Visual Story Telling

Saswat Kumar Dash

Visual Story Telling

by

Saswat Kumar Dash

Under the supervision

of

Dr. Debi Prosad Dogra & Dr. Manoranjan Satpathy

Thesis is submitted to the

Indian Institute of Technology Bhubaneswar

for award of the degree

of

Bachelor of Technology

SCHOOL OF ELECTRICAL SCIENCES

INDIAN INSTITUTE OF TECHNOLOGY BHUBANESWAR

NOVEMBER 2021

CERTIFICATE

Certified that the thesis entitled **"Visual Story Telling"**, submitted by **Mr. Saswat Kumar Dash** to the Indian Institute of Technology Bhubaneswar, for the award of the degree Bachelor of Technology has been accepted by the examiners and that the student has successfully defended the thesis in the viva-voce examination held today.

Date: Dr. Manoranjan Satpathy
(Supervisor)

Date: Dr. Debi Prosad Dogra
(Supervisor)

DECLARATION

I certify that

1. The work contained in the thesis is original and has been done by myself under the general supervision of my supervisor.
2. The work has not been submitted to any other Institute for any degree or diploma.
3. I have followed the guidelines provided by the Institute in writing the thesis.
4. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.
5. Whenever I have used materials (data, theoretical analysis, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.
6. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Signature of the Student

ACKNOWLEDGMENTS

I owe a debt of gratitude to all the people who helped, influenced and motivated me during my wonderful journey of four years at IIT Bhubaneswar.

First and foremost I would like to express my sincere gratitude to my thesis Supervisor, Dr. Debi Prosad Dogra for his vital support and assistance. It has been great experience learning with him and his moral support in every direction any time when I required inspite of his busy schedule. His guidance and encouragement made it possible to achieve the goal. His high level knowledge and enthusiastic behavior motivated me to tackle the research work. He is my source of inspiration. He is the pioneer of my life.

I extend my thanks to all peers, juniors and professors for guiding me well throughout this project work and helping me to make it a success.

Last but not the least I would like to take this opportunity to embrace my parents and my beloved ones for everything they have done for me. I am greatly indebted towards them. I thank my brother and sisters for their support throughout my life. They not only assisted me financially but also extended their support morally and emotionally. Hence again a special thanks to who extended their care and supportive hands in every aspects of my life in these two years of journey.

Saswat Kumar Dash

ABSTRACT

Since the beginning of human race, eyes have been a boon to mankind that help us perceive and apprehend our surrounding in best possible way. In this line, researcher around 2015 got interested in representing image sequence in the form of a story, and tried solving this complex problem with the help of natural language, computer vision, and deep learning. In early 2010, Automatic captioning of images and videos was gaining much popularity. But, soon researchers realised the visual story telling, a sub-domain of Sequence to Sequence Captioning is not so straight-forward like captioning. A story has some underlying hidden context, imaginative components and most of time emotions of the narrator. To generalize a story of best quality is a tough task. Some researchers argued that there is lack of proper quality metric for evaluating the story generated and the publicly celebrated VIST (Visual Story Telling) is biased.

A commercial grade model of visual story telling, can be used to help visually impaired people. It can be pivotal in revolutionizing robotics visual sensor system.

In this paper, we have tried to give a comprehensive analysis of various approaches used by researchers in the recent years and tried to achieve a chronological pattern between them. We found the VIST data-set is annotated by outsourcing the images to Amazon Mechanical Turk(AMT). This is also a serious issue as the pictures of events taken by a person is annotated by someone else. This bring in the issue of emotion biasing. Also, VIST data-set has most images of graduation, wedding ceremonies and picnic, adding event biased to dataset.

At end, we aim to make a new data-set that has much vivid range of self-annotated images of people, captures wide range of emotions, events and use it to run the state-of-the-art (SOTA) models to see how they perform.

Contents

Chapter 1

Introduction

1.1 Overview of Visual Story Telling

In recent years, visual captioning has been an active area of research. They focus on describing the content of image in descriptive sentences (mostly single line). Though it has achieved impressive results, its capability of performing human-like understanding is still restrictive. To further investigate machine's capabilities in understanding more complicated visual scenarios and composing structured expressions, visual storytelling has been proposed. Visual storytelling goes one step further ahead of visual captioning: it summarizes the idea of a photo stream and tells a story about it. Figure 1.1.1 shows an example of visual captioning and visual storytelling. It is observed that stories contain **rich emotions** (excited, happy, not want) and **imagination** (siblings, parents, school, car). Mostly, story-teller tries to add personal experience while describing a scene. It, therefore, requires the creative capability to associate with concepts that do not explicitly appear in the images. Mostly, stories are more subjective, so there barely exists standard templates for storytelling. As shown in Figure 1.1.1, the same photo stream can be paired with diverse stories, different from each other and each of them are correct in various context. A lack of proper evaluation metric make this task even tiresome.

The evolution of Artificial Intelligence, has made the task increasingly plausible to develop models that interpret vision and language in a human-like manner. A crucial element of such models is the capacity to not only match images with surface-level descriptions, but to infer deeper contextual meaning. Recent literature has begun to

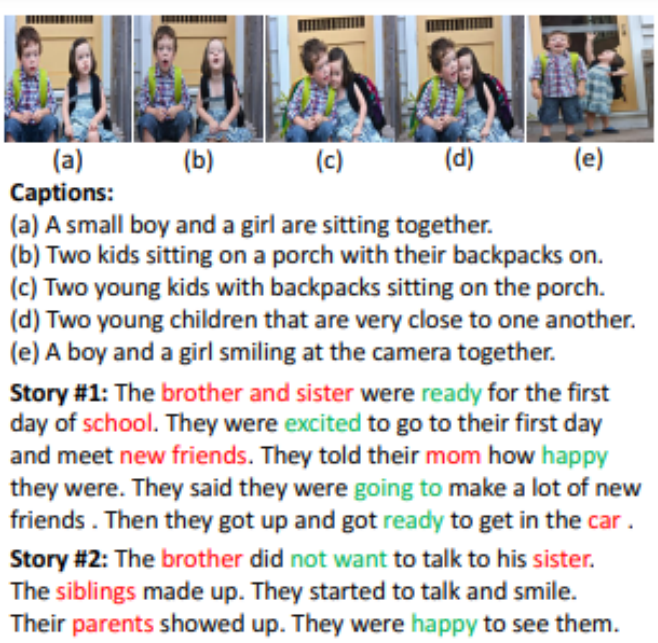


Figure 1.1.1: **An example of visual storytelling and visual captioning. Both captions and stories are shown here: each image is captioned with one sentence, and we also demonstrate two diversified stories that match the same image sequence.** [22]

refer to this task as **visual storytelling**: the generation of a cohesive, sequential set of natural-language descriptions across multiple images.

Visual Story telling can be viewed as **Sequence to Sequence** modelling problem. Whenever generating a set of stories, it should have the following properties:-

Relevance: The story should accurately describes what is happening in the image stream and covers the main objects appearing in the images. It should not misinterpret objects.

Expressiveness: The story generated should be coherent, grammatically and semantically correct, with no or minimal repetition. The generated sentences should be expressive in nature and convey certain emotions.

