE procedure.

Visualization Result of SEOE

To observe the tendency of learning with SEOE, we visualize them with t-SNE. Figure 5.4 shows the overall structure of event vectors e_t . In the dialogue text, sometime exactly the same sentences are shown such as greetings like ‘Hi, friends’ and ‘nice to meet you’. Some clusters are shown in Figure 5.4 (b) representing frequent sentences, events vectors in Figure 5.4 (a) have less number of clusters. Applying SEOE on the event vectors, we can visualize the overall structure built by all events, and trajectory-like embedding results as shown in Figure 5.5. To check how many events are follow this property, we analyze them with 2-nearest neighbors for each event vectors on the episodes. The percentages are around 99.4% of event vectors follow the property. (1-layer NN: 99.41%, 2-layer NN: 99.37%, No embedding: 2.035%)

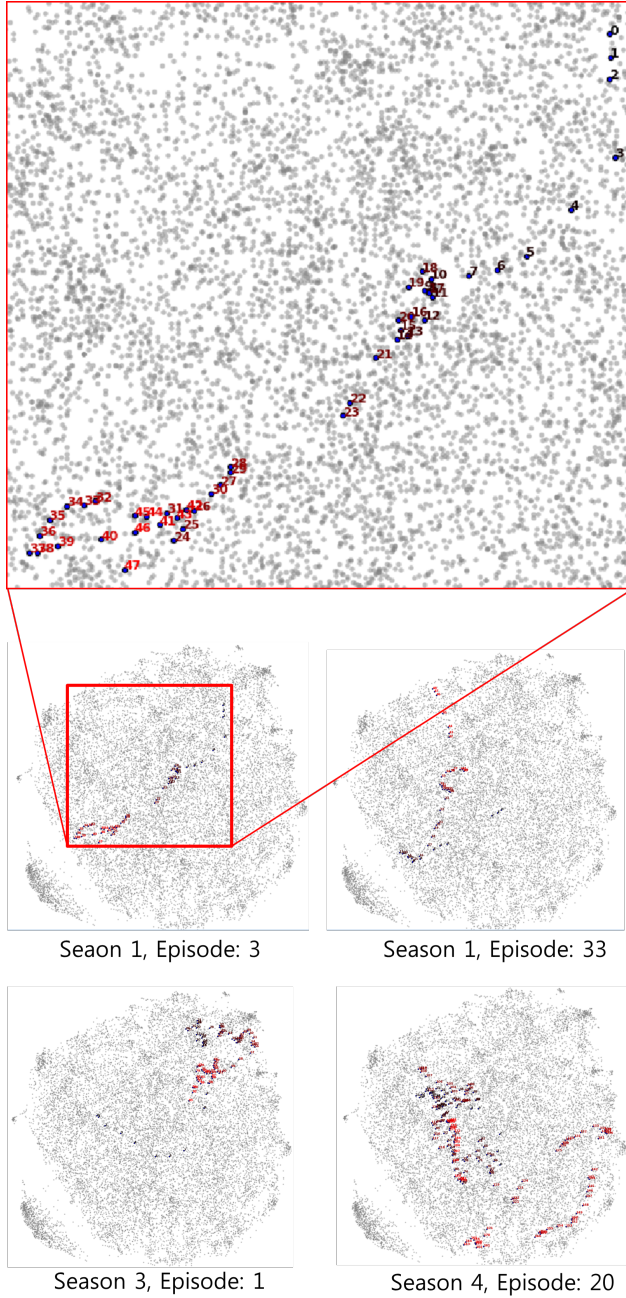


Figure 5.5 Visualization of episodes of the trajectory-like form in the embedded event space. The numbers designates the indices of events from the beginning. We color the points with black close to the start, and red close to the end. The tendency is maintained over all episodes without dependence of the length of episodes.

5.4.3 Sequence Models of the Event Space

Since our tasks need to generate sequences of events with multi-step prediction in arbitrary directions, we use bi-directional Long Short-term Memory (BiLSTM) adding the learning process following scheduled sampling methods (Bengio et al., 2015). We control the mixing rate with training data and the generated stories depending on the epoch. At the beginning, the portion of training data is high. As the epoch goes by, the generated stories are involved gradually more. After the certain epoch, the models are trained with the generated ones. This is helpful to alleviate generating wrong answers in open-loop LSTMs.

Bi-directional LSTM (BiLSTM): Neural language models with long short-term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) or Gated Recurrent Unit (GRU) (Chung et al., 2014) are very powerful to be used in sequence modeling to embrace long-term dependency information. BiLSTMs pursue to increase the amount of input information available using both of two directional hidden states. In our problems, we should predict arbitrary position of events such as early part of the sequence given the latter part. To utilize BiLSTMs, we use two LSTMs for forward direction and backward direction each. The input indices are rather different, from x_1 to x_{T-2} for the forward direction, from x_3 to x_T for the backward direction as shown in Figure 5.3. We combine those outputs with one fully connected layer of the concatenation of two inputs as follows:

$$f_{combine}(x_f, x_b) = \tanh(W \cdot (x_f || x_b) + b) \quad (5.10)$$

where x_f and x_b are two input vectors, W is a weight matrix and b is a bias vector.

To predict arbitrary missing parts, we can do similar to open-loop RNNs. To learn the sequences, the next step vectors are used as targets. $f_{combine}$ is the

function of the outputs of 2 LSTMs. (as eq. 5.5). Objective function is the mean squared error (MSE) loss:

$$\sum_x \sum_t (x_t - \hat{x}_t)^2 \quad (5.11)$$

Eventually, we use BiLSTMs as regression models for the embedded event vectors. As is done with mixture density networks (Graves, 2013) for generating sequences with RNNs, MSE loss can be seen in a probabilistic manner if a Gaussian distribution with fixed isotropic covariance on each output node.

5.5 Experimental Results

5.5.1 Experimental setup

We split all 171 episodes of the ‘PororoQA dataset’ into 90% training (154 episodes) / 10% test (17 episodes).² Lengths of episodes on trainset/testset are similar (mean: 93.1 vs 101.76, std: 72.13 vs 80.73). The evaluation methods are mean square error (MSE) between original vectors and generated ones in embedded vector spaces. We test 3 tasks: former-part prediction, mid-part prediction, latter-part prediction. MSE is advantageous to tune the overall difference from the previous / current / next desired vectors together than the difference at each time index. We check the effect of the masking lengths and scheduled sampling.

5.5.2 Quantitative Analysis

Figure 5.6 shows the performance of 3 tasks. Additionally, the results using scheduled sampling and the results of standard open-loop prediction only (normal). At first, single directional LSTMs show good for short-length prediction as the length of masking part are decreased. (Figure 5.6 (a)). In Figure 5.6 (b), the

²In the standard approach, validation set is often used for model selection. In ours, the learning curves show the convergence to the certain value. So, we split the dataset as above and we stop the iteration after the error variance is enough small.

