CSCI 544: Group 46 Final Project Report
ImagiNarrate: Building a Narrative with Images and Generated Captions

Asmita Chotani , **Apurva Gupta** , **Chetan Chaku** , **Rutuja Oza** and **Priya Nayak**
University of Southern California
[achotani, apurvagu, chaku, rgoza, psnayak] @usc.edu

Abstract

This paper introduces a new natural language processing (NLP) approach to solve the problem of visual storytelling that utilizes image features to generate captions and subsequently develop a coherent storyline for the images. By incorporating caption embeddings in the story generation process, the proposed approach aims to provide a more relevant and informative description of the images that can be used to build a cohesive and engaging narrative. Captions often utilize more expressive language and incorporate fine-grained concepts of the corresponding images, thereby providing additional contextual information that can potentially enhance the quality of story generation models. The paper evaluates the ImagiNarrate model compared to the AREL model (Wang et al., 2018) used for story generation based on traditional metrics like BLEU, CIDEr, and ROUGE as human evaluation of the generated stories.

1 Introduction

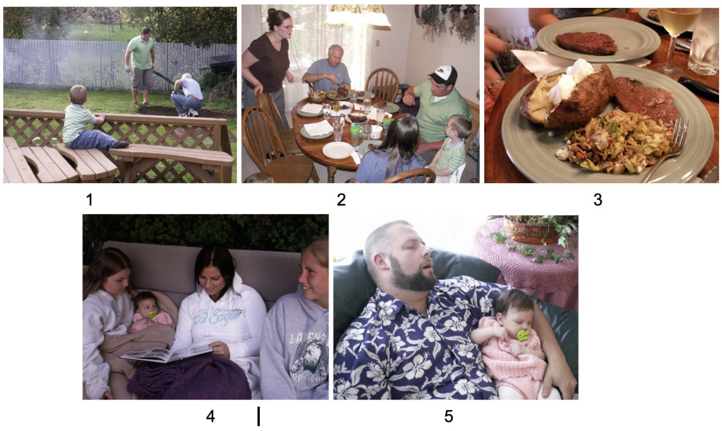
Storytelling is one of the few human capabilities that can be considered complicated for a machine. Humans have the ability to add emotions as well as imagination when creating a story. In an attempt to work on storytelling capabilities using NLP, we proposed ImagiNarrate, a framework to produce a visual narrative using a collection of images and captions generated from those images.

To create a compelling story, it is necessary to use techniques from both image processing and natural language processing. ImagiNarrate uses generated captions to build up the additional context for a story based on the series of images.

ImagiNarrate framework builds upon the state-of-art AREL model's framework introducing a caption generation model and further using a policy and a reward models for story generation(Wang et al., 2018). While the policy model performs basic actions to create a story sequence, the reward model learns from ground truth stories to determine the implicit reward function. The learned reward function is then used to optimize the policy, resulting in a more efficient and effective algorithm.

For evaluation, both automatic metrics and human evaluation were conducted. Comparing results of ImagiNarrate with stories created only using images confirmed the importance and usefulness of the augmentation of captions and images for storytelling.

ImagiNarrate gains performance improvement over the state-of-art model (AREL), which generates a story only based on the sequence of images. The effectiveness of our approach was evaluated by conducting experiments on the Visual Storytelling (VIST) dataset and comparing the results with the state-of-art model. This paper will delve into the details of our proposed framework and the experimental results, demonstrating the superiority of our approach over the baseline.



STORY GENERATED: " The family gathered for the family dinner . Everyone was having a great time . They had a lot of food . She was so happy to see the kids . After the night , we all had a great time ."

Figure 1: Generated Story Example

2 Related Work

(Ting-Hao et al., 2016) introduced the first dataset for sequential vision-to-language tasks and provided the first baselines for the task of visual storytelling on their proposed dataset - VIST, in 2016. Their model is trained using GRUs for the image sequence encoder and the story decoder. They also motivate using automatic metrics like Meteor to evaluate the generated stories by providing a comparative analysis against human judgment.

(Gonzalez-Rico and Fuentes-Pineda, 2018) proposed the neural vision storyteller, which extends the image description model by (Vinyals et al., 2014). The original model consisted of a CNN encoder and LSTM decoder. The newly proposed model uses 5 different decoders, one for each image in the sequence, and then consolidates the output from the decoders by concatenation.

