Visual Storytelling via Predicting Anchor Word Embeddings in the Stories

Bowen Zhang[1]   Hexiang Hu[1]   Fei Sha[2]
[1]University of Southern California, [2] Google AI
zhan734@usc.edu, hexiangh@usc.edu, fsha@google.com*

Abstract

We propose a learning model for the task of visual storytelling. The main idea is to predict anchor word embeddings from the images and use the embeddings and the image features jointly to generate narrative sentences. We use the embeddings of randomly sampled nouns from the ground-truth stories as the target anchor word embeddings to learn the predictor. To narrate a sequence of images, we use the predicted anchor word embeddings and the image features as the joint input to a seq2seq model. As opposed to state-of-the-art methods, the proposed model is simple in design, easy to optimize, and attains the best results in most automatic evaluation metrics. In human evaluation, the method also outperforms competing methods.

1. Introduction

Visual storytelling, ie, narrating a sequence of images, is a challenging task [8, 6]. It demands an understanding of the underlying storyline of the images. The process is naturally subjective. It often focuses more on conveying the narrator's own interpretation than describing the images in factual terms.

For example, as pointed out by the creators of the popular dataset VIST [6], concatenating the descriptions of the images does not give rise to a desirable narrative story. Table 1 illustrates the difference in the corpus statistics on the aforementioned dataset. Despite similar in length, story and caption use very different sets of words. At least 40% of words that appear in stories do not appear in captions.

While this discrepancy has been well-documented, it is unclear how this insight could be used to devising effective models for visual storytelling. The task seems naturally gravitating to the method of SEQ2SEQ where we learn a mapping to encode a sequence of image features then to decode by outputting a sequence of words [3]. This method met some successes and is followed by others [5, 11, 12].

In this paper, we take a step toward identifying what might be needed for generating a narrative story. We hy-

---

*On leave from U. of Southern California (feisha@usc.edu)

	Storys	**Captions**
Vocabulary Size	29,614	24,534
Avg. Sent. Length	11.4	11.9
# of Nouns	6,831	7,772
# of Verbs	5,217	3,202
# of Adjectives	2,089	2,018
# of Adverbs	1,505	286

Table 1. Statistics of stories and captions on the VIST dataset [6]

pothesize that each narrative story needs to have a sequence of *anchor* words. For simplicity, we assume one anchor word per image. The anchor words form a prior on what can be "said" about the images. To narrate a sequence of images, our learning model just needs to predict the anchor word embedding for each image in turn and then supply the embeddings to a SEQ2SEQ model to generate the story.

But then, *what are the anchor words*? They are not explicitly given in the annotated dataset. As a first step, we have shown that we can use the words in the ground-truth stories as anchor words and learn a predictive model (from the image features) to predict the anchor word embeddings when the ground-truth stories are not available.

As opposed to several best-performing models for the same task, our model is simple in design and does not need to use reinforcement learning to optimize [5, 11, 12]. Yet it attains the best performance in several evaluation metrics.

We describe the idea of using anchor words in section 3, supported by the evidence that such words, when added to a vanilla SEQ2SEQ model for story generation, significantly improve its performance. We then describe how to train a predictive model to predict its embedding. In section 4, we report our evaluation results and conclude in section 5.

2. Related Work

There is a large body of work in the intersection of vision and language, cf. [7, 10].

Image captioning is closely related to visual storytelling. SEQ2SEQ and its variants are among the most popular learning approaches for the task [13, 10].

From the very beginning, the creators of the dataset for visual storytelling highlighted the difference of captioning from narratives [6]. In essence, narrative stories go beyond
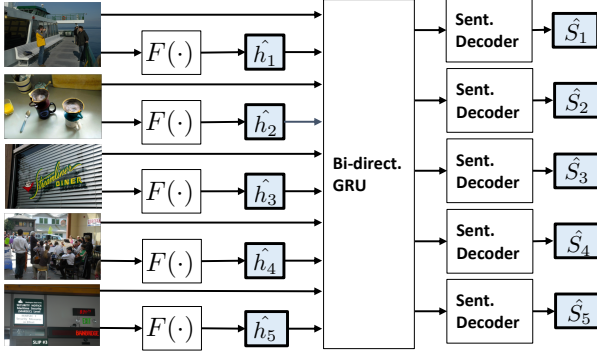
Figure 1. Conceptual diagram of our approach for visual storytelling. The key difference from a typical SEQ2SEQ model is the component of predicting anchor word embeddings from the images. The predictions are then fused with the image features as the inputs for generating desired narrative sentences.