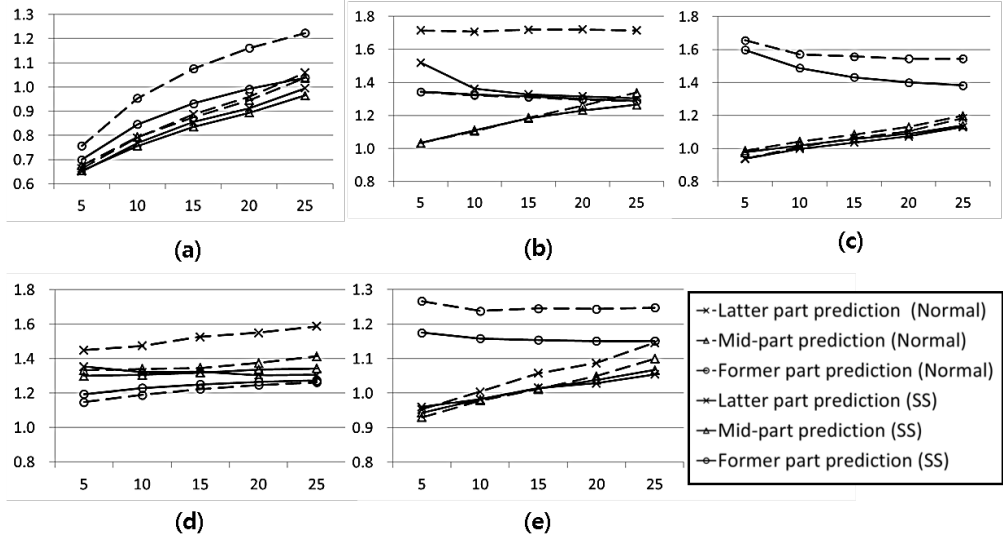


Figure 5.6 Plots for prediction errors (test data) on single directional LSTMs, open-loop output of 2 LSTMs, and combined nodes of Bidirectional LSTMs. Error metrics are the mean squared error (MSE). (solid line: scheduled sampling, dotted line: normal)

prediction error of the result of single directional open-loop LSTMs. They show similar tendencies with single LSTMs. In Figure 5.6 (c), BiLSTMs show low MSE and stable performance regardless of the length of masking part. In the area in longer than 200, the number of episodes is just 1, the result is so noisy. Entirely, scheduled sampling is helpful to be stable with respect to prediction error. From the results, we can generate stories.

5.5.3 Qualitative Analysis

To do surface realization, we train two decoders overfitted intentionally with the overall dataset, since we want to use decoders as the probe what events are encoded only, not a generalized surface realizer. The performance of two decoders is as shown in Table 5.1.

Interpolation via geometric mid-points: We can interpolate mid-point

Data	BLEU(1)	BLEU(2)	BLEU(3)	BLEU(4)	CIDEr	METEOR	ROUGE-L
Dialogue	0.989	0.987	0.986	0.984	8.901	0.849	0.882
Description	0.942	0.935	0.931	0.927	9.501	0.686	0.970

Table 5.1 Decoder performance evaluation results. They are trained overfitted intentionally. We want to use decoders as the probe what events are encoded only, not a generalized surface realizer. The generated descriptions shows lower scores, especially METEOR. Since the length of sentences of descriptions are longer, it is natural phenomenon.

events between two arbitrary ones. At first, we check the effects of SEOE with 1-step gap interpolation. Since the number of instances is limited, 1-step gap interpolation is not so interesting. But, we can see the noisy sentences are intervened between them if events are not ordered. As shown in Figure 5.7, we can observe that noisy events are appeared in case of without-SEOE. Also, the decoding boundary is more accurate in the using-SEOE case. Figure 5.8 shows the example of 5-step gap interpolation results, which goes out from the original storyline and come back. When we get the mid-points between two events, it would get better results if we get them to follow the trajectories. Since we use sequence models to track the trajectories in our system, it shows not clear results if we just find geometric midpoints.

Regeneration with sequence models: Figure 5.9 shows one of randomly chosen generated examples in the test data as the latter part generation problem. We observe that almost every sentences are grammatically correct, and whole story is reorganized and regenerated.

Figure 5.9 and 5.10 show randomly chosen one of relatively short episodes (smaller than 50 step). Also, it is marked with color boxes as shown in Figure 5.10 to check how many sentences are similar. We use yellow color for perfect-matched on the ground-truth. Blue one is for the matching case with on the 1-step shifted ones. Unless forcing to update with the cues, all of the stories

Interpolation Result without SEOE		
<div> <div>Inst. → 0</div> <div>(time t)</div> <div>Interpolation</div> </div>	0	Description: Pororo and Crong are about to go to sleep in their bed Dialogue Text: Crong
	1	Description: Pororo and Crong are about to sleep to sleep in their bed Dialogue Text: Crong
	2	Description: Pororo and Crong are about to sleep with their hands in their bed Dialogue Text: Crong
	3	Description: Pororo and Crong are having a conversation in his house to sleep and They are sleeping Dialogue Text: Crong
	4	Description: Pororo is shaking his head to Crong and Pororo are sleeping in their house Dialogue Text: Crong
	5	Description: Pororo is holding his withered plant, and Crong is still in his bed, and They are in Pororo's house Dialogue Text: Crong the flower
	6	Description: Pororo is holding his withered plant, while Crong is still in his bed, and They are in Pororo's house Dialogue Text: Crong the flower is
	7	Description: Pororo is holding his withered plant, while Crong is still in his bed, and They are in Pororo's house Dialogue Text: Crong the flower is dying
	8	Description: Pororo is holding his withered plant, while Crong is still in his bed, and They are in Pororo's house Dialogue Text: Crong the flower is dying
<div> <div>Inst → 9</div> <div>(time t+1)</div> </div>		
Interpolation Result with SEOE		
<div> <div>Inst → 0</div> <div>(time t)</div> <div>Interpolation</div> </div>	0	Description: Pororo and Crong are about to go to sleep in their bed Dialogue Text: Crong
	1	Description: Pororo and Crong are about to sleep to sleep in their bed Dialogue Text: Crong
	2	Description: Pororo and Crong are about to sleep with their hands in their bed Dialogue Text: Crong
	3	Description: Pororo and Crong are having a conversation in his house to sleep and They are sleeping Dialogue Text: Crong
	4	Description: Pororo is shaking his head to Crong and Pororo are sleeping in their house Dialogue Text: Crong
	5	Description: Pororo is holding his withered plant, and Crong is still in his bed, and They are in Pororo's house Dialogue Text: Crong the flower
	6	Description: Pororo is holding his withered plant, while Crong is still in his bed, and They are in Pororo's house Dialogue Text: Crong the flower is
	7	Description: Pororo is holding his withered plant, while Crong is still in his bed, and They are in Pororo's house Dialogue Text: Crong the flower is dying
	8	Description: Pororo is holding his withered plant, while Crong is still in his bed, and They are in Pororo's house Dialogue Text: Crong the flower is dying
<div> <div>Inst. → 9</div> <div>(time t+1)</div> </div>		
Description: Pororo is holding his withered plant, while Crong is still in his bed, and They are in Pororo's house Dialogue Text: Crong the flower is dying		

Figure 5.7 A comparative interpolation example of the 1-step gap with and without SEOE. In case of without SEOE, noisy events are observed.

