Seminar: Generating Narrative Paragraph for Photo Stream

Chen Zongqi
Matriculation Number 1564832
University of Mannheim, Germany
zchen@mail.uni-mannheim.de

Abstract

When machine learning techniques develop rapidly in image captioning, other extension relevant areas appeal to many researchers doing further study. Among them generating story from sequential photo stream becomes quite interesting and challenging task. Compared with single image captioning, this task needs to handle with two mainly problems facing to researchers, large visual variance in sequence and long-term language coherence among multiple sentences. Till now those two critical questions are solved by different approaches. In this paper, we mainly focus on one of them, generating story from sequential photos stream via Bidirectional Attention Recurrent Neural Network method, and discus several relevant approaches from first paper to state-of-the-art.

1 Introduction

With computing ability rapidly developing and new theory proposing, there are more and more intersection researches between computer vision area and natural language processing area occurring. This paper will focus on one specified topic from them, which is generating narrative paragraph for photo stream.

The basic task for this topic is that given sequential photos as input and return a story to describe as output. The core problems which need to be solved are how to generate corresponding sentence and how to organize sentences as human-level story. The previous question is based on the related field Image Captioning, such as generator with deep recurrent structure (Vinyals et al., 2014), deep visual-semantic alignments (Karpathy and Fei-Fei, 2017). However the different between this topic with image captioning is that image captioning mainly focus on individual image and this topic needs to handle with sequential photos. It would be more challenged that computer not only needs to understand the content of images but also need to find their inner relation.

This paper would introduce three approaches, Bidirectional Attention Recurrent Neural Network (Liu et al., 2017), Coherence Recurrent Convolutional Network (Park and Kim, 2015) and Generating Narrative Paragraph by Adversarial Training (Wang et al., 2018). Meanwhile we want to introduce some basic terminologies which could be beneficial for understanding in background part, such as basic network structures and some language evaluation methods. Furthermore this paper would introduce main-stream datasets special for this field including some information about scale, content and generation strategy. Finally, we would compare their difference and common characteristics and discuss which parts could improve our evaluation and how could they work in last two parts Evaluation and Discussion.

2 Background

In machine learning and natural language processing area, there are many high frequency terminologies appearing in almost every relevant paper and article. Here we list some significant terms mentioned in this paper having briefly explanation including different networks and retrieval methods.

Convolutional Neural Network CNN is first implemented for application in 1998 by Yann LeCun (LeCun et al., 1998) which applied for digit recognition experiment. CNN is developed from Neural Network (Hornik et al., 1989) which includes input layer, hidden layer and output layer. The basic CNN hidden layers contains convolu-

tional layers, pooling layers, fully connected layers and normalization layers. The reason why CNN could develop so rapidly is that researchers start to use GPU to improve calculation efficiency (Steinkraus et al., 2005). Later some famous CNN networks are invited, such as AlexNet (Krizhevsky et al., 2012), GoogleNet (Szegedy et al., 2015), VGGNet (Simonyan and Zisserman, 2014) and ResNet (He et al., 2016).

Recurrent Neural Network RNN is another neural network variant, which connects each unit in a cycle way. One of the successful network Long short-term memory is invited to use as a block for developing a larger RNN network (Hochreiter and Schmidhuber, 1997). The basic components include cell, input gate, output gate and forget gate. In particularly, the concept of cell is very important for LSTM, which means memory. Due to RNN has many variants, here we list two of them such as bidirectional recurrent neural networks (Schuster and Paliwal, 1997) and deep bidirectional recurrent neural networks (Graves et al., 2013). RNN has huge success in natural language processing because of its special recurrent component.

Generative Adversarial Networks GANs are quite special neural network variant which contains game theory to train two neural networks (Goodfellow et al., 2014). The basic idea is that two models, generative model an discriminative model, play a two-player game and leaning from each game result. GANs become more and more popular in deep learning area based on its powerful ability in handling with complex tasks.

BLEU Bilingual evaluation understudy (BLEU) is quick, inexpensive, and language-independent, that correlates highly with human evaluation, and that has little marginal cost per run (Papineni et al., 2002). BLEU is a popular evaluation score for machine translation quality since 2002. BLEU has approximate human judgement at a corpus-level, whereas it has bad performance sometimes on sentence-level. See formula 1 below :

$$\begin{aligned} Pn &= \frac{\sum\sum Countclip(n-gram)}{\sum\sum Count(n-gram)} \\ BLEU &= BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \end{aligned} \tag{1}$$

where Pn is the whole corpus modified precision score, N is the length, w_n is positive weights adding up to 1 and BP is the brevity penalty function 2 see below:

$$BP = \begin{cases} 1, & \text{if} c > r \\ e^{(1-r/c)}, & \text{otherwise} \end{cases} \tag{2}$$

where r is test corpus effective reference length and c is the total length of the candidate translation corpus.