(Wang et al., 2018) proposed the AREL model in 2018, which was based on Reinforcement learning and produced SOTA results on automatic metrics scores for Visual Storytelling. The paper uses an encoder/decoder-based policy model, which is trained with the help of another CNN-based reward model.

Further, (Kim et al., 2018) proposed GLAC Net, which generates visual stories by combining global-local (glocal) attention and context-cascading mechanisms. Some knowledge graph (Hsu et al., 2019) and common-sense (Chen et al., 2021) based models were also introduced over the years but weren't able to provide any significant improvement for the task at hand based on traditional metrics (Wang et al., 2022).

(Wang et al., 2022) performed a comparative study of some of the above-discussed models using traditional automatic metrics like Meteor and Bleu but concluded that they don't correlate enough with human evaluation and also proposed their own metric to overcome this hurdle.

3 Problem Description

Like other Storytelling frameworks, the ImagiNarrate framework expects the input to be a sequence of images, which are provided in the form of ResNet-152 features and outputs a story related to these images, i.e., a sequence of words.

The novelty in our framework is the inclusion of generated captions of the ResNet image features, which are then augmented with the image embeddings to form the input of the story generation model.

The additional captions augmented with the images help provide additional context to the model and help generate more coherent and fluent stories compared to the AREL model.

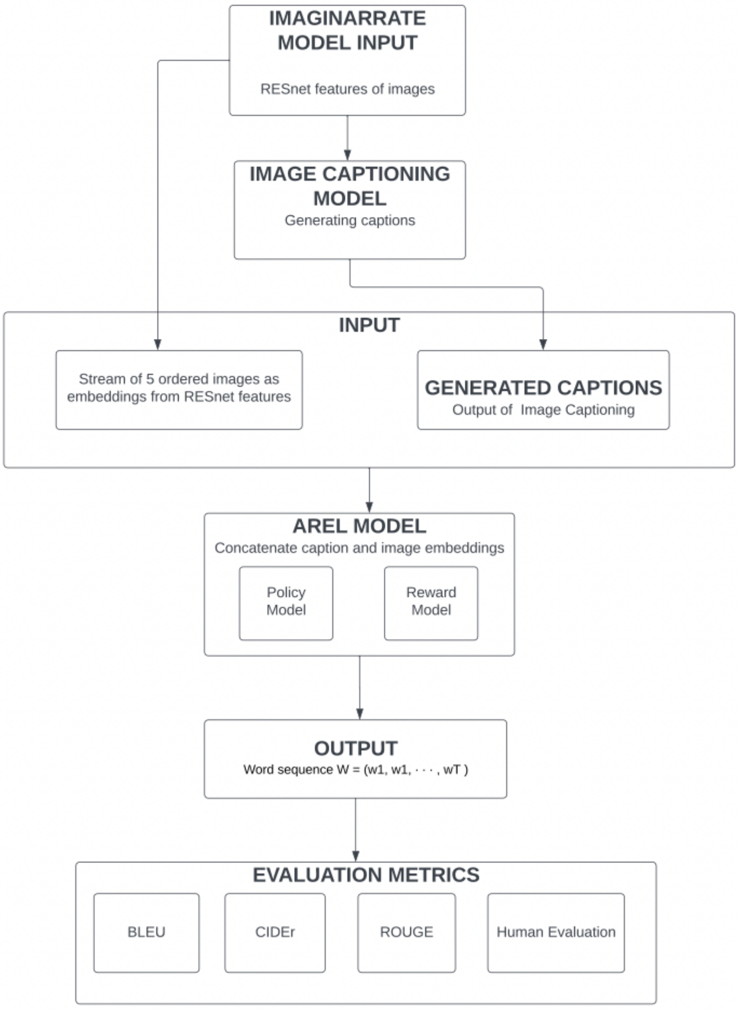

IMAGINARRATE MODEL INPUT
RESnet features of images
IMAGE CAPTIONING MODEL
Generating captions
INPUT
Stream of 5 ordered images as embeddings from RESnet features
GENERATED CAPTIONS
Output of Image Captioning
AREL MODEL
Concatenate caption and image embeddings
Policy Model
Reward Model
OUTPUT
Word sequence W = (w1, w1, · · · , wT)
EVALUATION METRICS
BLEU
CIDEr
ROUGE
Human Evaluation

Figure 2: Proposed Model Architecture

4 Methodology

4.1 Overview

ImagiNarrate tackles the task of generating a story consisting of a sequence of words, where each word belongs to the vocabulary V obtained from the training set. The input to our model is a sequence of ResNet features of five images, denoted as I.