Inst. (time t)	(0, 'Generated dialogue:', ['cro .'])
	(0, 'Generated description:', ['Crong does not know where Pororo went, and Crong is in Pororo's house.'])
	(1, 'Generated dialogue:', ['cro .'])
	(1, 'Generated description:', ['Pororo gets hit by Crong's toy arrow, and Pororo is in his house.'])
	(2, 'Generated dialogue:', ['cro .'])
	(2, 'Generated description:', ['Pororo opens the door and says hi to Poby, Loopy, and Crong is sleeping.'])
	(3, 'Generated dialogue:', ['huh loopy is here too .'])
	(3, 'Generated description:', ['Pororo opens the door and says in his house.'])
	(4, 'Generated dialogue:', ['loopy crong .'])
	(4, 'Generated description:', ['Pororo asks to Loopy that Crong did try hula hoops, and Crong is pointing at the hoop.'])
	(5, 'Generated dialogue:', ['loopy is in danger .'])
	(5, 'Generated description:', ['Loopy asks to Pororo whether or not Pororo finishes fixing her for her clock.'])
	(6, 'Generated dialogue:', ['loopy crong .'])
Inst. (time t+5)	(6, 'Generated description:', ['Loopy enters into Pororo's sleigh and Loopy are talking to each other in a snow forest.'])
	(7, 'Generated dialogue:', ['come in .'])
	(7, 'Generated description:', ['Loopy enters into Pororo's house, and The woods are covered with snow, and The sky is blue and clear.'])
	(8, 'Generated dialogue:', ['come in .'])
	(8, 'Generated description:', ['Loopy enters into Pororo's house, and The woods are covered with snow, and The sky is blue and clear.'])
	(9, 'Generated dialogue:', ['come in .'])
	(9, 'Generated description:', ['Loopy enters into Pororo's house, and The woods are covered with snow, and The sky is blue and clear.'])

Figure 5.8 An interpolation example of the 5-step gap with SEOE.

are changed a little. There are some pairs that semantically same sentences or only some words are different. We mark them as green one. We can observe that large part of regenerated result of dialogues is recovered and the descriptions are relatively small part is.

5.6 Summary

We propose story learning and regeneration framework for kids videos as surrogate data of everyday lives. This type of datasets is meaningful to research context understanding in real life. Descriptive story generators also are introduced using the framework. Successive Event Order Embedding (SEOE) builds composite representation of order and semantics, which shows the structure of episodes and give stable regeneration result. We observe the potential to interpolate events between arbitrary events, and we can get better results with sequence models to span the event space. Note that it is still limited due to the relatively small number of instances.

<Start of Episode>

0	Generated [Diag]:	crong crong .	[Desc]: Pororo and his friends are standing outside the wooden house.
0	Right Ans [Diag]:	where is it .	[Desc]: Eddy is looking for something in Eddy's library.
1	Generated [Diag]:	where is it .	[Desc]: Eddy is walking on the snow, and Eddy suddenly moves his face and Eddy smiles.
1	Right Ans [Diag]:	i saw it somewhere around here .	[Desc]: Eddy is standing on the ladder and Eddy is finding something on the bookshelves.
2	Generated [Diag]:	i saw it somewhere around here .	[Desc]: Eddy is walking on the snow, and Eddy suddenly hear something and comes to the sky.
2	Right Ans [Diag]:	ah i found it .	[Desc]: Eddy found the book, and Eddy climbs down a ladder.
3	Generated [Diag]:	what should i do .	[Desc]: Eddy says that it is the historical moment to invent novel chemicals.
3	Right Ans [Diag]:	what should i make today .	[Desc]: Eddy looks at the book and questions himself.
4	Generated [Diag]:	what should i make today .	[Desc]: Eddy looks at the book and questions himself.
4	Right Ans [Diag]:	okay that looks good .	[Desc]: Eddy came up with an idea and Eddy decides to make something.
5	Generated [Diag]:	eddy .	[Desc]: Eddy came up with an idea.
5	Right Ans [Diag]:	eddy .	[Desc]: Pororo and Crong visit Eddy.
6	Generated [Diag]:	you came down .	[Desc]: Pororo and crong are making a car, and Poby holds a help in the other hand.
6	Right Ans [Diag]:	eddy .	[Desc]: Pororo and crong calls out eddy.
7	Generated [Diag]:	i can put it out .	[Desc]: Eddy says first will stop right friends.
7	Right Ans [Diag]:	yes come in .	[Desc]: Eddy allows Pororo and Crong to come in, and Crong opens the door.
8	Generated [Diag]:	hi .	[Desc]: Petty and Loopy think that the doll looks delicious.
8	Right Ans [Diag]:	hi .	[Desc]: Pororo and Crong say hi to Eddy.
9	Generated [Diag]:	eddy what are you doing .	[Desc]: Eddy and Pororo are surprised, and Eddy is little bit unpleasant.
9	Right Ans [Diag]:	eddy what are you doing .	[Desc]: Behind Eddy, there is a car, and Pororo asks him a question.
10	Generated [Diag]:	oh i am making a new toy .	[Desc]: Eddy touches Eddy's head and explains that Eddy is making a new toy.
10	Right Ans [Diag]:	oh i am making a new toy .	[Desc]: Eddy touches Eddy's head and explains that Eddy is making a new toy.
11	Generated [Diag]:	oh i am making a new toy .	[Desc]: Eddy looks at Petty and smiles.
11	Right Ans [Diag]:	toy crong .	[Desc]: Pororo and Crong look at the car and run toward the car.
12	Generated [Diag]:	toy crong .	[Desc]: The sides of mailbox is colored by pink, and the word, POST, is also written on the front side of the mailbox.
12	Right Ans [Diag]:	it looks great .	[Desc]: Pororo and Crong stand in front of the car, and they think the car is cool.
13	Generated [Diag]:	i will take it to the playground .	[Desc]: Eddy comes over and says Eddy's plan.
13	Right Ans [Diag]:	i will take it to the playground .	[Desc]: Eddy comes over and says Eddy's plan.
14	Generated [Diag]:	i will take it to the playground .	[Desc]: Eddy is looking for eddy friends.
14	Right Ans [Diag]:	to the playground .	[Desc]: Crong is looking around the car, and Pororo asks a question.
15	Generated [Diag]:	to the playground .	[Desc]: Petty and Loopy think that the situation is strange.
15	Right Ans [Diag]:	then all of us can ride in it .	[Desc]: Eddy stretches Eddy's arms, and Eddy is taking the car to the playground for everyone.
16	Generated [Diag]:	i will bring something to drink something .	[Desc]: Pororo agrees with Eddy and wants to help Eddy.
16	Right Ans [Diag]:	i will bring it there .	[Desc]: Pororo agrees with Eddy and wants to help Eddy.
17	Generated [Diag]:	let go .	[Desc]: Pororo asks friends if Eddy did go.
17	Right Ans [Diag]:	really .	[Desc]: Eddy is surprised, and Pororo comes closer to the head of the car.
18	Generated [Diag]:	i will be there first .	[Desc]: Pororo and Crong are waiting for Loopy.
18	Right Ans [Diag]:	i will be there first .	[Desc]: Pororo takes car to the playground, and Eddy and Crong are surprised.
19	Generated [Diag]:	i will be there first .	[Desc]: Pororo goes out of Eddy's house, and Poby is surprised by Pororo.
19	Right Ans [Diag]:	crong .	[Desc]: Pororo goes out of Eddy's house, and Eddy is standing still, and Crong follows Pororo.
20	Generated [Diag]:	crong .	[Desc]: Pororo goes out of Eddy's house, and is approaching to Pororo.
20	Right Ans [Diag]:	crong crong .	[Desc]: Pororo runs with the car, and Crong follows Pororo.
21	Generated [Diag]:	we are safe .	[Desc]: The night arrives in the forest.
21	Right Ans [Diag]:	it is hard .	[Desc]: Pororo stops at the top of the hill, and Crong stops too.
22	Generated [Diag]:	it is hard .	[Desc]: Poby says that Harry is playing with the ball, and Harry is excited.
22	Right Ans [Diag]:	crong .	[Desc]: Crong and Pororo are surprised, and Crong swings Crong's arms.
23	Generated [Diag]:	crong .	[Desc]: Crong and Pororo are surprised, and Crong swings Crong's arms.
23	Right Ans [Diag]:	hey stop there .	[Desc]: The car is going down a hill, and Pororo follows, and Crong stamps Crong's feet repeatedly.
24	Generated [Diag]:	hey stop there .	[Desc]: pororo runs toward crong with angry face, and crong and pororo smile at pororo.
24	Right Ans [Diag]:	crong .	[Desc]: Crong thinks the situation is weird.
25	Generated [Diag]:	pororo crong .	[Desc]: Crong says it is not a luck.
25	Right Ans [Diag]:	crong .	[Desc]: Eddy calls Crong, and Crong looks back.
26	Generated [Diag]:	what are you doing here .	[Desc]: Eddy asks Pororo what Crong is doing.
26	Right Ans [Diag]:	what are you doing here .	[Desc]: Eddy asks Crong what Crong is doing.
27	Generated [Diag]:	what are you doing here .	[Desc]: Eddy asks Crong what Crong is doing.
27	Right Ans [Diag]:	crong .	[Desc]: Crong points something to Eddy.
28	Generated [Diag]:	it looks fun .	[Desc]: Eddy looks angry, and Pororo says something to Eddy.
28	Right Ans [Diag]:	oh no pororo disappeared in a new toy car .	[Desc]: Pororo is disappearing with a new toy car, and Eddy and Crong are surprised.
29	Generated [Diag]:	oh no pororo disappeared crong .	[Desc]: Pororo is disappearing with a new toy car, and Eddy and Crong are surprised.
29	Right Ans [Diag]:	lalalala .	[Desc]: Loopy is walking while singing.
30	Generated [Diag]:	i will make you a sorry .	[Desc]: Eddy, Eddy and Pororo are standing in front of the door, and Pororo is waving to Eddy saying what is going to his house.
30	Right Ans [Diag]:	hi loopy .	[Desc]: On a new toy car, Pororo say hi to Loopy, and Pororo is going fast.
31	Generated [Diag]:	hi loopy .	[Desc]: On a new toy car, crong talks to Loopy, and Eddy is little bit surprised.
31	Right Ans [Diag]:	hey pororo .	[Desc]: Loopy says hi, but Pororo is too fast.
32	Generated [Diag]:	what are you doing .	[Desc]: Pororo is surprised at Pororo's house, and Pororo is also smiling.
32	Right Ans [Diag]:	what are you doing .	[Desc]: Loopy is looking at Pororo, and Loopy is surprised.
33	Generated [Diag]:	there you go .	[Desc]: Loopy and Eddy are amazed that Crong feels something.
33	Right Ans [Diag]:	this is a new toy eddy made .	[Desc]: Pororo yells at Loopy that it is a Eddy's new car.
34	Generated [Diag]:	this is a new toy eddy made .	[Desc]: Pororo yells at Loopy that it is a Eddy's new car.
34	Right Ans [Diag]:	yahoo .	[Desc]: Pororo is enjoying a ride, and Pororo yells with joy.
35	Generated [Diag]:	yahoo .	[Desc]: Pororo is angry and coming to crong with a sleigh.
35	Right Ans [Diag]:	lalalala lalalala .	[Desc]: Loopy is walking merrily on a snowy land.