METEOR Metric for Evaluation of Translation with Explicit ORdering (METEOR) is based on a generalized concept of uni-gram matching between the machine produced translation and human-produced reference translations (Banerjee and Lavie, 2005). Compared with BLEU, METEOR consider both precision and recall value on whole corpus. And METEOR has better performance in both corpus-level and sentence-level. Here is the basic formula 3:

$$\begin{aligned} Fmean &= \frac{10PR}{R+9P} \\ Pen &= \gamma\left(\frac{ch}{m}\right)^{\theta} \\ METEOR &= (1-Pen)\cdot Fmean, \end{aligned} \tag{3}$$

where P is precision value, R is recall value, ch is the number of chunks, γ and θ are tuned to maximize correlation with human judgements.

CIDEr Consensus-based Image Description Evaluation (CIDEr) a novel paradigm for evaluating image descriptions that uses human consensus (Vedantam et al., 2015). CIDEr treats each sentence as document using its tf-idf vector and compute caption with generative caption by cosine similarity. See equation 4 below:

$$g(s_{ij}) = \frac{h_k(s_{ij})}{\sum h_l(s_{ij})} \log\left(\frac{|I|}{\sum \min(1, \sum_q h_k(s_{pq}))}\right), \tag{4}$$

where $h_k(s_{ij})$ means the number of times an n-gram w_k occurs in a reference sentence s_{ij}, $|I|$ is the set of all images in the dataset and $g(s_{ij})$ is tf-idf weighting for each n-gram w_k. From equation 4 we can find that the left part is actually to compute term frequency and right part $\log$ is to calculate idf. Finally, we can use g as tf-idf weight to compute generative sentence and reference sentence cosine similarity.

If candidate sentence and reference sentence are more similar in vector space then their cosine sim-

ilarity value is larger. See the formula below:

$$CIDEr_n(c_i, S_i) = \frac{1}{m}\sum_j \frac{g^n(c_i)\cdot g^n(s_{ij})}{|g^n(c_i)||g^n(s_{ij})|}, \quad (5)$$

where $g^n(c_i)$is a vector formed by $g(c_i)$ corresponding to all n-grams of length n and $|g^n(c_i)|$ is the magnitude of the vector $g^n(c_i)$; $g^n(s_{ij})$ as well.

Finally we combine them from n-grams of varying lengths for the following equation 6 as below:

$$CIDEr(c_i, S_i) = \sum_{n=1}^{N} w_n CIDEr_n(c_i, S_i), \quad (6)$$

where uniform weights $w_n = \frac{1}{N}$work the best.

3 Approach

In this part, we mainly introduce our topic model Bidirectional Attentional Recurrent Neural Networks, and then briefly explain two related but important models, Coherence Recurrent Convolutional Network (Park and Kim, 2015) and Generating Narrative Paragraph by Adversarial Training (Wang et al., 2018). Particularly the paper which proposes Coherence Recurrent Convolutional Network is the first paper in this field and the paper which proposes using Adversarial Training is the newest paper till now.

3.1 Bidirectional Attention Recurrent Neural Network

In order to generate narrative paragraph from photo stream, we need to not only learn from photos but also analyse from sentences. Based on this idea, the approach Bidirectional Attention Recurrent Neural Network is separated by two main parts, one is aimed to combine sentence embedding space with image embedding space called Semantic Space Embedding and another one is aimed to predict sentence embedding features using Bidirectional Attention Recurrent Neural Network. In general, photo streams and corresponding sentences as input would be manipulated by CNN and Word2Vecs. For image stream, each image would be transformed into 4096-dimension VGG features (Simonyan and Zisserman, 2014)and then mapped into 300-dimension image embedding space. For sentence, each sentence would be transformed by Word2Vecs (Mikolov et al., 2013)into 300-dimension sentence semantic embedding space. There is a very important assumption that 300-dimension image embedding space and 300-dimension sentence semantic embedding space share one embedding space, which means if the contents of image and sentence are similar then their euclidean distance is close.