	B@4	M	R	C
Image Only	$13.9_{\pm 0.4}$	$35.2_{\pm 0.1}$	$29.5_{\pm 0.3}$	$8.4_{\pm 0.9}$
Anchoring				
with **Noun**	$\mathbf{17.2}_{\pm 0.2}$	$\mathbf{39.0}_{\pm 0.2}$	$33.8_{\pm 0.1}$	$\mathbf{15.7}_{\pm 0.4}$
with **Verb**	$16.5_{\pm 0.1}$	$37.9_{\pm 0.2}$	$\mathbf{34.7}_{\pm 0.2}$	$13.0_{\pm 0.2}$
with **Adj.**	$15.2_{\pm 0.3}$	$36.6_{\pm 0.1}$	$31.9_{\pm 0.2}$	$11.3_{\pm 0.2}$
with **Adv.**	$14.9_{\pm 0.3}$	$35.9_{\pm 0.1}$	$31.0_{\pm 0.2}$	$10.4_{\pm 0.3}$

Table 2. Adding ground-truth words as anchor words to a SEQ2SEQ model significantly improves its performance where only image features are used. The higher numerical value indicates better performance.

the factual enumeration of objects and activities depicted in the images, which is often adequate for image captioning.

Recent approaches for visual storytelling have been using reinforcement learning (RL) to optimize complicated models [5, 12]. The approach proposed in this paper has the advantage of a simplified design and learning procedure, yet attains the best performance on several evaluation metrics.

3. Approach

The task of visual storytelling is to generate a sequence of narrative sentences $\{ \mathbf{S}_i, i = 1, 2, \ldots, N \}$, one for each of the N images $\{ \mathbf{I}_i, i = 1, 2, \ldots, N \}$. The order of the images is important and is fixed. Each of the generated sentences $\mathbf{S}_i$ could contain a variable length of words.

The main idea behind our approach is straightforward. For each image, we learn and apply a model to predict its anchor word embedding. The predicted embedding is then concatenated with the image feature. The combined feature is fed into a SEQ2SEQ [9] where the narrative sentence is generated as output. Fig. 1 illustrates the model design.

The key challenge is to learn the anchor word prediction model when the dataset does not provide anchor words explicitly. We begin by describing how we overcome this challenge. Then we introduce our model in detail.

3.1. What is an anchor word?

We are inspired by the comparison between narrative stories and captions on the same sequence of images, shown in Table 1. In particular, a large number of words used in narration do not appear in captions. Intuitively speaking, they are less likely visually grounded.

Thus, we conjecture that possible candidates for anchor words are the words in the narrative sentences. The analysis in Table 2 confirms the usefulness of this hypothesis.

Specifically, we train a model as in Fig. 1 with two variants. In the first variant, we supply only the image features. The results are reported in the row labeled as "Image Only". In the second variant ("Anchoring"), we select all the noun (alternatively, verb, adjective, or adverb) words as anchor words – one word per sentence in the story. We then train the SEQ2SEQ model by combining the image feature and the word embedding end to end. The results are reported in rows labeled with the part-of-speech (POS) tags of the selected words. For simplicity, all anchor words have the same POS tags. If there are multiple words with the same POS tags in each sentence, we randomly select one.

There are two points worth making. First, adding anchor words, irrespective of their types, significantly improves the performance of the SEQ2SEQ model with image features only. Note that the results in "Image Only" is on par with state-of-the-art results [12]. Secondly, among all POS tags, nouns as anchor words seem to be the most beneficial ones on all metrics except R(OUGE) where verbs improve more.

In the rest of this paper, we use nouns in the stories as the anchor words.

3.2. Model and Learning

The data for our learning task is augmented with a list of anchor words $\{w_i \mid i = 1, 2, \ldots, N\}$ corresponding to the images. Next, we explain how to learn each component.

Anchor word embedding predictor We learn a model $F(\mathbf{I}_i)$ to predict w_i. $F(\cdot)$ is parameterized by a one-hidden-layer multi-layer perception (MLP) with ReLU non-linearity. The input could be the features for the ith image or all the images in the same sequence. In practice, there is no significant difference.

To be able to generalize to new anchor words, we predict its embedding and cast learning $F(\cdot)$ as a regression problem. To obtain the target (ie, the "ground-truth" embedding) for the word w_i, we take the embeddings from the "Anchoring" model in Table 2. $F(\cdot)$ is then optimized to reduce the mean squared error between the predictions and the target anchor words embeddings.