Concreteness: The story should narrate precisely what is in the image rather than giving very general descriptions. It should be able to infer plots by surfing through various image sequences.

In real-life when we talk about narrative of a story, it has a personal and emotion touch with the stories we frame. A person liking sunset may fill the story of a sunset scene with excited and emotional words like (soothing, pleasant, beautiful), whereas a neutral man might just use simple words like (the sun is setting, sky is yellowish). Here comes a massive difference in terms of stories we frame.

The art of story narrative has emotion, imagination specific to a person. In recent years, people have tried adopting models that incorporates emotion as a parameter in model [15]. But, we found out the dataset they train on is highly biased. So, they fail to generate human specific languages.

From previous related work we found that the state-of-the-art dataset, **VIST** (Visual Story Telling), previously known as **SIND** (Sequential Image Narrative Dataset) is highly biased [22]. In fact, it was found no automatic metric is perfect for evaluating the quality of the model [22]. The metrics are very misleading at times and needs a human judge to monitor the quality of the sentences framed.

When, we analyzed how the VIST dataset was formed [28] (also discussed in section 2.2.2), the study revealed that the dataset was formed by outsourcing the images collected from Flickr with open source community. This leads to a serious problem in the formation of dataset. The pictures taken by some other person is annotated by someone else. Thus, the emotions described in the sentences have become biased. In this paper, we are proposing formation of a new dataset where the images collected will be annotated by the person themselves and those will be used for training. We are going to analyze the performance of the various proposed models (in recent years) on our dataset and compare it with the VIST dataset.

1.2 Problem Statement

Given a sequence of images as input, the visual storytelling task is about building a model that can generate a coherent textual narrative as output. Formally, it can be viewed as a **Sequence to Sequence Modelling** problem. Given a sequence of images

$$I = \{I_1, I_2, \ldots, I_n\},$$

the aim is to generate a sequence of correlated sentences

$$S = \{S_1, S_2, \ldots, S_n\},$$

to depict the visual representations of the sequence. The problem can be mapped to a supervised learning problem, where the input label is images, and the output is a sequence of correlated sentences, that best represents the idea of the image sequence. The first and only curated dataset for visual storytelling task, to date, is the **VIST dataset**, (which is detailed under Section 2.2.2) and released by the work that popularized visual storytelling [28]. All the work that has followed has heavily relied on using the same dataset and proposed architectures that are pseudo dependent on the composition of the data.

The main aim is to study this VIST dataset and various solution approach proposed in last five year for visual story telling.

1.3 Main Challenges in Visual Story-Telling

1.3.1 Embedding

The images cannot be directly given as input. Thus, the image is first converted into a form of embedding that establishes the relation among various objects present in the image using some common features. The image is converted into real numbers vector by using these embedding, that preserves the spatial relationship.
Forming a vector representation of the input image is a tough task.
Some solution

1. Using the last fully connected (fc) layers of the CNN architecture to use as the vector representation of the input image.

2. Use graph Neural Network (GNN) to generate a scene graph as embedding of the input image [31].

1.3.2 Preserving the temporal dependencies

Since we are forming a model of sequence to sequence transformation, the temporal dependencies between images must be learned by the model and the sentences formed should also preserve these dependencies.

Some Proposed solution

Figure 1.3.1: **Images of a skating event. The model should learn that the finishing line in the last image is not any random event but its end of the skating event depicted in first image**

1. To use LSTM, (Long Short Term Memory) [9] units to keep track of the temporal dependencies between various images.

1.3.3 Generate meaningful sentences

Sentences formed should be coherent and relevant in context of the images. Moreover, generating semantically correct sentences with less repetitive words is a tough task. For Example:-

1. "A girl is wearing a pink dress. The girl is moving towards the car. The girl is dancing."
   In this example, 'The girl' in later sentences should be replaced with 'she' in order to make it more human-like sentence.

2. "The sun that is in west is setting in the the west that is beautiful."
   Semantically and grammatically wrong sentences should be given lower score.

1.3.4 Lack of a perfect evaluation metric

Mostly, automatic metrics like METEOR, BLEU fail to give a proper quality estimate of the sentence generated. They normally check the semantics analysis of the chunks

of words(n-gram analysis) instead of considering its entire meaning as a whole.
For Example: - **"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."**
The above machine-generated sentence has a high METEOR score of 40 [22]. But, its meaningless.
Conversely, when using some other metrics (e.g. BLEU, CIDEr) to evaluate the stories, we observe an opposite behavior: many relevant and coherent stories are receiving a very low score (nearly zero).
Some proposed Solution: -

1. Need a human judge to give rewards to the generated sentences [22].

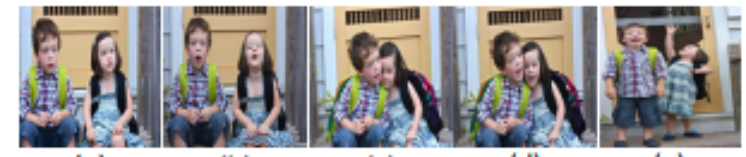
1.3.5 The publicly available dataset are biased

Recent research has found that the highly used public dataset VIST, is biased [22], such as gender bias and event bias. In the training set, the ratio of male and female's appearances is 2.06:1, and it is 2.16:1 in the test set. The Deep Learning models aggravate the gender bias to 3.44:1. Besides, because all the images are collected from Flickr, there is also an event bias issue. We count three most frequent events: party, wedding, and graduation, whose ratios are 6.51:2.36:1 on the training set and 4.54:2.42:1 on the test set. However, their ratio on the testing results is 10.69:2.22:1. Clearly, the models tend to magnify the influence of the largest majority [22].

Consequence

1. Mostly, the model prefer female filling words (like she, her) over male filling words (like he or his) whenever there are group of people and the subject is not clear.

2. More often the sentence generated have positive outcomes like (excited, happy, beautiful, merry) as the dataset has more samples of wedding and graduation which are joyful events.



(a) (b) (c) (d) (e)

Captions:
(a) A small boy and a girl are sitting together.
(b) Two kids sitting on a porch with their backpacks on.
(c) Two young kids with backpacks sitting on the porch.
(d) Two young children that are very close to one another.
(e) A boy and a girl smiling at the camera together.

Story #1: The brother and sister were ready for the first day of school. They were excited to go to their first day and meet new friends. They told their mom how happy they were. They said they were going to make a lot of new friends . Then they got up and got ready to get in the car .

Story #2: The brother did not want to talk to his sister. The siblings made up. They started to talk and smile. Their parents showed up. They were happy to see them.

Figure 1.3.2: - An automatic evaluation metrics will give more weightage to Story-1 as it has more matching words than story-2. This is a case of bias as both the story are equally true (as context is unknown).

1.3.6 Lack of imaginative and creative component in story generation

The proposed model mostly work on supervised learning and lack any imaginative and creative components. Story telling should have a personal touch in it. Then it feels more natural and real. Every story have some imaginative and hidden components which viewers apprehend by their own analytical and logical reasoning skills. Most evaluation metrics do not keep emotion of the viewer as a parameter for quality checking.