Joint Embedding for Semantic Space To create this semantic space embedding, we could use a loss function and then use learning algorithm to approach the minimum loss value. Here we use a contrastive loss function equation7:

$$\begin{aligned} C^{remb}(x, v) = &\sum_{x\in X, v\in V, v'\in V'} max(0, \alpha - xv + xv') \\ &+ \sum_{x\in X, x'\in X', v\in V} max(0, \alpha - xv + x'v), \end{aligned} \quad (7)$$

where X is the image embedding vector and V is the sentence embedding vector, X' and V' are negative paired image and sentence sample, α is entropy for judging positive image-sentence pair similarity. Contrastive loss function (Chopra et al., 2005) is aimed to make gradient of loss function successfully approach to minimum value. The principle of contrastive is that we want to introduce a negative pair to confirm that the result we get is what we expect but not stochastic error. For example we get a result from equation and I want to know this equation that is correctly. So we input a wrong number into the equation, if we get correct result then we can confirm this equation does not work.

Bidirectional Attention Recurrent Neural Network for Textual Story Generation The goal of this part is to use Bidirectional RNN with Attention modelling to predict sentence embedding features. The input of BARNN model is image and sentence embedding vectors. The output h is sentence embedding features with image stream inputting.

The idea of Attention modelling (Itti et al., 1998) is that when human focus on one object, they would ignore something non-relevant,so that we can introduce new attention weight to balance what should be concentrated on or what should be ignored. Meanwhile, under the assumption that same content has close distance in embedding space, we introduce equation 8:

$$R_{pt} = x_p x_t, \quad (8)$$

where x_p and x_t are p and t time inputting images, and R_{pt} is the relation of image x_p and image x_t.

We use R_{pt} as our attention weight to focus on objects that are more important, then we add them more weight.

Later we have a new designed Gated Recurrent Unit with skip gate we called Skip-GRU. The classic GRU(Chung et al., 2014) has update gate and reset gate. In order to introduce our attention modelling, we add a skip gate. The advantage of skip-gate is that we can add attention weight to current hidden state and furthermore it can influence the main object features in our semantic embedding space. See our Skip-GRU equations 9:

$$\begin{aligned} z_t &= \sigma(W_{zx}x_t + W_{zh}h_{t-1}) \\ r_t &= \sigma(W_{rx}x_t + W_{rh}h_{t-1}) \\ s_t &= \sigma(W_{sx}x_t + W_{sh}h_p) \\ \tilde{h} &= \tanh(W_{hx}x_t + W_{hh}rt \odot h_{t-1} \\ &\quad + \sum_{p<t} R_{pt} \cdot W_{hp}s_t \odot h_p) \\ h_t &= z_t\tilde{h} + (1 - z_t h_{t-1}), \end{aligned} \tag{9}$$

where t and p are times, x_t and x_p are t time and p time input image, z_t, r_t and s_t are t time update gate, reset gate and skip gate, $\tilde{h}$ and h_t are current hidden state and t time output, $\odot$ means element-wise multiplication and $\tanh$ means hyper tangent function.

For Bidirectional Framework, we apply our new skip-GRU into our framework in both forward and backward pass. See equation 10below:

$$\begin{aligned} (z_t^f, r_t^f, s_t^f, \tilde{h}^f, h_t^f) &= sGRU(x_t, h_{t-1}^f, R, h_p^f; W^f) \\ (z_t^b, r_t^b, s_t^b, \tilde{h}^b, h_t^b) &= sGRU(x_t, h_{t-1}^b, R^T, h_p^b; W^b) \\ h_t &= W_h^f h_t^f + W_h^b h_t^b, \end{aligned} \tag{10}$$

where f means forward pass and b means backward pass. We learn $W = (W^f, W^b, W_h^f, W_h^b)$ as our parameters to learn. And h is what we expect that predict sentence embedding features.

The contrastive loss function is quite same with equation 7 see equation 11 below:

$$\begin{aligned} C^{cpt}(h, v) = &\sum_{v' \in V'} max(0, \gamma - hv + hv') \\ &+ \sum_{h' \in H'} max(0, \gamma - hv + h'v), \end{aligned} \tag{11}$$

where h' and v' are negative image-sentence pairs, and γ is contrastive margin. By using equation 11 we can learn parameter W combined with corresponding image captioning.