Story generation model Similar to state-of-the-art visual storytelling methods [12, 6], we use a SEQ2SEQ model [9] as story generator. Concretely, a bidirectional gated recurrent neural network [2](GRU) is used to encode the concate-

Method	B@1	B@2	B@3	B@4	M	R	C
AREL [12]	63.8	39.1	23.2	14.1	35.0	29.5	9.4
Show, Reward and Tell [11]	43.4	21.4	10.4	5.2	-	-	11.4
HSRL w/ Joint Training [5]	-	-	-	12.3	35.2	30.8	10.7
SEQ2SEQ+Heuristics [6]	-	-	-	-	31.4	-	-
H-Attn-Rank[14]	-	-	20.8	-	33.9	29.8	7.4
StoryAnchor: Image Only	$62.2_{\pm 2.5}$	$38.3_{\pm 1.7}$	$22.7_{\pm 0.8}$	$13.9_{\pm 0.4}$	$35.2_{\pm 0.1}$	$29.5_{\pm 0.3}$	$8.4_{\pm 0.9}$
StoryAnchor: w/ Predicted Nouns	$65.1_{\pm 0.3}$	$40.0_{\pm 0.2}$	$23.4_{\pm 0}$	$14.0_{\pm 0.1}$	$35.5_{\pm 0.0}$	$30.0_{\pm 0.1}$	$9.9_{\pm 0.1}$

Table 3. Comparison of state-of-the-art method for the visual storytelling task on the VIST dataset. Our "Image Only" model is a reimplementation of XE+SS [12] with the authors' public available codes.

nated feature of the image and the predicted anchor word embedding and to produce a sequence of hidden states

$$\boldsymbol{v}_i = \text{BiGRU}(\mathbf{I}_i, F(\mathbf{I}_i)) \qquad (1)$$

The sequence of the hidden states is then decoded by a one-layer GRU. Both the encoder and the decoder are trained to maximize the likelihood of ground-truth stories.

3.3. Other Implementation Details

Visual and textual representation We extracted the 2048 dimension feature from the penultimate layer of ResNet-152 [4] as visual representations. The 512-dimensional word embedding is randomly initialized, which are fine-tuned in the training. Note that the anchor words are sharing the word embeddings with the words in the vocabulary.

Model details The concatenated features of the image and the anchor word embedding are projected into a 2048 dimensional feature with a one-hidden-layer MLP. Then, a one-layer BiGRU with 256-dimensional hidden states generates contextual embedding $\boldsymbol{v}_i$ of 512 dimensions, to serve as hidden states representation. A standard SEQ2SEQ decoder with one-layer GRU with 512 hidden dimensions is used on top of these hidden states to generate a story.

Optimization As mentioned, the model is trained in two stages. In the first stage, ground-truth anchor words (nouns in the stories) are used to train the encoder-decoder as well as the embeddings end to end. The model is trained with mini-batches and ADAM for 100 epochs. Each mini-batch contains 64 sampled stories. The learning rate is initialized as 4e-4 and schedule sampling [1] has been used. The probability of schedule sampling is first set to be 0.05, increased by 0.05 every 5 epochs till 25 epochs. In the second stage, the predictor $F(\cdot)$ is trained. Specifically, we use the model that achieves the highest Meteor score on the validation set in the first stage training as a pre-trained model. We use the same optimization hyper-parameter to train the predictor with encoder-decoder model in an end-to-end way. The encoder-decoder and the word embeddings are kept fixed.

Inference At the inference time (ie, narrating a sequence of images), we perform beam search for sentence decoding with a beam size of 3.

4. Experiments

4.1. Experimental Setups

Dataset We use the VIST dataset [6] for evaluation. It contains 10,032 visual albums with 50,136 stories. Each story contains five narrative sentences, corresponding to five grounded images respectively.

Evaluation We follow the evaluation setup used in [6, 12, 14, 5]. For each testing album, we sample one image sequence and generate a story based on that image sequence. The story is then scored against all 5 reference stories of that album. We use the evaluation code provided by the [14][1]. We report results with average BLEU, METEOR, ROUGE, and CIDER over the test split. We evaluate over 3 random runs and compute the means and variances of the metrics.

Identifying anchor words We use NLTK POS tagger to get the tags. Each sentence contains on average 2.63 nouns, 2.0 verbs, 0.8 adjectives, and 0.5 adverbs. We use 'UNK' as the anchor word when there is no corresponding POS tag.

4.2. Main Results

We compare our method (*StoryAnchor*) to several state-of-the-art methods [6, 14, 12, 11, 5]. Figure 3 shows that our model performs significantly better than others in almost all evaluation metrics. In ROUGE and CIDER, approaches of using reinforcement learning seem to perform well.

We also conduct human evaluations to compare the outputs of our model and AREL [12]. We follow [12] and design three questions to evaluate the relevance, concreteness, and coherence of generated stories and image sequences. 150 generated stories from the test splits are evaluated. For each story, 5 AMT workers are assigned. The reports are reported in Table 5. Our approach performs better.