Solution

Egocentric Approach

Instead of asking a third person to annotate the images, it's better to make a dataset of images that people had annotated by themselves. Then it would have diverse emotions and model training can be in a holistic manner.

1.4 Motivation

Computer vision, together with text and language processing, has become pivotal in many disciplines. With a significant surge in the availability of multimedia content everywhere, large-scale annotated data is becoming a reality. Researchers can hire people like workers at Amazon Mechanical Turk (AMT) on paid basis to build strong datasets.

Some direct applications of these visual descriptions are understanding of images on social media platforms for better recommendations and captioning of videos on broadcasting mediums for the hearing or visually impaired. Other applications include describing signs and symbols in different levels of detail for the interpret-ability of robots in autonomous systems.

Although many works [16] have addressed the standard image and video captioning problems, the specific task of visual storytelling is a relatively new domain. Often more than not, the standard captioning models fail to interpret the non-obvious structure in the visual input. They do not account for different moments and hidden imaginative components within the image or across a given sequence of images. The VIST models are expected to generate stories with a balance between creativity and actuality of the data without loss of critical semantics.

An intelligent commercial grade application of visual story-telling will be very useful in describing social events and when integrated with audio tools can act as a strong software in social media apps like Facebook, Instagram.

Chapter 2

Literature Review

Visual storytelling belongs to the family of captioning with connections between sequences. Automatically generated descriptions can assist people with visual or hearing impairment to perceive multimedia content. Real-time closed captioning on social media and broadcasting platforms solve the problems of language barriers and improve outreach. To fully understand how the research field arrived at VIST as a task, it is essential to study into respective parent and sibling domains. This chapter details each of the related topics and respective motivations behind them. A timeline view on the evolution of the space between vision and language and, thereby, visual storytelling is provided.

2.1 Image Captioning

Image captioning is a task of automatically producing textual descriptions for given visual data. The conception of captioning as a task is often primarily traced back to success in the field of visual object detection [13]. Humans tend to perceive and learn visually, but communicate and share through text, language. This is rather evident considering many aspects of life from advertisements, social media to multimedia platforms. Therefore, the emphasis on language processing and understanding grew with the advent of deep language models utilizing neural networks [1].

Inspired by the developments in the machine translation domain, Kiros [12], further proposed an encoder-decoder architecture based on RNN to leverage the high dimensional distributed representation space. The model comprises an encoder of a CNN-

LSTM setup in which the CNN extracts image features while the long short-term memory network (LSTM) [9], encodes the textual input. These representations are then projected into a multimodal space to achieve a joint embedding. Subsequently, a neural language model, another LSTM network, reads the content vector (image or text embedding) and structure vector (part-of-speech tags) and generates each word conditioned on the auxiliary content and provided structure. The work extensively reports on the multimodal vector space emphasizing that similar underlying concepts should have similar spatial representations.

The work of Vinyal's and Xu [32] applied an attention mechanism to compensate for the dominance of any particular modality. These works also mainstreamed the adaptation of natural language measures like Bilingual Evaluation Understudy (BLEU) [20] and Metric for Evaluation of Translation with Explicit Ordering (METEOR) [14] for evaluating the automatically generated captions.

2.1.1 Baseline Architecture

From the above-mentioned developments, it is evident that a standard underlying architecture has shaped the approach towards image captioning in recent times. Hence, on a conceptual level, it would be fitting to state that all neural image captioning models follow an encoder-decoder style architecture [2], as depicted in Figure 2.1.1. Neural sequence to sequence models is another consequential term to these models. Concerning the perspective of captioning, the modules of encoder and decoder handle dependent but different objectives.

The **encoder**, which is the first component in the play, is typically a CNN owing

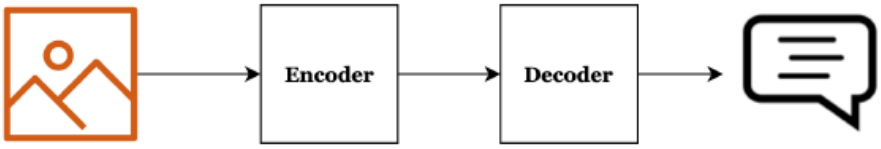


Figure 2.1.1: Illustration of encoder-decoder framework with visual input to the encoder and textual output from the decoder. Encoder and decoders are deep networks.

to their viability in successful detection and summarization of visual semantics. The

CNN would normally be a pre-trained image classifier network trained with a classification task as the objective, using datasets such as ImageNet [4]. There are a variety of popular well-trained CNNs available, such as AlexNet [13], VGG [26], and Resnet [8].

The **decoder** module of the model is typically a standard recurrent neural network. RNN being inherently sequential is a natural fit for the generation of language or text. Consequently, the decoder networks are autonomously called language models. At every time-step the decoder receives individual word vector embedding as input. Along with the words, the decoder gets a subject vector (also called a context vector) which is essentially the output of the encoder module.

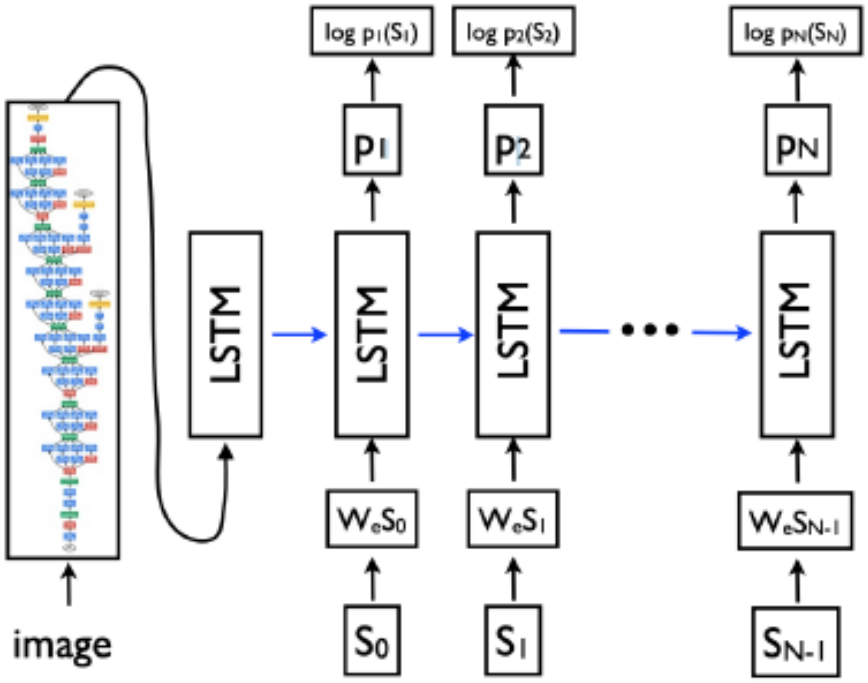


Figure 2.1.2: Model proposed by Vinyals [30]

2.2 Sequence to Sequence Captioning

Sequence captioning is a term used for referring to automated captioning tasks dealing with more than one visual. Of course, videos are the most popular sequential data

in vision, but tasks involving any form of an ordered collection of images or frames can be called a sequence, e.g.an album of photographs. In the world where photo collages, GIFs, and videos dominate a majority of today's internet, the motivation toward automatically generating descriptions for them is on the surge.