Figure 5.9 One of latter part prediction examples with test data (given: 0 ~ 20). Whole story is reorganized and generated.

<Start of Episode>		
0 Generated [Dia]	crong crong .	[Desc]: Pororo and his friends are standing outside the wooden house.
0 Right Ans [Dia]	where is it .	[Desc]: Eddy is looking for something in Eddy's library.
1 Generated [Dia]	where is it .	[Desc]: Eddy is walking on the snow, and Eddy suddenly moves his face and Eddy smiles.
1 Right Ans [Dia]	I saw it somewhere around here .	[Desc]: Eddy is standing on the ladder and Eddy is finding something on the bookshelves.
2 Generated [Dia]	I saw it somewhere around here .	[Desc]: Eddy is walking on the snow, and Eddy suddenly hear something and comes to the sky.
2 Right Ans [Dia]	ah i found it .	[Desc]: Eddy found the book, and Eddy climbs down a ladder.
3 Generated [Dia]	what should i do .	[Desc]: Eddy says that it is the historical moment to invent novel chemicals.
3 Right Ans [Dia]	what should i make today .	[Desc]: Eddy looks at the book and questions himself.
4 Generated [Dia]	what should i make today .	[Desc]: Eddy looks at the book and questions himself.
4 Right Ans [Dia]	okay that looks good .	[Desc]: Eddy came up with an idea and Eddy decides to make something.
5 Generated [Dia]	eddy .	[Desc]: Eddy came up with an idea.
5 Right Ans [Dia]	eddy .	[Desc]: Pororo and Crong visit Eddy.
6 Generated [Dia]	you came down .	[Desc]: Pororo and crong are making a car, and Poby holds a help in the other hand.
6 Right Ans [Dia]	eddy .	[Desc]: Pororo and crong calls out eddy.
7 Generated [Dia]	I can put it out .	[Desc]: Eddy says first will stop right friends.
7 Right Ans [Dia]	yes come in .	[Desc]: Eddy allows Pororo and Crong to come in, and Crong opens the door.
8 Generated [Dia]	hi .	[Desc]: Petty and Loopy think that the doll looks delicious.
8 Right Ans [Dia]	hi .	[Desc]: Pororo and Crong say hi to Eddy.
9 Generated [Dia]	eddy what are you doing .	[Desc]: Eddy and Pororo are surprised, and Eddy is little bit unpleasant.
9 Right Ans [Dia]	eddy what are you doing .	[Desc]: Behind Eddy, there is a car, and Pororo asks him a question.
10 Generated [Dia]	oh i am making a new toy .	[Desc]: Eddy touches Eddy's head and explains that Eddy is making a new toy.
10 Right Ans [Dia]	oh i am making a new toy .	[Desc]: Eddy touches Eddy's head and explains that Eddy is making a new toy.
11 Generated [Dia]	oh i am making a new toy .	[Desc]: Eddy looks at Petty and smiles.
11 Right Ans [Dia]	toy crong .	[Desc]: Pororo and Crong look at the car and run toward the car.
12 Generated [Dia]	toy crong .	[Desc]: The sides of mailbox is colored by pink, and the word, POST, is also written on the front side of the mailbox.
12 Right Ans [Dia]	it looks great .	[Desc]: Pororo and Crong stand in front of the car, and They think the car is cool.
13 Generated [Dia]	I will take it to the playground .	[Desc]: Eddy comes over and says Eddy's plan.
13 Right Ans [Dia]	I will take it to the playground .	[Desc]: Eddy comes over and says Eddy's plan.
14 Generated [Dia]	I will take it to the playground .	[Desc]: Eddy is looking for eddy friends.
14 Right Ans [Dia]	to the playground .	[Desc]: Crong is looking around the car, and Pororo asks a question.
15 Generated [Dia]	to the playground .	[Desc]: Petty and Loopy think that the situation is strange.
15 Right Ans [Dia]	then all of us can ride in it .	[Desc]: Eddy stretches Eddy's arms, and Eddy is taking the car to the playground for everyone.
16 Generated [Dia]	I will bring something to drink something .	[Desc]: Pororo agrees with Eddy and wants to help Eddy.
16 Right Ans [Dia]	I will bring it there .	[Desc]: Pororo agrees with Eddy and wants to help Eddy.
17 Generated [Dia]	let go .	[Desc]: Pororo asks friends if Eddy did go.
17 Right Ans [Dia]	really .	[Desc]: Eddy is surprised, and Pororo comes closer to the head of the car.
18 Generated [Dia]	I will be there first .	[Desc]: Pororo and Crong are waiting for Loopy.
18 Right Ans [Dia]	I will be there first .	[Desc]: Pororo takes car to the playground, and Eddy and Crong are surprised.
19 Generated [Dia]	I will be there first .	[Desc]: Pororo goes out of Eddy's house, and Poby is surprised by Pororo.
19 Right Ans [Dia]	crong .	[Desc]: Pororo goes out of Eddy's house, and Eddy is standing still, and Crong follows Pororo.
20 Generated [Dia]	crong .	[Desc]: Pororo goes out of Eddy's house, and is approaching to Pororo.
20 Right Ans [Dia]	crong crong .	[Desc]: Pororo runs with the car, and Crong follows Pororo.
21 Generated [Dia]	we are safe .	[Desc]: The night arrives in the forest.
21 Right Ans [Dia]	it is hard .	[Desc]: Pororo stops at the top of the hill, and Crong stops too.
22 Generated [Dia]	it is hard .	[Desc]: Poby says that Harry is playing with the ball, and Harry is excited.
22 Right Ans [Dia]	crong .	[Desc]: Crong and Pororo are surprised, and Crong swings Crong's arms.
23 Generated [Dia]	crong .	[Desc]: Crong and Pororo are surprised, and Crong swings Crong's arms.
23 Right Ans [Dia]	hey stop there .	[Desc]: The car is going down a hill, and Pororo follows, and Crong stamps Crong's feet repeatedly.
24 Generated [Dia]	hey stop there .	[Desc]: pororo runs toward crong with angry face, and crong and pororo saile at pororo.
24 Right Ans [Dia]	crong .	[Desc]: Crong thinks the situation is weird.
25 Generated [Dia]	pororo crong .	[Desc]: Crong says it is not a luck.
25 Right Ans [Dia]	crong .	[Desc]: Eddy calls Crong, and Crong looks back.
26 Generated [Dia]	what are you doing here .	[Desc]: Eddy asks Pororo what Crong is doing.
26 Right Ans [Dia]	what are you doing here .	[Desc]: Eddy asks Crong what Crong is doing.
27 Generated [Dia]	what are you doing here .	[Desc]: Eddy asks Crong what Crong is doing.