$$I = (i_1, i_2, ..., i_5)$$

To obtain suitable captions for each of these images, a caption generation model first generates captions for each of the five images denoted as C .

$$C = (c_1, c_2, ..., c_5)$$

where T is the number of words in the caption and each word belongs to the vocabulary V. The embeddings of the image ResNet features and generated captions are further concatenated, where embeddings E are in the form of

$$E(I, C) = \sum_{k=1}^{5} (i_k \oplus c_k)$$

creating a sequence of augmented embeddings that are inputted to the story generation model. The story generation model then generates the desired sequence of words,

$$W = (w_1, w_2, ..., w_T) \forall w_i \in V$$

The story generation model employs a reinforcement learning framework inspired by the state-of-art model(Wang et al., 2018) and consists of a generative model called the Policy Model and a discriminative model called the Reward model, both with the same architecture. The generative model generates a story from the sequence of images provided as input, while the discriminative model computes a reward based on the similarity of the generated story to a human-annotated ground truth story.

4.2 Models

4.2.1 Captioning Model

The captioning model employs an LSTM-based architecture.The ResNet-152 features of the image are fed to the LSTM decoders as inputs, which generates captions for each image where each word of the caption belongs to the Vocabulary V. The cleaned and pre-processed captions are then converted to a list of respective indexes from V and sent to the story generation model.

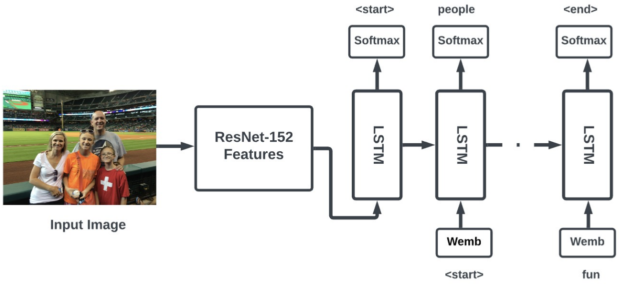
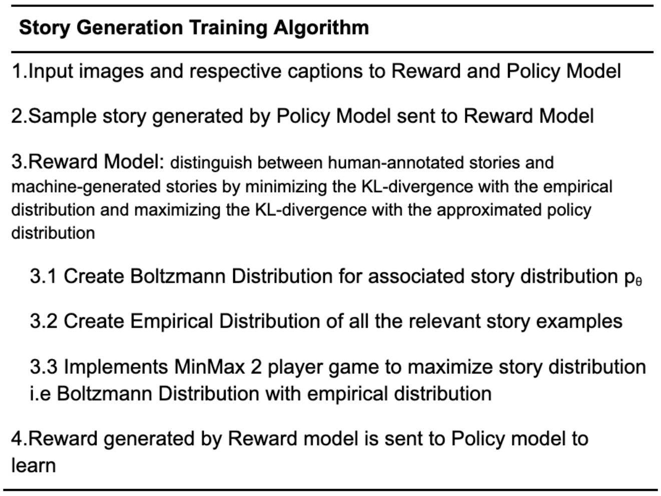


Figure 3: Captioning Model Architecture

4.2.2 Story Generation Model

- **Policy model :** In our proposed approach, the generative model comprises a bidirectional gated recurrent unit (GRU) encoder that takes in a sequence of ResNet features augmented with word embeddings to create a high-level context vector. The decoder generates each substory W_i using the context vectors generated by the encoder, and the final story output is obtained by concatenating all the intermediate substories.

- **Reward Model :** The discriminative model rewards each of the substories generated by the encoder model by first querying the word embeddings of the substory, followed by extracting n-gram features using multiple convolutional layers with different kernel sizes. These features are then projected into the sentence-level representation space by pooling layers. We combine the sentence representation with the visual feature of the input image and the generated captions through concatenation and feed them into the final fully connected decision layer. The reward model outputs an estimated reward value.