Valid object recognition is obviously at the core of the task, but aspects related to activity detection such as event inference, salient relationships, and adequately addressing diversity, influence the quality of a model. Nevertheless, there has been significant research interest that has continuously been growing towards sequence captioning and related sub-fields such as video captioning, visual story-telling.

2.2.1 Visual Story-Telling

The necessity to generate narrative style texts for image sequences that reflect experiences, rather than simple elements, has motivated the task of visual storytelling. Visuals in the input sequence typically adhere to an ascending time-frame order.

The initial work on visual storytelling used NYC and Disney image sets crawled from blog posts over the web [21]. Later, in 2016, Huang [28] formed a dataset VIST which is today widely used in research community.

Figure 2.2.1: An example sequence from VIST and a respective generated story [28]. Words like great, proud, ready can be tagged as subjective concepts within the narrative.

2.2.2 VIST Dataset

The VIST dataset includes 10,117 Flickr albums with 210,819 unique photos. The release comprises three tiers of language for the same set of images; descriptions of images-in-isolation (DII), descriptions of images-in-sequence (DIS), and stories for

images-in-sequence (SIS). Shortlisted albums were from Flickr, which had storyable events, like John's birthday party, or Friends' visit. Subsequently, crowd-workers through Amazon Mechanical Turk extracted stories for grouped photo-sequences within the albums as depicted Figure 2.13. The obtained stories were post-processed by to-

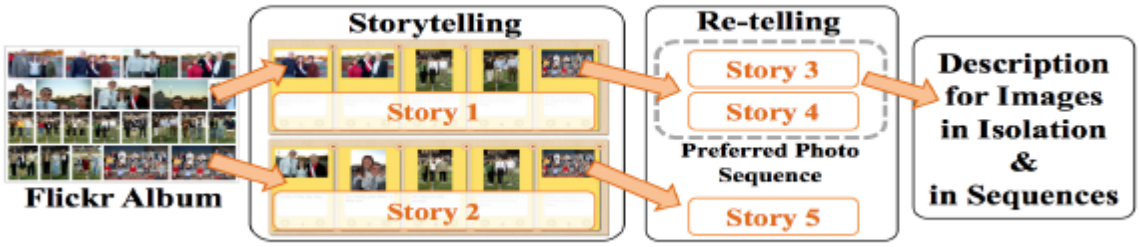


Figure 2.2.2: Dataset crowd-sourcing workflow of the VIST dataset from [28]. For each album two workers perform storytelling and three workers perform retelling on the photo-sequences selected in the storytelling phase.

kenizing them using the CoreNLP toolkit [18] to replace people's names, specific locations, and other identifiers with generic tokens. Eventually, the final data release comprised training, validation, and test splits following 80%, 10%, 10% proportions, respectively. The DIS data tier uses the same procedures and interfaces with an additional instruction for the workers to follow MS COCO [17] description styles, like to describe all the essential parts". The DII data tier leverages the complete MS COCO captioning interface. In the SIS data tier, each sequence has five images with corresponding descriptions, which together make up for a whole story. Furthermore, for each Flickr album, there are five permutations of a selected set of its images.

In the overall available data, there are 40,071 training, 4,988 validation, and 5,050 usable testing stories.

2.2.3 Evaluation Measures

Human evaluation of automatically generated or translated text is the gold standard for judging the robustness of a model. However, it is impractical, primarily owing to the expensiveness of human labor. Additionally, human judgment is not reusable or generalize to minor perturbations in use-cases.

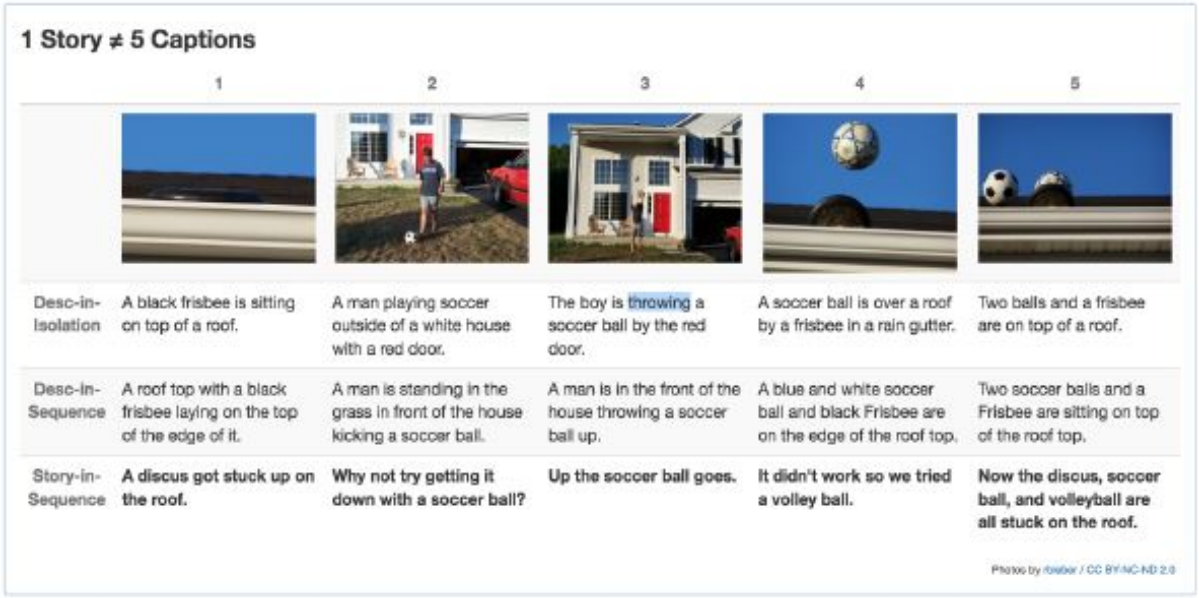


Figure 2.2.3: Sample descriptions of the DII, DIS and SIS language tiers of the VIST data-set from [28]

1. **BLEU** (Bilingual Evaluation Understudy Score)

   To handle this bottleneck, Papineni et al [20] proposed the BLEU metric focused on evaluating machine translated hypotheses. The rationale behind is to mimic the human way of judging the relevance of a sentence, given the expected true sentence. Considering the candidate translation length c and reference corpus length r, BLEU score is computed.

   Nevertheless, certain shortcomings of the BLEU metric surfaced overtime. Importantly, BLEU does not account for recall, which might result in misleading scores.

2. **METEOR** (Metric for Evaluation of Translation with Explicit ORdering)

   To address the challenges with BLEU, METEOR [14] automatic evaluation metric was conceived. Alignments between words and phrases in METEOR are based on the stem, synonym, or paraphrase matching between hypothesis-reference pairs.

It accounts for the correlation of longer matches by considering the ratio of contiguous chunks to unigrams.Additionally, the metric provides parameters for handling punctuation, tokenizing and weighting modules. These options are customizable to reflect several human biases for tasks such as machine translation and captioning.