27 Right Ans [Dia]	crong .	[Desc]: Crong points something to Eddy.
28 Generated [Dia]	it looks fun .	[Desc]: Eddy looks angry, and Pororo says something to Eddy.
28 Right Ans [Dia]	oh no pororo disappeared in a new toy car .	[Desc]: Pororo is disappearing with a new toy car, and Eddy and Crong no are surprised.
29 Generated [Dia]	oh no pororo disappeared crong .	[Desc]: Pororo is disappearing with a new toy car, and Eddy and Crong are surprised.
29 Right Ans [Dia]	lalalala .	[Desc]: Loopy is walking while singing.
30 Generated [Dia]	i will make you a sorry .	[Desc]: Eddy, Eddy and Pororo are standing in front of the door, and Pororo is waving to Eddy saying what is going to his house.
30 Right Ans [Dia]	hi loopy .	[Desc]: On a new toy car, Pororo say hi to Loopy, and Pororo is going fast.
31 Generated [Dia]	hi loopy .	[Desc]: On a new toy car, crong talks to Loopy, and Eddy is little bit surprised.
31 Right Ans [Dia]	hey pororo .	[Desc]: Loopy says hi, but Pororo is too fast.
32 Generated [Dia]	what are you doing .	[Desc]: Pororo is surprised at Pororo's house, and Pororo is also smiling.
32 Right Ans [Dia]	what are you doing .	[Desc]: Loopy is looking at Pororo, and Loopy is surprised.
33 Generated [Dia]	there you go .	[Desc]: Loopy and Eddy are amazed that Crong feels something.
33 Right Ans [Dia]	this is a new toy eddy made .	[Desc]: Pororo yells at Loopy that it is a Eddy's new car.
34 Generated [Dia]	this is a new toy eddy made .	[Desc]: Pororo yells at Loopy that it is a Eddy's new car.
34 Right Ans [Dia]	yahoo .	[Desc]: Pororo is enjoying a ride, and Pororo yells with joy.
35 Generated [Dia]	yahoo .	[Desc]: Pororo is angry and coming to crong with a sleigh.
35 Right Ans [Dia]	lalalala lalalala .	[Desc]: Loopy is walking merrily on a snowy land.
<end of Episode>		

Figure 5.10 Comparative cover map of the generated result and ground-truth. It is marked with color boxes to check how many sentences are similar. Yellow color for perfect-matched on the ground-truth, blue one for the matching case with on the 1-step shifted ones, and green one for the case that semantically same sentences or only some words are different.

Chapter 6

Concluding Remarks

6.1 Summary of Methods and Contributions

As research vision, I pursue building situation-aware AI agents. The proposed methods can be applied to situation expression as visual storytelling, situation inference as open story generation and situation inference from partially observed stories.

We propose new architectures GLAC Nets for visual storytelling, SEOE for embedding story to have the trajectory form, ViStoryNets for video story regeneration, RERMs for open-ended story generation.

Also, this dissertation proposes several technical issues as follows: Embedding story with the form of trajectories can be used for composite representation of order and semantics. The scheduled sampling technique is helpful to multi-step prediction in BiLSTM. Vision-to-language translation setting and global-local attention setting is powerful not only to learn overall structures but also to deliver information to the decoder. Cascading mechanism is useful to serial

generation of sentences. Automated metric scores should be used very carefully for the story generation tasks. Recurrent event retrieval models (RERMs) can be trained in an self-supervised manner.

6.2 Limitation and Outlook

While some part of components can be covered with the proposed methods in this dissertation, lots of parts are still remained to be developed. In particular, a study on video story learning with kids videos suffers from the small size of the dataset.

Recently, some researchers focus on some promising issues such as human-intervened interface (visual dialog¹ (Das et al., 2017) and visual object discovery via dialogue (guessWhat)² (de Vries et al., 2017)), learning via navigating in the environment (Room-to-Room (R2) navigation³ (Anderson et al., 2018), and embodied QA⁴ (Das et al., 2018)). Those works can make good synergy effect with the techniques of visual-linguistic story understanding and generation.

6.3 Suggestions for Future Research

In the our visionary scenario introduced in Chapter 1, we can see the further direction to do more. The first promising topic is to build generalized situation representation of image sequences. Even though the size of VIST dataset is good, it is not enough to be able to transfer to other task except for event-like picture streams. It needs to gather more data to be used for analyzing everyday lives freely, it is not tested to be potentiality yet.

The second topic is to gear neural conversational models newly developed

¹<https://visualdialog.org/>

²<https://guesswhat.ai/>

³<http://bringmeaspoon.org/>

⁴<https://embodiedqa.org/>

recently. Currently, the proposed system is based on story generation without interaction with humans.

The last suggestion is to build capable to take **the Strong Story Hypothesis** and **the Directed Perception Hypothesis** proposed by Winston (2011).

The Strong Story Hypothesis: The mechanisms that enable humans to tell, understand, and recombine stories separate human intelligence from that of other primates.

The Directed Perception Hypothesis: The mechanisms that enable humans to direct the resources of their perceptual systems to answer questions about real and imagined events account for much of commonsense knowledge.

As we mentioned in related work, the ability of story understanding is an innate function, which makes humans unique. Important point of those hypothesis is the capability to recombine stories as if two words are merged (Chomsky, 2010).