Finally, we combine two contrastive loss functions, joint embedding semantic space and bidirectional attention RNN, as one equation 12 see below:

$$C = \sum_{X,V} C^{emb}(x, v) + \sum_{H,V} C^{cpt}(h, v), \tag{12}$$

where X is sequential image embedding vectors, V is corresponding sentence embedding vectors and H is predict corresponding sentence vectors from BARNN model.

3.2 Coherence Recurrent Convolutional Network

For the first paper in this field, the model Coherence Recurrent Convolutional Network has a simple idea to implement the task that generates corresponding sentences from image stream. They divide CRCN model in three parts, first is Bidirectional Recurrent Neural Network (BRNN) which learns image captioning crawling from user blog, second is Convolutional Neural Network (CNN) which learns image representation then transforms into 4,096 image VGG vectors and the third is local coherence model (Barzilay and Lapata, 2008) which learns sentence patterns and structures in order to combine several sentences as story or paragraph. For CNN part they directly use VGGNet (Simonyan and Zisserman, 2014), so in this approach, we mainly focus on Bidirectional Recurrent Neural Network (BRNN) and local coherence model.

Bidirectional Neural Network model The goal of BRNN is to learn corresponding sentences content and also consider previous and next sentence content as we already mention in BARNN approach. In BRNN model, there are five layers, input layer, forward layer, backward layer, output layer and ReLU activation layer. See the equation 13 below:

$$\begin{aligned} x_t^f &= f(W_i^f p_t + b_i^f) \\ x_t^b &= f(W_i^b p_t + b_i^b) \\ h_t^f &= f(x_t^f + W_f h_{t-1}^f + b_f) \\ h_t^b &= f(x_t^b + W_b h_{t-1}^b + b_b) \\ o_t &= W_o(h_t^f + h_t^b) + b_o, \end{aligned} \tag{13}$$

where p_t is 300-dimension embedding vectors from sentence as our input, x_f and x_b are forward and backward units activated input, h_f and h_b are forward and backward units output, o_t is

in time the output of ReLU layer as the content of this sentence, f is ReLU activation function $f(x) = max(0, x)$ and W and b are weights and bias term.

The local coherence model In order to understand the sentence structure and content, they introduce local coherence model. The basic idea is that they make a table which contains discourse entity as row and sentence as column. The discourse entity grid is from sequential parse trees where extract from sequential text by using Stanford core NLP library (Manning et al., 2014). And then in grammatical aspect, they separate into four categories: S denotes subject, O denotes object, X denotes others except for subject or object, and $-$ denotes absent. Finally we computer their ratio by occurrence frequency and transform them into 300-dimension vector by zero-padding.

Combination of CNN, BRNN, and Coherence Model To combine BRNN and coherence model, they use two fully connected layers (FC) (Barzilay and Lapata, 2008). The two FC layers can combine both sentence content and sentence structure then map them into 4,096-dimension vector which is same dimension as image vectors. To compute sentence vector and image vector they use this score function 14 (Karpathy and Fei-Fei, 2015) as follow:

$$S_{kl} = \sum_{t=1...N} s_t^k \cdot v_t^l + g^k \cdot v_t^l, \tag{14}$$

where S_{kl} means score between image k and sentence l, v_t^l is CNN output of stream l in time t, s_t^k means the set of words s for image k, and g^k is the set of image fragments in image k (Karpathy et al., 2014). Finally, they create a loss function 15 to learn their Coherence Recurrent Convolutional Network model as follow:

$$\begin{aligned} C(\theta) = \sum_k \langle \sum_l \max(0, 1 + S_{kl} - S_{kk}) \\ + \sum_l \max(0, 1 + S_{lk} - S_{kk}) \rangle, \end{aligned} \tag{15}$$

where S_{kk} is positive image and sentence pair, S_{kl} and S_{lk} are negative image and sentence pair. Our first approach BARNN uses this loss function but they use γ instead of 1. However both of them are inspired by (Karpathy et al., 2014).

3.3 Adversarial Training

The model generating paragraph from stream photos with adversarial training is the newest model in this field to the best of my knowledge. The basic idea for this model is that they create a hierarchical story generator with Convolutional Neural Network (CNN) and Hierarchical Recurrent Neural Network (HRNN) to generate story from sequential photos inputting, and then they build two discriminators which critic determines whether the image and generated sentence are similar and critic determines whether the generated sentence is human-level story. Finally, by adversarial training, the model would be more and more powerful.