3. **CIDEr** (Consensus-based Image Description Evaluation)

   It is a novel paradigm for evaluating image descriptions that uses human consensus. CIDEr [29], considers the vocabulary of the overall corpus during evaluation and give better results most of time.

Chapter 3

Comparison of existing models

3.1 Visual Story Telling Baseline

Huang et al [28] mainstreamed the domain of visual storytelling with their work. They released the VIST (visual storytelling) data-set (detailed in Section 2.3) which is the first and only full-edged dataset available for the visual storytelling task till date. Along with the dataset, this work presented various measures and results of baseline experiments on the task. The primary intent was to introduce the problem statement of visual storytelling rather than solving it. It elaborated how visual story telling is different from visual captioning. Therefore they proposed a sequence-to-sequence recurrent neural network (RNN) architecture shown in Figure 3.1, extending the single-image captioning technique of [5] and [30] to multiple images.

The encoder module reads the image sequence features extracted using a pretrained convolutional neural network based feature extractor. The images in the sequence were read in the reverse order. The authors do not explicitly provide reasoning behind such reversal, but it can be seen as a way to inculcate a futuristic dependency between events of the sequence. The features then sequentially pass through the RNN, yielding a context-vector Z as shown in Figure 3.1.1. The context vector is then passed both as the initial hidden state and as the first input to the decoder RNN module by concatenating it with the ¡start¿ token embedding. The decoder module then learns to produce the story word-by-word, at every time-step. Gated Recurrent Units (GRU) [3] were used for both the image-sequence encoder and the story decoder.

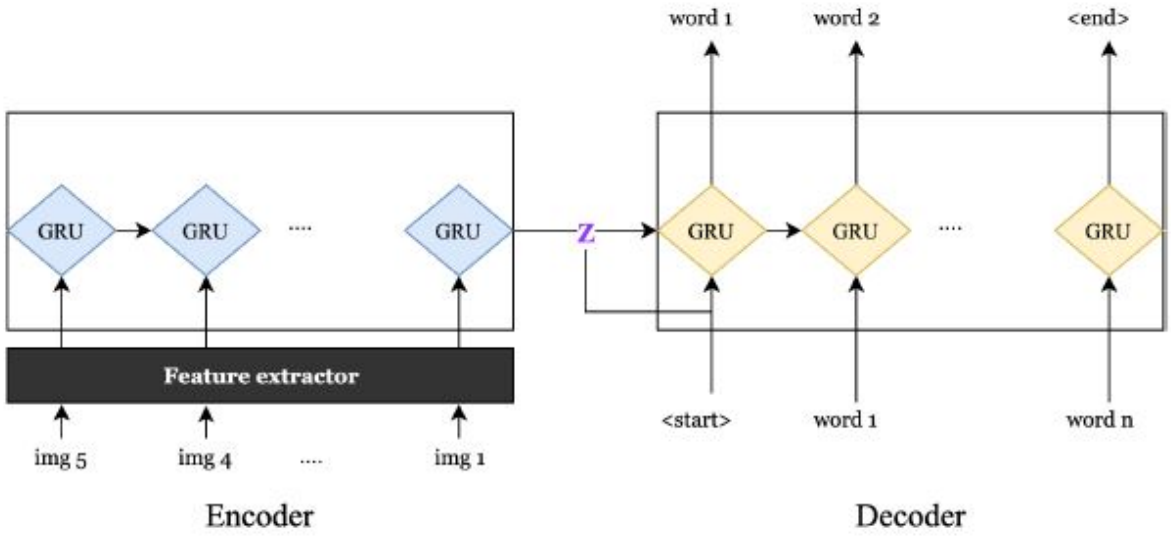


Figure 3.1.1: Visual storytelling baseline architecture (inferred based on the description in [28]).

3.2 Adversarial Reward Learning (AREL)

The words in the vocabulary are considered as classes and employing a multi-class log-loss function like cross-entropy for accessing the predictions has its implications. Fundamentally, the function rewards or punishes the model solely based on the probability of correct classes, by design. The loss value calculated is completely independent of the remaining probability split between the incorrect classes. Training mechanism based on such a criteria works very well for mutually exclusive multi-class classification scenarios like image classification [25]. However, for the use-case of visual storytelling or rather image captioning, the correctness of the output is subjective in nature.

Drawback of Baseline model:

The training of baseline model based on cross entropy criterion tries to maximize the likelihood of the observed stories ans suffer from exposure bias. Another problem with modeling using the methods the above method is that the training and testing phases are asynchronous in terms of their driving objectives. Cross-entropy loss trains the model but NLP metrics like METEOR test the quality of the trained model. To

Figure 3.1.2: Image sequences and corresponding stories generated by the Visual storytelling baseline [28]

address the problem of exposure bias and the state of asynchronicity an optimization approach called self-critical sequence training (SCST), based on reinforcement learning was proposed [23].

Nevertheless, the self-critical loss criterion and related variations utilize a hand-crafted scorer (like METEOR) for rewarding and optimizing the model. Although this approach solves the issues with cross entropy, it brings with it the implicit limitations of automatic evaluation metrics which prevent the model to learn more intrinsic semantic details. Therefore handcrafted methods are either too biased or too sparse to drive the search for optimal policy [22].

AREL architecture

An Adversarial REward Learning (AREL) framework is proposed to learn an implicit reward function from human demonstrations, and then optimize policy search with the learned reward function. Specifically, a Boltzmann distribution is first incorporated to associate reward learning with distribution approximation and then adversarial process is designed with two models – a policy model and a reward model. The policy model performs the primitive actions and produces the story sequence, while the reward model is responsible for learning the implicit reward function from human demonstrations. The learned reward function would be employed to optimize the policy in return.

For evaluation, both automatic metrics and human evaluation is conducted but observe a poor correlation between them.

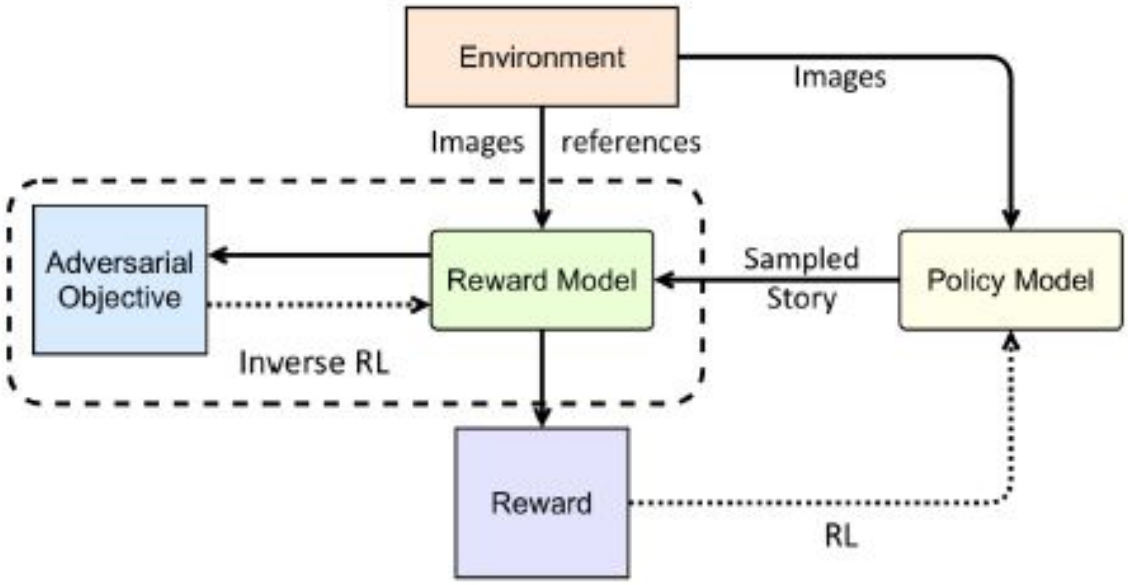


Figure 3.2.1: Adversarial Reward Learning framework [22].