Since I believe that human-level AI should have the ability of story understanding, it needs to do research on the system to combine 'video story learning', 'story generation', and 'recombining stories' as concept blending.

Bibliography

- Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., and Vijayanarasimhan, S. (2016). Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*.
- Agarwal, S., Wills, J., Cayton, L., Lanckriet, G., Kriegman, D., and Belongie, S. (2007). Generalized non-metric multidimensional scaling. In *Artificial Intelligence and Statistics*, pages 11–18.
- Ailon, N. (2012). An active learning algorithm for ranking from pairwise preferences with an almost optimal query complexity. *Journal of Machine Learning Research*, 13(Jan):137–164.
- Akata, Z., Perronnin, F., Harchaoui, Z., and Schmid, C. (2016). Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1425–1438.
- Amit, Y., Fink, M., Srebro, N., and Ullman, S. (2007). Uncovering shared structures in multiclass classification. In *Proceedings of the 24th international conference on Machine learning*, pages 17–24. ACM.
- Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., and van den Hengel, A. (2018). Vision-and-language navigation:

- Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.
- Arias-Castro, E. et al. (2017). Some theory for ordinal embedding. *Bernoulli*, 23(3):1663–1693.
- Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.
- Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.
- Belz, A., Berg, T. L., and Yu, L. (2018). From image to language and back again. *Journal of Natural Language Engineering*, 24(3):325–362.
- Bengio, S., Vinyals, O., Jaitly, N., and Shazeer, N. (2015). Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179.
- Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.
- Borg, I. and Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

- Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R., and Bengio, S. (2016). Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21.
- Breazeal, C. (2004). Social interactions in hri: the robot view. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(2):181–186.
- BRUNER, J. S. (1986). *Actual Minds, Possible Worlds*, volume 1. Harvard University Press.
- Caba Heilbron, F., Escorcia, V., Ghanem, B., and Carlos Nibbles, J. (2015). Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970.
- Camacho-Collados, J. and Pilehvar, T. (2018). From word to sense embeddings: A survey on vector representations of meaning. *arXiv preprint arXiv:1805.04032*.
- Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4724–4733. IEEE.
- Cayton, L. (2005). Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, 12(1-17):1.
- Chambers, N. and Jurafsky, D. (2008). Unsupervised learning of narrative event chains. *Proceedings of ACL-08: HLT*, pages 789–797.
- Changpinyo, S., Chao, W.-L., Gong, B., and Sha, F. (2016). Synthesized

- classifiers for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5327–5336.
- Chaturvedi, S., Peng, H., and Roth, D. (2017). Story comprehension for predicting what happens next. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1603–1614.
- Chen, Y.-N., Celiyilmaz, A., and Hakkani-Tür, D. (2017). Deep learning for dialogue systems. *Proceedings of ACL 2017, Tutorial Abstracts*, pages 8–14.
- Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111.
- Chomsky, N. (2010). Some simple evo devo theses: How true might they be for language. *The evolution of human language*, pages 45–62.
- Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.
- Dai, B., He, N., Pan, Y., Boots, B., and Song, L. (2017). Learning from conditional distributions via dual embeddings. In *Artificial Intelligence and Statistics*, pages 1458–1467.
- Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., and Batra, D. (2018). Embodied Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M., Parikh, D.,

- and Batra, D. (2017). Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2.
- Dawes, R. M. (1999). A message from psychologists to economists: mere predictability doesn’t matter like it should (without a good story appended to it). *Journal of Economic Behavior & Organization*, 39(1):29–40.
- de Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., and Courville, A. C. (2017). Guesswhat?! visual object discovery through multi-modal dialogue. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Deepak, P., Garg, D., and Shevade, S. (2017). Latent space embedding for retrieval in question-answer archives. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 855–865.
- Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of CVPR*, pages 2625–2634.
- Finlayson, M. M. A. (2012). *Learning narrative structure from annotated folktales*. PhD thesis, Massachusetts Institute of Technology.
- Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al. (2013). Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129.
- Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., and Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468.

- Gao, Y., Beijbom, O., Zhang, N., and Darrell, T. (2016). Compact bilinear pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 317–326.
- Geirhos, R., Janssen, D. H. J., Schütt, H. H., Rauber, J., Bethge, M., and Wichmann, F. A. (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. *CoRR*, abs/1706.06969.
- Gervás, P., Díaz-Agudo, B., Peinado, F., and Hervás, R. (2005). Story plot generation based on cbr. *Knowledge-Based Systems*, 18(4-5):235–242.
- Goyal, P. and Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 151:78–94.
- Graves, A. (2013). Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.
- Ha, J., Kim, K.-M., and Zhang, B.-T. (2015). Automated construction of visual-linguistic knowledge via concept learning from cartoon videos. In *AAAI*, pages 522–528.
- Härdle, W. and Simar, L. (2015). *Applied Multivariate Statistical Analysis*. Springer Berlin Heidelberg.
- He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.
- Heo, M.-O., Han, D.-S., Kim, K.-M., and Zhang, B.-T. (2015a). A deep learning-based story learning framework for kids videos. *KIISE Winter Conference*, pages 725–727.

- Heo, M.-O., Jo, S.-H., Lee, S.-W., and Zhang, B.-T. (2015b). Learning sparse higher-order markov random fields for human activity analysis. *Proceedings of KIIS Spring Conference*, 25(1):62–63.
- Heo, M.-O., Kang, M., and Zhang, B.-T. (2010). Visual query expansion via incremental hypernetwork models of image and text. In *Pacific Rim International Conference on Artificial Intelligence*, pages 88–99. Springer.
- Heo, M.-O., Kang, M.-G., Lim, B.-K., Hwang, K.-B., Park, Y.-T., and Zhang, B.-T. (2012). Real-time route inference and learning for smartphone users using probabilistic graphical models. *Journal of KIISE: Software and Applications*, 39(6):425–435.
- Heo, M.-O., Kim, K.-M., and Zhang, B.-T. (2016). Deep learning-based techniques for learning video stories. *Journal of Korean Multimedia Society*, 20:23–40.
- Heo, M.-O., Kim, K.-M., and Zhang, B.-T. (2018). Vistorynet: Neural networks with successive event order embedding and bilstms for video story regeneration. *KIISE Transactions on Computing Practices*, 24(3):138–144.
- Heo, M.-O., Lee, S.-W., Lee, J., and Zhang, B.-T. (2013). Learning global-to-local discrete components with nonparametric bayesian feature construction. In *NIPS workshop on Constructive Machine Learning*.
- Heo, M.-O. and Zhang, B.-T. (2016). Character facial sentiment learning methods based on r-cnns for kids animations. *Korea Computer Congress 2016*, pages 909–911.
- Herman, D. (2013). *Storytelling and the Sciences of Mind*. MIT press.

- Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.
- Huang, T.-H. K., Ferraro, F., and Misra, I., editors (2018a). *Proceedings of 1st Workshop on Storytelling (New Orleans, Louisiana, June 2018)*, New Orleans, Louisiana, USA. Association for Computational Linguistics (ACL).
- Huang, T.-H. K., Ferraro, F., and Misra, I. (2018b). Visual storytelling challenge. <http://www.visionandlanguage.net/workshop2018/>. Accessed: 2018-05-30.
- Huang, T.-H. K., Ferraro, F., Mostafazadeh, N., Misra, I., Devlin, J., Agrawal, A., Girshick, R., He, X., Kohli, P., Batra, D., et al. (2016). Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*.
- Jamieson, K. G. and Nowak, R. D. (2011). Low-dimensional embedding using adaptively selected ordinal data. In *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, pages 1077–1084. IEEE.
- Jang, Y., Song, Y., Yu, Y., Kim, Y., and Kim, G. (2017). Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). Honolulu, Hawaii*, pages 2680–8.