Story Generation Training Algorithm

1.Input images and respective captions to Reward and Policy Model

2.Sample story generated by Policy Model sent to Reward Model

3.Reward Model: distinguish between human-annotated stories and machine-generated stories by minimizing the KL-divergence with the empirical distribution and maximizing the KL-divergence with the approximated policy distribution

3.1 Create Boltzmann Distribution for associated story distribution p_θ

3.2 Create Empirical Distribution of all the relevant story examples

3.3 Implements MinMax 2 player game to maximize story distribution i.e Boltzmann Distribution with empirical distribution

4.Reward generated by Reward model is sent to Policy model to learn

4.3 ImagiNarrate Framework

The ImagiNarrate framework is a novel approach to generating stories based on a sequence of 5 images. The framework follows a step-by-step process starting with the use of a captioning model to generate a caption for each of the images in the input sequence. In the next step, the ResNet-152 features of each image in the sequence are combined with the word embedding features of the respective caption. This step aims to capture both the visual and semantic information from the images and captions. The augmented image + caption features are then fed into the story generation model, which is responsible for generating stories. Finally, the story generation model uses a reward-based approach to optimize the generated stories based on the input images and captions.

5 Experiments and Analysis

5.1 Experimental Setup

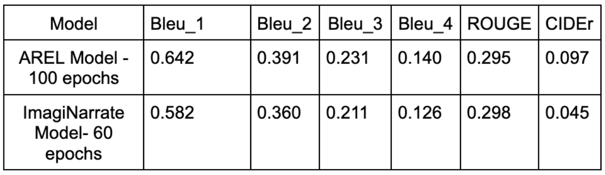
VIST Dataset : The project uses the ResNet features of the VIST dataset (vis) for training the image to caption model and further using the captions along with images for the storytelling task. VIST is specifically for sequential vision-to-language tasks, which makes it perfect for our use case. A single VIST sample consists of an image sequence paired with a description for individual images and a description of a human-constructed narrative over the sequence. The dataset consists of a total of 10,117 Flickr albums and 210,819 unique images which have been divided into 40,098 training, 4,988 validation, and 5,050 testing samples. Each sample contains one story that describes 5 selected images from a photo album (mostly one sentence per image). And the same album is paired with 5 different stories as references.

Training Details : Image Features extracted using a pre-trained ResNet-152 model (He et al., 2016) were used as an input for the image-to-caption model as well as the storytelling task. A vocabulary of size 9,838 was created to include words appearing more than three times in the training set to support the caption generation task and story generation task.

Evaluation Metrics : In order to comprehensively evaluate our method on the dataset, we evaluated our model against the baseline AREL model which only uses the images to generate the stories on the basis of the above metrics using automatic evaluation techniques including BLEU, ROUGE, and CIDEr.

5.2 Results

In Figure 4, we compare ImagiNarrate with state-of-art AREL model (Wang et al., 2018) which report achieving the best-known results on the VIST dataset. We first implement the baseline model, which shares the same architecture as AREL, and then make changes accommodating an additional input which is the generated captions. As shown in Figure 4, our framework reaches quite close to the AREL results for majority metrics and outperforming AREL on ROUGE, even though the number of epochs runs to train the ImagiNarrate model was limited to 60 epochs. Human evaluation was performed for the framework by circulating a google form with the image sequences and generated story and asked participants to rank the story on how related it was with the image sequence, their engagement, flow and fluency. The average score for relevance was 4/5, engagement 3.5/5 , flow 3/5 and fluency was 3/5.



Model	Bleu_1	Bleu_2	Bleu_3	Bleu_4	ROUGE	CIDEr
AREL Model - 100 epochs	0.642	0.391	0.231	0.140	0.295	0.097
ImagiNarrate Model- 60 epochs	0.582	0.360	0.211	0.126	0.298	0.045

Figure 4: Automatic evaluation on the VIST dataset. We report BLEU, ROUGE, and CIDEr scores of the SOTA framework and our proposed framework