4.3. Analysis

Is visual storytelling fundamentally out of reach of machines? Are the metrics being used now to guide the design of our systems the right ones?

---

[1]https://github.com/lichengunc/vist_eval. This is the most commonly used evaluation script nowadays.

	B@1	B@2	B@3	B@4	M	R	C
Human	51.2 $_{\pm 0.2}$	25.0 $_{\pm 0.2}$	11.7 $_{\pm 0.2}$	5.6 $_{\pm 0.2}$	28.4 $_{\pm 0.1}$	24.5 $_{\pm 0.1}$	7.8 $_{\pm 0.1}$
StoryAnchor: Image Only	58.6 $_{\pm 0.2}$	34.7 $_{\pm 0.2}$	20.0 $_{\pm 0.1}$	11.2 $_{\pm 0.1}$	34.0 $_{\pm 0.1}$	28.3 $_{\pm 0.1}$	8.8 $_{\pm 0.1}$
StoryAnchor: w/ Predicted Nouns	60.7 $_{\pm 0.2}$	35.8 $_{\pm 0.2}$	20.3 $_{\pm 0.1}$	11.9 $_{\pm 0.1}$	34.5 $_{\pm 0.0}$	28.9 $_{\pm 0.0}$	10.1 $_{\pm 0.1}$
StoryAnchor: w/ Ground-truth Nouns	65.1 $_{\pm 0.2}$	40.3 $_{\pm 0.1}$	23.9 $_{\pm 0.1}$	14.7 $_{\pm 0.0}$	37.7 $_{\pm 0.1}$	32.3 $_{\pm 0.1}$	16.2 $_{\pm 0.1}$

Table 4. Evaluating human performance by automatic evaluation procedures. Machine outperforms human in all metrics.

StoryAnchor: w/ Predicted Nouns	AREL	Tie
Relevance	53.2%	40.4%	6.4%
Concreteness	45.1%	38.1%	16.8%
Coherence	48.9%	42.3%	8.8%

Table 5. Human evaluation on the generated stories

StoryAnchor: w/ Predicted Nouns	AREL	Human	Unsure
19.9%	18.0%	**57.8%**	4.2%

Table 6. Which stories are preferred by human readers

The results in Table 4 highlight the issues. There, we assess *how well human storyteller would do*. For each album, we randomly select one human-written ground-truth story as "generated" story and the other 4 as "reference" stories. We then evaluate human performance by scoring the generated story. For a fair comparison, we re-evaluated all of the learning models with 4 sampled reference stories. Mean evaluation performances over five random runs are reported.

Clearly, the learning models outperform human storyteller significantly in every metric! Yet, our "Turing test" suggests the opposite. In Table 6, over 450 stories (3 for each of the 150 sequences of images), we report the percentages of 150 AMT workers' preference of stories by two learning models and one human annotator. Human storytelling is much more preferred. The misalignment between human evaluation and automatic evaluation metrics is likely a bottleneck for developing new methods for this task.

5. Conclusion

The proposed StoryAnchor model is simpler in design. Yet, it attains the best results on most automatic evaluation metrics. The key insight is to use "anchor words" to model the evolvement of the underlying storyline. Crudely, those words are the "topics" or "states" of the narrators. While those notions are not explicitly annotated in the current dataset, we have selected the nouns in the ground-truth stories as targets for learning an anchor word predictor.

References

[1] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179, 2015.

[2] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[3] D. Gonzalez-Rico and G. Fuentes-Pineda. Contextualize, show and tell: a neural visual storyteller. *arXiv preprint arXiv:1806.00738*, 2018.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Q. Huang, Z. Gan, A. Celikyilmaz, D. Wu, J. Wang, and X. He. Hierarchically structured reinforcement learning for topically coherent visual story generation. *arXiv preprint arXiv:1805.08191*, 2018.

[6] T.-H. K. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, et al. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, 2016.

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[8] C. C. Park and G. Kim. Expressing an image stream with a sequence of natural sentences. In *Advances in neural information processing systems*, pages 73–81, 2015.

[9] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[10] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[11] J. Wang, J. Fu, J. Tang, Z. Li, and T. Mei. Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. AAAI, 2018.

[12] X. Wang, W. Chen, Y.-F. Wang, and W. Y. Wang. No metrics are perfect: Adversarial reward learning for visual storytelling. *arXiv preprint arXiv:1804.09160*, 2018.

[13] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[14] L. Yu, M. Bansal, and T. L. Berg. Hierarchically-attentive rnn for album summarization and storytelling. *arXiv preprint arXiv:1708.02977*, 2017.