- Johnson, J., Karpathy, A., and Fei-Fei, L. (2016a). Denscap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2016b). Google’s multilingual neural machine translation system: Enabling zero-shot translation. Technical report, Google.
- Kapadia, M., Poulakos, S., Gross, M., and Sumner, R. W. (2017). Computational narrative. In *ACM SIGGRAPH 2017 Courses*, page 4. ACM.
- Karpathy, A. and Fei-Fei, L. (2017). Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676.
- Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *CVPR*.
- Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.
- Kilickaya, M., Erdem, A., Ikizler-Cinbis, N., and Erdem, E. (2017). Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 199–209.
- Kim, J.-H., Jun, J., and Zhang, B.-T. (2018). Bilinear attention networks. *arXiv preprint arXiv:1805.07932*.

- Kim, J.-H., Lee, S.-W., Kwak, D., Heo, M.-O., Kim, J., Ha, J.-W., and Zhang, B.-T. (2016). Multimodal residual learning for visual qa. In *Advances in Neural Information Processing Systems*, pages 361–369.
- Kim, K.-M., Heo, M.-O., Choi, S.-H., and Zhang, B.-T. (2017). Deepstory: video story qa by deep embedded memory networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2016–2022.
- Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.
- Kiros, R., Salakhutdinov, R., and Zemel, R. (2014). Multimodal neural language models. In *International Conference on Machine Learning*, pages 595–603.
- Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.
- Kleindessner, M. and Luxburg, U. (2014). Uniqueness of ordinal embedding. In *Conference on Learning Theory*, pages 40–67.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS’12*, pages 1097–1105, USA. Curran Associates Inc.
- Kybartas, B. and Bidarra, R. (2017). A survey on story generation techniques for authoring computational narratives. *IEEE Transactions on Computational Intelligence and AI in Games*, 9(3):239–253.
- Lebowitz, M. (1985). Story-telling as planning and learning. *Poetics*, 14(6):483–502.

- Lei, J., Yu, L., Bansal, M., and Berg, T. L. (2018). Tvqa: Localized, compositional video question answering. In *EMNLP*.
- Li, B. and Riedl, M. O. (2015). Scheherazade: crowd-powered interactive narrative generation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 4305–4306. AAAI Press.
- Li, X., Vilnis, L., and McCallum, A. (2017). Improved representation learning for predicting commonsense ontologies. *arXiv preprint arXiv:1708.00549*.
- Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.
- Lin, Z., Feng, M., Santos, C. N. d., Yu, M., Xiang, B., Zhou, B., and Bengio, Y. (2017). A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.
- Liu, Y., Fu, J., Mei, T., and Chen, C. W. (2017). Let your photos talk: Generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks. In *AAAI Conference on Artificial Intelligence*.
- Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of EMNLP*, pages 1412–1421.
- Ma, Y. and Fu, Y. (2011). *Manifold Learning Theory and Applications*. CRC Press.
- Mani, I. (2013). *Computational Modeling of Narrative*, volume 18. Morgan & Claypool Publishers.

- Martin, L. J., Ammanabrolu, P., Hancock, W., Singh, S., Harrison, B., and Riedl, M. O. (2017). Event representations for automated story generation with deep neural nets. *arXiv preprint arXiv:1706.01331*.
- Mateas, M. and Sengers, P. (2003). *Narrative intelligence*. J. Benjamins Pub.
- McCann, B., Bradbury, J., Xiong, C., and Socher, R. (2017). Learned in translation: Contextualized word vectors. In *NIPS*, pages 6294–6305.
- McIntyre, N. and Lapata, M. (2009). Learning to tell tales: A data-driven approach to story generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 217–225. Association for Computational Linguistics.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P., and Allen, J. (2016). A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of NAACL-HLT*, pages 839–849.
- Mostafazadeh, N., Roth, M., Louis, A., Chambers, N., and Allen, J. (2017). Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51.
- Mun, J., Seo, P. H., Jung, I., and Han, B. (2017). Marioqa: Answering questions by watching gameplay videos. In *ICCV*, pages 2886–2894.

- Na, S., Lee, S., Kim, J., and Kim, G. (2017). A read-write memory network for movie story understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 677–685.
- O’Neill, B. and Riedl, M. (2014). Dramatis: A computational model of suspense. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
- Park, C. C. and Kim, G. (2015). Expressing an image stream with a sequence of natural sentences. In *NIPS*, pages 73–81.
- Perone, C. S., Silveira, R., and Paula, T. S. (2018). Evaluation of sentence embeddings in downstream and linguistic probing tasks. *arXiv preprint arXiv:1806.06259*.
- Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of NAACL-HLT*, volume 1, pages 2227–2237.
- Pichotta, K. and Mooney, R. J. (2016). Learning statistical scripts with lstm recurrent neural networks. In *30th AAAI Conference on Artificial Intelligence*.
- Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250.
- Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. (2016). Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069.

- Riedl, M. O. (2016). Computational narrative intelligence: A human-centered goal for artificial intelligence. In *Proceedings of CHI’16 Workshop on Human-Centered Machine Learning*.
- Riedl, M. O. and Young, R. M. (2010). Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research*, 39:217–268.
- Rohrbach, A., Rohrbach, M., Tandon, N., and Schiele, B. (2015). A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212.
- Schank, R. C. (1990). *Tell me a story: A new look at real and artificial memory*. Charles Scribner’s Sons.
- Shaw, B. and Jebara, T. (2009). Structure preserving embedding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 937–944. ACM.
- Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *ICLR*.
- Socher, R., Ganjoo, M., Manning, C. D., and Ng, A. (2013). Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943.
- Sukhbaatar, S., Weston, J., Fergus, R., et al. (2015). End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448.
- Sutskever, I., Vinyals, O., and Le, Q. V. (2014a). Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

- Sutskever, I., Vinyals, O., and Le, Q. V. (2014b). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.
- Swanson, R. and Gordon, A. S. (2012). Say anything: Using textual case-based reasoning to enable open-domain interactive storytelling. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):16.
- Tapaswi, M., Zhu, Y., Stiefelhausen, R., Torralba, A., Urtasun, R., and Fidler, S. (2016a). Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4631–4640.
- Tapaswi, M., Zhu, Y., Stiefelhausen, R., Torralba, A., Urtasun, R., and Fidler, S. (2016b). MovieQA: Understanding Stories in Movies through Question-Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Terada, Y. and Luxburg, U. (2014). Local ordinal embedding. In *International Conference on Machine Learning*, pages 847–855.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *NIPS*, pages 5998–6008.
- Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.
- Vendrov, I., Kiros, R., Fidler, S., and Urtasun, R. (2016). Order-embeddings of

- images and language. In *International Conference on Learning Representations*.
- Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE.
- Vondrick, C., Pirsiavash, H., and Torralba, A. (2016). Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, pages 613–621.
- Wang, B., Liu, K., and Zhao, J. (2017). Conditional generative adversarial networks for commonsense machine comprehension. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4123–4129.
- Wang, D. and Nyberg, E. (2015). A long short-term memory model for answer sentence selection in question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 707–712.
- Wang, J., Fu, J., Tang, J., Li, Z., and Mei, T. (2018a). Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. In *AAAI Conference on Artificial Intelligence*.
- Wang, X., Chen, W., Wang, Y.-F., and Wang, W. Y. (2018b). No metrics are perfect: Adversarial reward learning for visual storytelling. In *The annual meeting of the Association for Computational Linguistics (ACL)*.
- Wauthier, F., Jordan, M., and Jojic, N. (2013). Efficient ranking from pairwise