Hierarchical Story Generator To generate story-style narrative paragraph, they use train two RNN as encoder and decoder which both are built on Gated Recurrent Units (GRUs) (Chung et al., 2014). In order to train Hierarchical Story generator, they treat the generator as agent, and they build the loss function 16 as below:

$$\begin{aligned} L(\theta) &= -\sum_{n=1}^{N}\sum_{t=1}^{T} p_\theta(y_{n,t}|x_{1:n}; y_{n,1:t-1}) R(y_n) \\ &= -\sum_{n=1}^{N} \mathbb{E}_{y_n \backsim p_\theta}[R(y_n)], \end{aligned} \tag{16}$$

where $R(x)$ is reward function which will be mentioned later, generator parameter θ defined a policy $p_\theta(y_{n,t}|x_{1:n}; y_{n,1:t-1})$, x means observed image and y represents generated sentence.

Multi-modal Discriminator This discriminator is to tell how similar between image and generated sentence is. The multi-model discriminator contains fusion mechanism is inspired by (Antol et al., 2015) and a fully connected layer followed by a softmax layer, the equation 17 see below:

$$\begin{aligned} v_{x_n} &= W_x \cdot x_n + b_x \\ v_{y_n} &= W_y \cdot LSTM_\eta(y_n) + b_y \\ f_n &= \tanh(v_{x_n}) \odot \tanh(v_{y_n}) \\ C_n^m &= softmax(W_m \cdot f_n + b_m), \end{aligned} \tag{17}$$

where x_n is image x, y_n is sentence, v_{x_n} denotes embedded image feature by linear layer, v_{y_n} denotes sentence vector by sentence inputting into Long-short Term Memory network, f_n denotes fused vector, C_n^m is our multi-model discriminator where $C_n^m(c|x_n, y_n)$ can be seen probability and $c \in paired, unpaired, generated$. For example, $C_n^m(paired|x_n, y_n)$ means what is the probability of image and generated sentence are paired.

Language-style Discriminator The other discriminator is language-style discriminator, which

is aimed to check if the generated sentence is human-level understanding sentence. See the equation 18 as follow:

$$\begin{aligned} v_p &= LSTM_\phi(\overline{p}) \\ C^s &= softmax(W_p \cdot v_p + b_p), \end{aligned} \tag{18}$$

where v_p denotes the last hidden state as the encoded paragraph vector, $\overline{p}$ denotes paragraph embedding and C^s denotes the probability $C^s(gt|y)$ which means how similar with $groundtruth(gt)$. Finally, they create the reward function 19 defined below:

$$R(y_n|\cdot) = \lambda C_n^m(paired|x_n, y_n) + (1-\lambda)C_n^s(gt|y), \tag{19}$$

where λ denotes tradeoff parameter.

4 Dataset

In this part, we briefly introduce three main stream dataset, SIND, Disney and NYC, in generating photos description story field. In particularly, Disney and NYC are special crawling blog posts with topic "Travel".

SIND is the first dataset for sequential vision to language and explore how this data may be used for the task of visual storytelling (Huang et al., 2016). It contains 48,043 stories with 210,819 unique photos. The image streams are extracted from Flickr and the text stories are written by AMT. Each story consists of 5 images and 5 corresponding sentences for a story. The dataset has been split into 38,386 (80%) stories as training set, 4,837 (10%) as test set and 4,820 (10%) as validation set.

Disney is the dataset using special web crawling strategy generating from Disneyland database (Kim et al., 2015). It contains 7,717 blog posts and 60,545 images, in which 80% for training, 10% for validation and 10% for testing.

NYC is the dataset using special web crawling strategy generating from NYC database (Kim et al., 2015). It contains 11,861 unique blog posts and 78,467 images, in which 80% are used for training, 10% for validation and 10% for testing.

5 Evaluation

Due to three approaches appearing in different time, we just exacting some results from different paper to compare and discuss. However, although they are in different time-line, their relation could be summarized like that, Coherence Recurrent Convolutional Network and Bidirectional Attention Recurrent Neural Network could be compared together because BARNN has done the experiment for both model comparison.