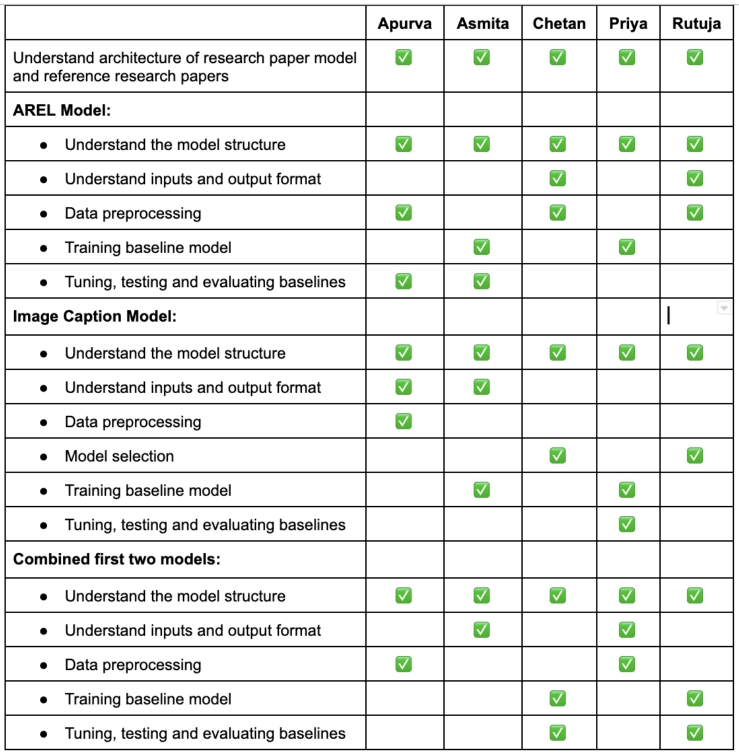
6 Future Scope

One of the largest impacts of this technique if successful would be creating novels from comic books automatically. Visually impaired people can not access comics in the usual way others do and translating comics to novels is a hard and time-consuming manual task. Using our model these comics can directly be converted into textual stories which are easy to convert to braille, making them accessible to the visually impaired.

7 Conclusion

In this paper, we introduce a novel approach to present a reward learning model that can generate human-like stories through the utilization of image sequences and captions. Notably, the captions utilized in this model are produced by a separate model. The generated stories' performance is evaluated by well-known metrics such as BLEU, CIDEr, and ROUGE. Since the stories generated are quite diverse and creative, human evaluation was also performed. Although our research has produced promising results, we believe there is still considerable room for improvement in the narrative paragraph generation task, particularly in terms of better simulating human imagination to create more diverse and vivid stories.

Distribution of Work



	Apurva	Asmita	Chetan	Priya	Rutuja
Understand architecture of research paper model and reference research papers	✅	✅	✅	✅	✅
AREL Model:					
• Understand the model structure	✅	✅	✅	✅	✅
• Understand inputs and output format			✅		✅
• Data preprocessing	✅		✅		✅
• Training baseline model		✅		✅	
• Tuning, testing and evaluating baselines	✅	✅			
Image Caption Model:					
• Understand the model structure	✅	✅	✅	✅	✅
• Understand inputs and output format	✅	✅			
• Data preprocessing	✅				
• Model selection			✅		✅
• Training baseline model		✅		✅	
• Tuning, testing and evaluating baselines				✅	
Combined first two models:					
• Understand the model structure	✅	✅	✅	✅	✅
• Understand inputs and output format		✅		✅	
• Data preprocessing	✅			✅	
• Training baseline model			✅		✅
• Tuning, testing and evaluating baselines			✅		✅

References

Github repository - imaginarrate code.

Vist dataset.

Hong Chen, Yifei Huang, Hiroya Takamura, and Hideki Nakayama. 2021. Commonsense knowledge aware concept selection for diverse and informative visual storytelling. arXiv.

Diana Gonzalez-Rico and Gibran Fuentes-Pineda. 2018. Contextualize, show and tell: A neural visual storyteller. arXiv.

Ansar Hani, Najiba Tagougui, and Monji Kherallah. 2019. Image caption generation using a deep architecture. pages 246–251.

Chao-Chun Hsu, Zi-Yuan Chen, Chi-Yang Hsu, Chih-Chia Li, Tzu-Yuan Lin, Ting-Hao 'Kenneth' Huang, and Lun-Wei Ku. 2019. Knowledge-enriched visual storytelling. arXiv.

Chi-Yang Hsu, Yun-Wei Chu, Ting-Hao 'Kenneth' Huang, and Lun-Wei Ku. 2021. Plot and rework: Modeling storylines for visual storytelling. arXiv.

Taehyeong Kim, Min-Oh Heo, Seonil Son, Kyoung-Wha Park, and Byoung-Tak Zhang. 2018. Glac net: Glocal attention cascading networks for multi-image cued story generation. arXiv.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft coco: Common objects in context. arXiv.

Ting-Hao, Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. arXiv.

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

Eileen Wang, Caren Han, and Josiah Poon. 2022. Rovist:learning robust metrics for visual storytelling. arXiv.

Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. 2018. No metrics are perfect: Adversarial reward learning for visual storytelling. arXiv.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. arXiv.