- comparisons. In *International Conference on Machine Learning*, pages 109–117.
- Wehrmann, J., Mattjie, A., and Barros, R. C. (2018). Order embeddings and character-level convolutions for multimodal alignment. *Pattern Recognition Letters*, 102:15–22.
- Weinberger, K. Q. and Chapelle, O. (2009). Large margin taxonomy embedding for document categorization. In *Advances in Neural Information Processing Systems*, pages 1737–1744.
- Weston, J., Bordes, A., Chopra, S., and Mikolov, T. (2015). Towards ai-complete question answering: A set of prerequisite toy tasks. *CoRR*, abs/1502.05698.
- Winston, P. H. (2011). The strong story hypothesis and the directed perception hypothesis. In *AAAI Fall Symposium: Advances in Cognitive Systems*.
- Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., and van den Hengel, A. (2017). Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40.
- Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., and Schiele, B. (2016). Latent embeddings for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 69–77.
- Xu, J., Mei, T., Yao, T., and 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*, pages 5288–5296.
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In Bach, F. and Blei, D., editors, *Proceedings of the*

32nd International Conference on Machine Learning, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France. PMLR.

Yu, L., Bansal, M., and Berg, T. (2017). Hierarchically-attentive rnn for album summarization and storytelling. In *Proceedings of EMNLP*, pages 966–971.

Yu, L., Hermann, K. M., Blunsom, P., and Pulman, S. (2014). Deep learning for answer sentence selection. *arXiv preprint arXiv:1412.1632*.

Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., and Metaxas, D. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE Int. Conf. Comput. Vision (ICCV)*, pages 5907–5915.

Zhang, Z. and Saligrama, V. (2016). Zero-shot learning via joint latent similarity embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6034–6042.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

초록

스토리를 이해하는 능력은 동물들 뿐만 아니라 다른 유인원과 인류를 구별짓는 중요한 능력이다. 인공지능이 일상생활 속에서 사람들과 함께 지내면서 그들의 생활 속 맥락을 이해하기 위해서는 스토리를 이해하는 능력이 매우 중요하다. 하지만, 기존의 스토리에 관한 연구는 언어처리의 어려움으로 인해 사전에 정의된 세계 모델 하에서 좋은 품질의 저작물을 생성하려는 기술이 주로 연구되어 왔다. 기계학습 기법을 통해 스토리를 다루려는 시도들은 대체로 자연어로 표현된 데이터에 기반할 수 밖에 없어 자연어 처리에서 겪는 문제들을 동일하게 겪는다. 이를 극복하기 위해서는 시각적 정보가 함께 연동된 데이터가 도움이 될 수 있다. 최근 딥러닝의 눈부신 발전에 힘입어 시각과 언어 사이의 관계를 다루는 연구들이 늘어나고 있다. 연구의 비전으로서, 인공지능 에이전트가 주변 정보를 카메라로 입력받는 환경 속에 놓여있는 상황을 생각해 볼 수 있다. 이 안에서 인공지능 에이전트는 주변을 관찰하면서 그에 대한 스토리를 자연어 형태로 생성하고, 생성된 스토리를 바탕으로 다음에 일어날 스토리를 한 단계에서 여러 단계까지 예측할 수 있다. 본 학위 논문에서는 사진 및 비디오 속에 나타나는 스토리(visual story)를 학습하는 방법, 내러티브 텍스트로의 변환, 가려진 사건 및 다음 사건을 추론하는 연구들을 다룬다.

첫 번째로, 여러 장의 사진이 주어졌을 때 이를 바탕으로 스토리 텍스트를 생성하는 문제(비주얼 스토리텔링)를 다룬다. 이 문제 해결을 위해 글랙넷(GLAC Net)을 제안하였다. 먼저, 사진들로부터 정보를 추출하기 위한 컨볼루션 신경망, 문장을 생성하기 위해 순환신경망을 이용한다. 시퀀스-시퀀스 구조의 인코더로서, 전체적인 이야기 구조의 표현을 위해 다계층 양방향 순환신경망을 배치하되 각 사진별 정보를 함께 이용하기 위해 전역적-국부적 주의집중 모델을 제안하였다. 또한, 여러 문장을 생성하는 동안 맥락정보와 국부정보를 잃지 않게 하기 위해 앞선 문장

정보를 전달하는 메커니즘을 제안하였다. 위 제안 방법으로 비스트(VIST) 데이터 집합을 학습하였고, 제 1 회 시각적 스토리텔링 대회(visual storytelling challenge)에서 사람 평가를 기준으로 전체 점수 및 6 항목 별로 모두 최고점을 받았다.

두 번째로, 스토리의 일부가 문장들로 주어졌을 때 이를 바탕으로 다음 문장을 예측하는 문제를 다룬다. 임의의 길이의 스토리에 대해 임의의 위치에서 예측이 가능해야 하고, 예측하려는 단계 수에 무관하게 작동해야 한다. 이를 위한 방법으로 순환 사건 인출 모델(Recurrent Event Retrieval Models)을 제안하였다. 이 방법은 은닉 공간 상에서 현재까지 누적된 맥락과 다음에 발생할 유력 사건 사이의 거리를 가깝게 하도록 맥락누적함수와 두 개의 임베딩 함수를 학습한다. 이를 통해 이미 입력되어 있던 스토리에 새로운 사건이 입력되면 쌍선형적 연산을 통해 기존의 맥락을 개선하여 다음에 발생할 유력한 사건들을 찾는다. 이 방법으로 락스토리(ROCStories) 데이터집합을 학습하였고, 스토리 클로즈 테스트(Story Cloze Test)를 통해 평가한 결과 경쟁력 있는 성능을 보였으며, 특히 임의의 길이로 추론할 수 있는 기법 중에 최고성능을 보였다.

세 번째로, 비디오 스토리에서 사건 시퀀스 중 일부가 가려졌을 때 이를 복구하는 문제를 다룬다. 특히, 각 사건의 의미 정보와 순서를 모델의 표현 학습에 반영하고자 하였다. 이를 위해 은닉 공간 상에 각 에피소드들을 꺾적 형태로 임베딩하고, 이를 바탕으로 스토리를 재생성을 하여 스토리 완성을 할 수 있는 모델인 비스토리넷(ViStoryNet)을 제안하였다. 각 에피소드를 꺾적 형태를 가지게 하기 위해 사건 문장을 사고벡터(thought vector)로 변환하고, 연속 이벤트 순서 임베딩을 통해 전후 사건들이 서로 가깝게 임베딩되도록 하여 하나의 에피소드가 꺾적의 모양을 가지도록 학습하였다. 뽀로로QA 데이터집합을 통해 실험적으로 결과를 확인하였다. 임베딩 된 에피소드들은 꺾적 형태로 잘 나타났으며, 에피소드들을 재생성 해본 결과 전체적인 측면에서 유사한 결과를 보였다.

위 결과물들은 카메라로 입력되는 주변 정보를 바탕으로 스토리를 이해하고 일부 관측되지 않은 부분을 추론하며, 향후 스토리를 예측하는 방법들에 대응된다.

주요어: 시각적 스토리텔링, 서사 텍스트 생성, 다음 사건 예측, 스토리 완성, 전역-

국부 주의집중, 순환 사건 인출 모델, 연속 이벤트 순서 임베딩

학번: 2005-21534