NYC				
Methods	**R@1**	**R@5**	**R@10**	**Medr**
Random	0.17	0.25	0.59	763
1NN	5.95	13.57	20.71	63.5
BARNN-sGRU	16.23	28.7	39.53	19
BARNN-EMB	17.27	29.42	38.97	19
BCLSTM	15.10	29.91	41.07	18
CRCN	11.67	31.19	43.57	14
BARNN	**29.37**	**45.43**	**52.10**	**8**

Table 1: Sentence retrieval evaluation on NYC (Liu et al., 2017)

Disney				
Methods	**R@1**	**R@5**	**R@10**	**Medr**
Random	0.26	1.17	1.95	332
1NN	9.18	19.05	27.21	45
BARNN-sGRU	19.97	37.48	46.04	14
BARNN-EMB	21.57	39.24	46.50	12
BCLSTM	19.77	38.92	45.20	14
CRCN	14.29	31.29	43.2	16
BARNN	**35.01**	**49.07**	**57.83**	**6**

Table 2: Sentence retrieval evaluation on Disney (Liu et al., 2017)

SIND				
Methods	**R@1**	**R@5**	**R@10**	**Medr**
Random	0.0	0.04	0.10	2753
1NN	4.8	13.00	21.07	74
BARNN-sGRU	21.39	38.72	46.96	14
BARNN-EMB	21.63	38.54	47.01	14
BCLSTM	21.47	37.30	47.39	18
CRCN	9.87	28.74	39.51	21
BARNN	**24.07**	**44.29**	**53.06**	**9**

Table 3: Sentence retrieval evaluation on SIND (Liu et al., 2017)

We can see table 1, table 2 and table 3, they are retrieval task comparisons, recall@K and median rank value, doing in (Liu et al., 2017), which BARNN-sGRU and BARNN-EMB mean BARNN without sGRU and without Embedding, BCLSTM means bidirectional connection LSTM. The main point in those three tables are CRCN and BARNN comparison in three different datasets.

The table 4 shows that the language generating quality score comparison using different methods which we already mentioned in background part.

Table 5 is user study from 30 volunteers (15 males and 15 females) with 150 photo streams (5

Method	BLUE(N)	CIDEr(N)	BLUE(D)
CRCN	26.83	30.9	28.15
BARNN	39.3	41.6	37.7

Table 4: METEOR score of our generation approach and baseline (Liu et al., 2017)

photos for each). The education background distribution is: economics (16.7%), computer science (33.3%), linguistics (6.7%), engineering (13.3%), biology (20%) and art (10%). The age distribution is: 21-25 (23.3%), 26-30 (33.3%), 31-35 (20%), 36-40 (13.3%) and 41-45 (10%). The volunteers give subjective score according to two criteria:1. Relevance (C1), the question like whether the story is relevant to the photo stream? and 2. Story-style (C2), the question like whether the story has expressive and story-style language?

Method	C1	C2
CRCN	2.05	3.95
BARNN	3.49	6.23
Adversarial Training	6.63	6.68

Table 5: User study results on real-world photo streams (Wang et al., 2018).

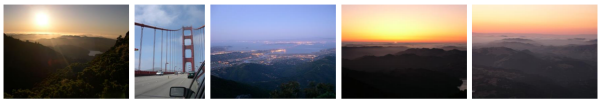
Finally, the results from three different approaches show in figure 1 and figure 2.

Ground truth: The flag corps led the parade down the streets. The mayor closed up the parade waving at the crowd as he passed. The flag twirlers came out, but had some problems staying in sync. Onlookers gathered on the sides of the street for the parade. Then the local marching band entertained us.
CRCN: They're much trickier than a normal kite. This group finally got theirs in the air long enough to place second. It's time for the annual kite flying contest in location. One of the contests is managing to fly a box kite. Some of the kites go pretty high in strong wind.
Proposed: Quite a lot of participants and spectators. Participants gather the night before the wildflower triathlon. Getting ready to start the first leg. Participants in good spirits for the event. Another wave gets ready to swing.

Figure 1: Examples of visual storytelling result on SIND. Three stories are generated for each photo stream: story by GT, story by baseline CRCN and story by the proposed BARNN. The colored words indicate the semantic matches between the generation results with the GT (Liu et al., 2017)

6 Discussion

Compared with three different datasets, we can see from table 1, table 2 and table 3, BARNN model performs better than CRCN model in each dataset and each retrieval task. For example in table 3, BARNN R@1 value is 24.07 where CRCN



Ground truth: The view of the sunset was beautiful. We decided to get back on the road before it got too dark. When we hopped on the plane we could see the entire landscape. The view of the sunset was beautiful from the plane. I had a great time.
Proposed: We went on vacation to a beautiful country. We drove through a tunnel that we saw a lot of traffic on our way through the city. We took a helicopter ride to the top of the mountain to see the beautiful view. The sunset was the best part of the day. The sun was setting and it was a beautiful day for a hike.