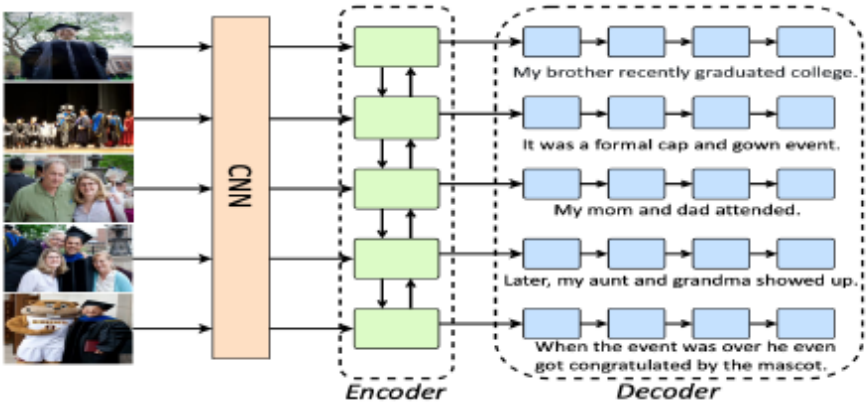


Figure 3.2.2: AREL policy model [22].

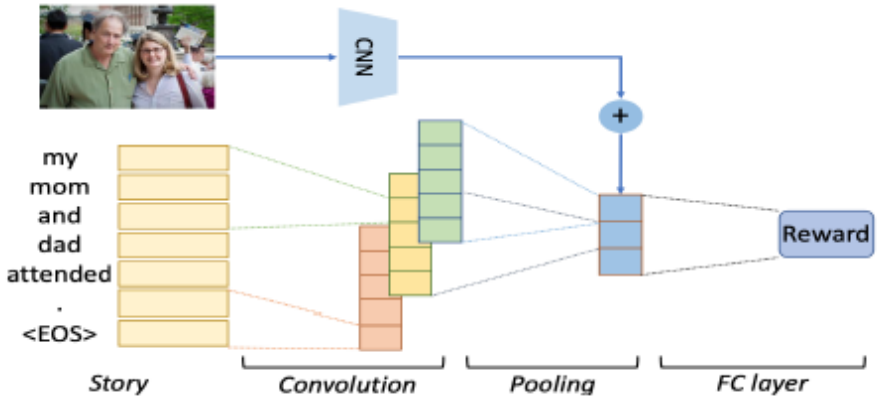


Figure 3.2.3: AREL reward model [22].

3.3 Diverse and Relevant Visual Storytelling with Scene Graph Embedding

The problem in automatically generated stories for image sequences is that the proposed models use overly generic vocabulary and phrase structure and fail to match the distributional characteristics of human-generated text. This problem is addressed by introducing explicit representations for objects and their relations by extracting scene graphs from the images [10]. A scene graph is a symbolic representation of structural information where entities are nodes and their relations are edges. Utilizing an embedding of this scene graph enables the model to reason more accurately over objects and their relations during story generation, compared to the global features from an

object classifier used in previous work. Metrics that account for the diversity of words and phrases of generated stories as well as for reference to narratively-salient image features are incorporated. Experiments on scene graph embedding also indicate that the model obtained competitive results on reference-based metrics.

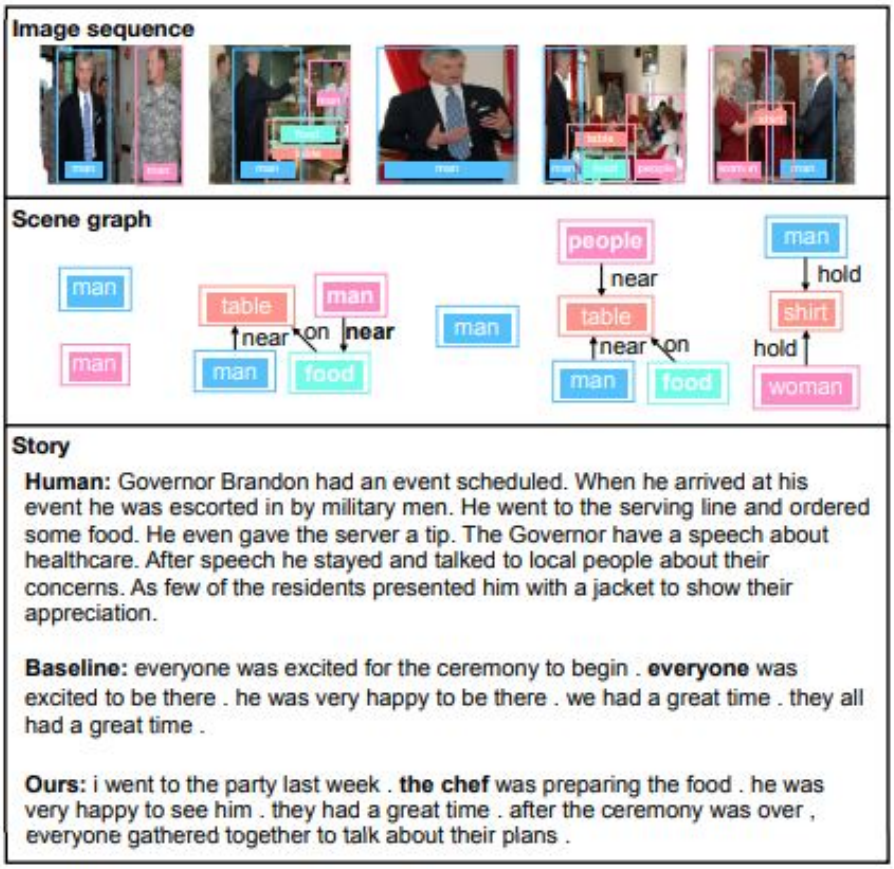


Figure 3.3.1: Example of extracting scene graphs from images and their relationship to content words and phrases in the stories. The first story (Baseline) is generated by AREL [22]. The second story (Ours) is generated by our proposed model [10]. The Human story comes from the VIST data-set [28].

The author performed the first fine-grained analysis on the distributions of words and phrases in generated stories which showed that scene graph embedding increase word and phrase diversities and bring the distributions closer to that of humans. They have also shown that the diverse noun phrases the model generate are more relevant to the objects in the images.