Figure 2: Examples for narratives generated by the proposed model and stories from ground truth. The proposed is adversarial training model. The words in same colors indicate the correct semantic matches between the generated narratives and those from ground truth (Wang et al., 2018).

just is 9.87. The bigger recall top K value is, the better performance model has. It's the same as R@5 and R@10, the value 44.29 versus 28.74 and 53.06 versus 39.51. For median rank value, the less value is, the better performance the model has. Compared with BARNN median rank 9 with CRCN median rank 21, it confirms BARNN has better performance as well.

The reason why BARNN performs better than CRCN is that BARNN using bidirectional RNN and also creates embedding space for images and narrative sentences, which could handle with large image variance. This advantage is CRCN lack of. The evidence is from table 3 methods BARNN-sGRU and BARNN-EMB compared with CRCN. We can see that BARNN without sGRU or embedding space could perform better than CRCN as well and their value, for example R@1, without sGRU is 21.39 and without embedding space is 21.63, CRCN is just 9.87 and BARNN is 24.07. The number shows BRNN network performs quite better in this task than recurrent convolutional network. BRNN with sGRU and embedding space has micro-improvement than without.

For table 4, BARNN also has better quality of language generating in BLEU score and CIDEr value. Although these methods would be influenced by different language generating model, these values could partially confirm how generating sentence correlates with ground truth. See the figure 1, the colored words indicate the semantic matches between generation results with ground truth, and BARNN model has more similar semantic words mentioned in ground truth sentence compared with CRCN. For example the yellow word in ground truth *local marching band*, in CRCN there

is nothing related word to match it, but in BARNN *Another wave* could be matched it.

Due to adversarial training model did not compare with CRCN and BARNN directly, so we just discuss their user study result in table 5. This table shows that in both two criteria, relevance and story-style, BARNN performs better than CRCN, which to accord with our previous conclusion. However, Adversarial training model performs better than BARNN. In relevance criteria, adversarial training model's value is almost twice larger than BARNN's, but in story-style there is no quite large difference, which means relevance value could show how information density the generation sentence has and story-style value could show how languae model performance is. So it could be partially confirmed that Adversarial learning model has more powerful ability in image captioning than BARNN.

In my perspective of view, I think why adversarial model could perform better than the others is that introducing game theory would approach more to human brain operation. As rational agent, we learn different things from different games every moment. And learning is also a way like creating assumption and doing confirmation. This is quite similar as generative model and discriminative model.

At the end, we introduce threes different approaches in images story generating field and explain their basic principles. And then we discuss them difference and performance, explain the reason why they perform differently. We can see this images story generating task is solving step by step, from the initial CRCN approach to using BARNN, later stay at best performance Adversarial model till now. This interesting procedure of evolution also tells us when we use more artificial principles, we could be closer to the truth.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Satanjeev Banerjee and Alon Lavie. 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*, volume 29, pages 65–72.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Comput. Linguist.*, 34(1):1–34, March.

S. Chopra, R. Hadsell, and Y. LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546 vol. 1, June.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. 1989. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*.

L. Itti, C. Koch, and E. Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, Nov.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Andrej Karpathy and Li Fei-Fei. 2017. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676, April.

Andrej Karpathy, Armand Joulin, and Li Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 1889–1897, Cambridge, MA, USA. MIT Press.

Gunhee Kim, Seungwhan Moon, and Leonid Sigal. 2015. Ranking and Retrieval of Image Sequences from Multiple Paragraph Queries. In *28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Yu Liu, Jianlong Fu, Tao Mei, and Chang Wen Chen. 2017. Let your photos talk: Generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks. In *AAAI*, pages 1445–1452.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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.

Cesc C Park and Gunhee Kim. 2015. Expressing an image stream with a sequence of natural sentences. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 73–81. Curran Associates, Inc.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

K. Simonyan and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Dave Steinkraus, I Buck, and PY Simard. 2005. Using gpus for machine learning algorithms. In *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*, pages 1115–1120. IEEE.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator. cite arxiv:1411.4555.

Jing Wang, Jianlong Fu, Jinhui Tang, Zechao Li, and Tao Mei. 2018. Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. February.