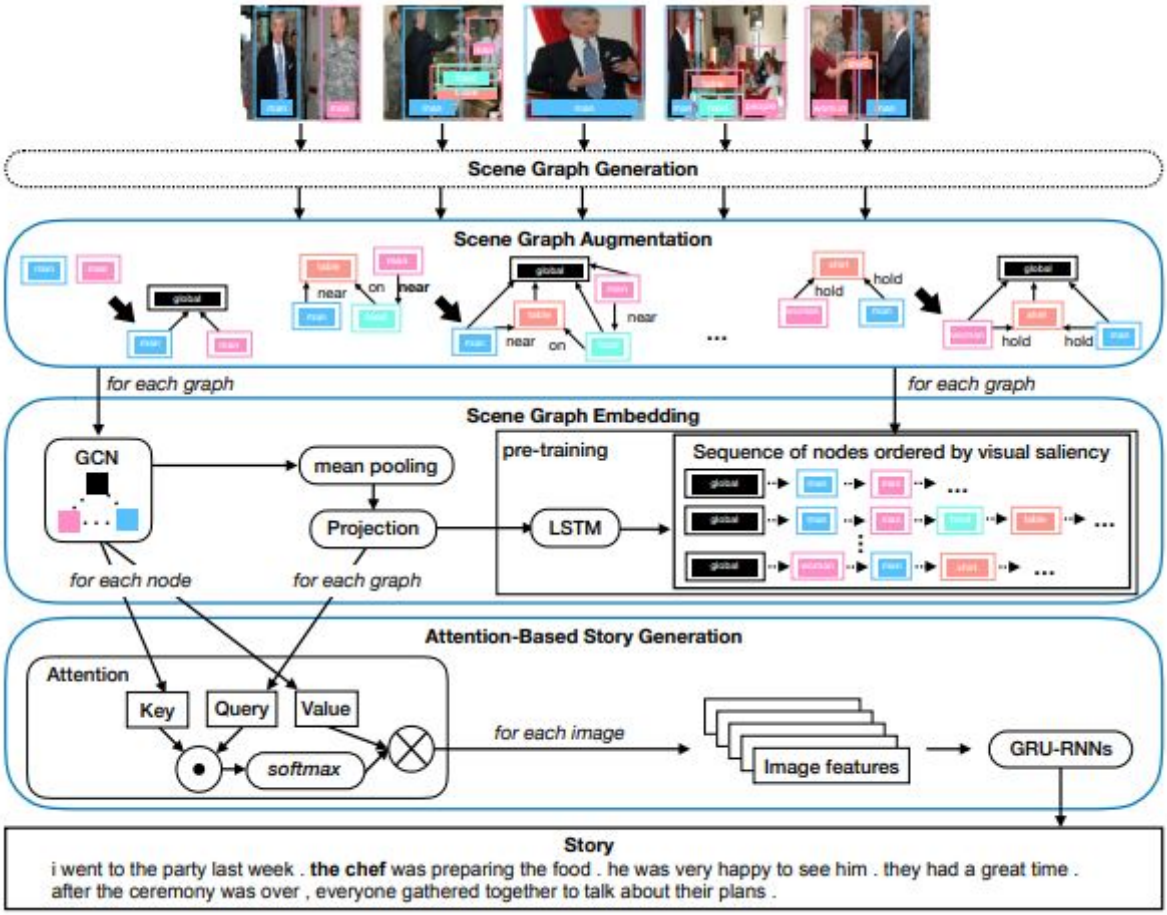


Figure 3.3.2: The pipeline used in Visual Story Telling using scene graph [10].

Chapter 4

Proposed Work

4.1 Reversing Text to Image Model

A model for the generation of image sequence from a sequence of input sentences was proposed by Rohit Agarwaal [Auto-CGAN: A Hybrid Learning Framework for Correlated Image Generation from Texts]. The framework is referred to as Auto-CGAN. It is a combination of auto-encoders and conditional-GAN that aims at learning these relations between the images and sentences. A modular approach was adopted to train the individual module. Interconnected LSTM layers had been used to implement the auto-encoders module. It captured the relations among the sentences in a story. A smaller dimensional representation of a sentence was generated using the encoder. The encoded representations acted as conditions for the generation of images. Based on each condition, the CGAN module generated images that were having a similar high dimensional distributions of the training set. GANs was used for generation of natural images that were hard to be distinguished from real photos. The conditional-GAN (CGAN) was built using CNNs for generation of images.

My contribution

I tried to take the same model and reverse the input as image and output as text. The main difficulty in this approach was the model used Universal Encoder to get the vectorized representation of the text paragraph. But, in case of images this task

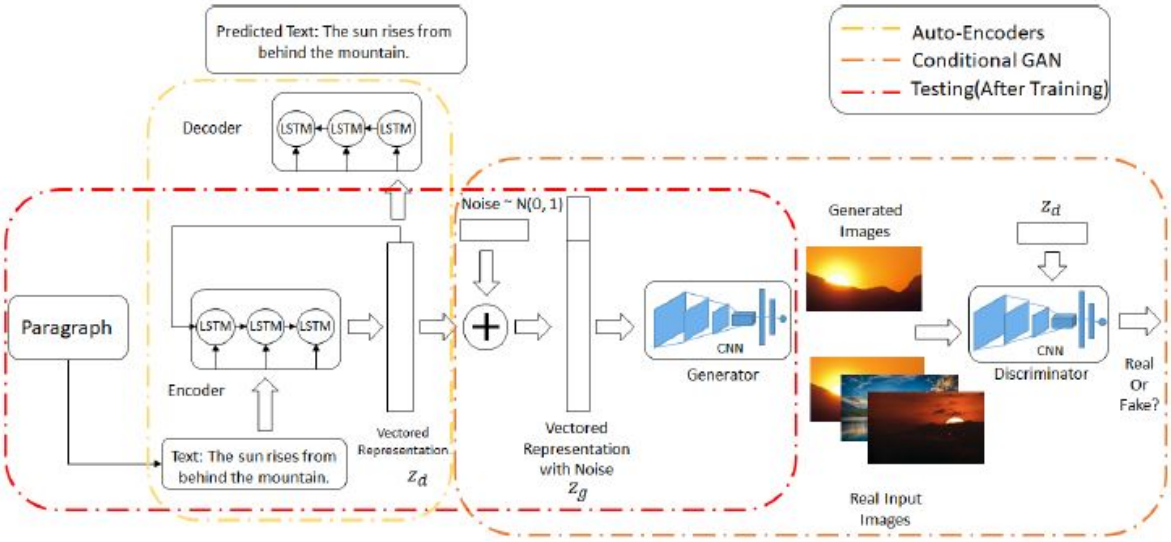


Figure 4.1.1: The pipeline used in Auto C-GAN model for text to image generation.

is not so straight-forward. Images have various components, objects and orientation and embedding them into a vector is not easy.

4.1.1 Approach

I proposed a non-deep learning model that uses the approach of Bag of Visual Words [11]. The SIFT (Scale-invariant feature transform) [7], a computer vision algorithm to detect, describe, and match local features in images was proposed to extract the feature vector.

I presumed this will be a good representation of the input image and after getting vectorized representation of five images, we can fed it as five different time-step to LSTM model to get the temporal dependencies between the image sequence. At end, the latent vector could have been reduced using PCA (Principal Component Analysis) [24] to a lower dimension space and used for traing C-GAN(Generative Adversial Neural Network) module [6, 19].

Drawback of the proposed model

1. SIFT is an old approach of image-classification, so perhaps it wont be able to correlate intra-image dependencies like postures, orientation of objects.

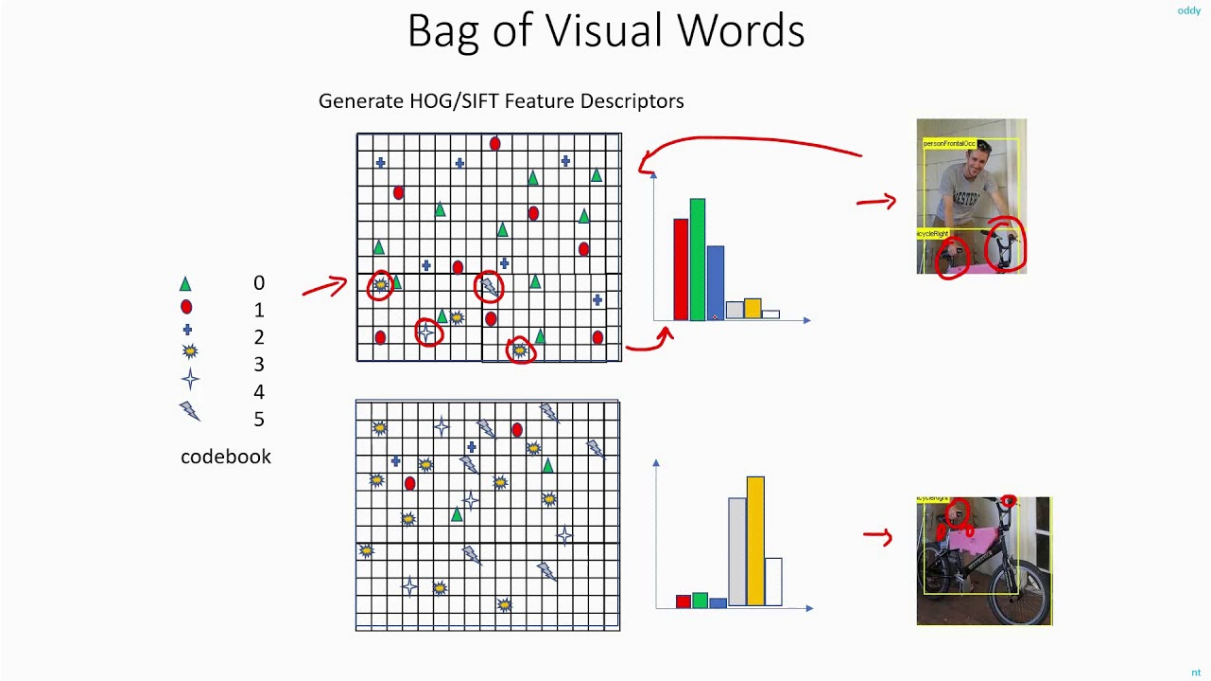


Figure 4.1.2: Bag of Visual Words technique

2. SIFT algorithm are quite slow, costs long time, hence cannot be used in responsive applications like robotics industry

3. The GAN model may not always give best result. Training a GAN is a costly affair and other better models might be used.

Keeping, all these points in mind we shifted our attention in performing a comprehensive analysis and found that scene graph embedding is the latest and state-of-the-art model.**A deep neural network model will learn subtle features and can give better results.**

4.2 Rethink of VIST Dataset

As discussed in section 2, VIST dataset was created by outsourcing the collected images to Amazon Mechanical Turk(AMT) workers for annotation. Thus, the data-set suffer from severe bias.

The VIST dataset is not perfect. It has some inherent flaws like character bias, baseless abstract words, and limited size vocabulary [27]. Nevertheless, being the only

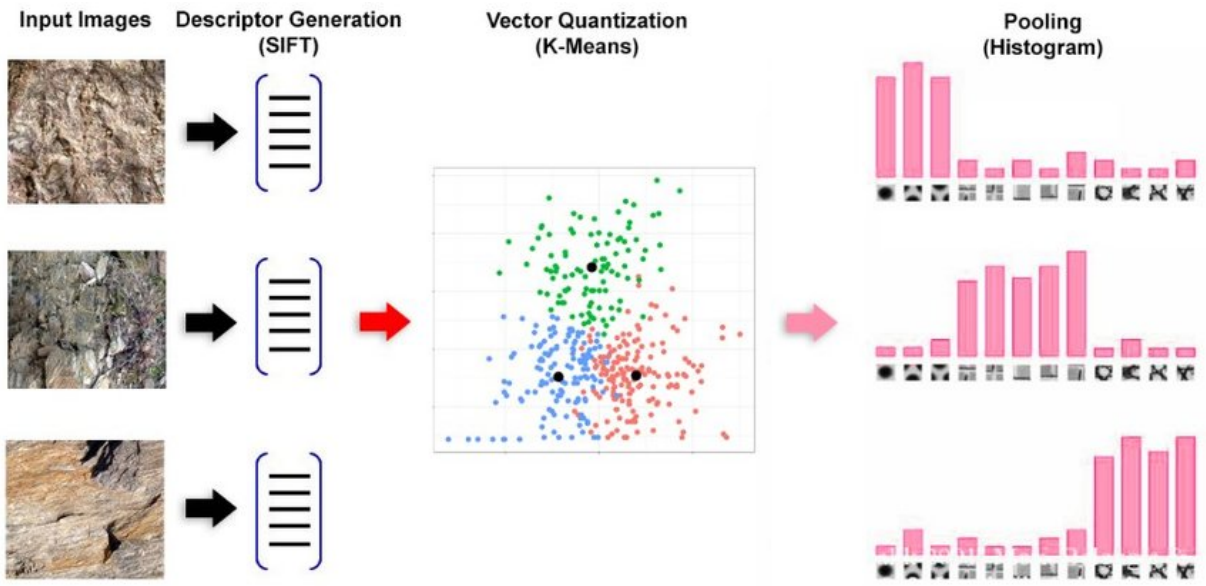


Figure 4.1.3: The bag-of-visual-words based image representation. Here, we can clearly see that 3 cluster formed represents 3 different feature and can be a used as a vectorized representation of input image

available straightforward dataset for the storytelling task, it is understandable that many published models trains on this dataset.

Mostly, in training phase, uncorrelated images are given in a single sequence, in leu to make the make the model more robuts and learn hidden distribution in image space. But, this contradicts the basic purpose of the problem statement, that we need to work on correlated images.

The VIST dataset has high event bias with mainly having higher occurrence of graduation, wedding and picnic events. Apart, from that the male to female ratio of text in the dataset is highly biased with nearly 2.01:1 (female:male) in training set and 2.16:1 in test set [22]. All these problems are described in section 1.3.4 and section 1.3.5.

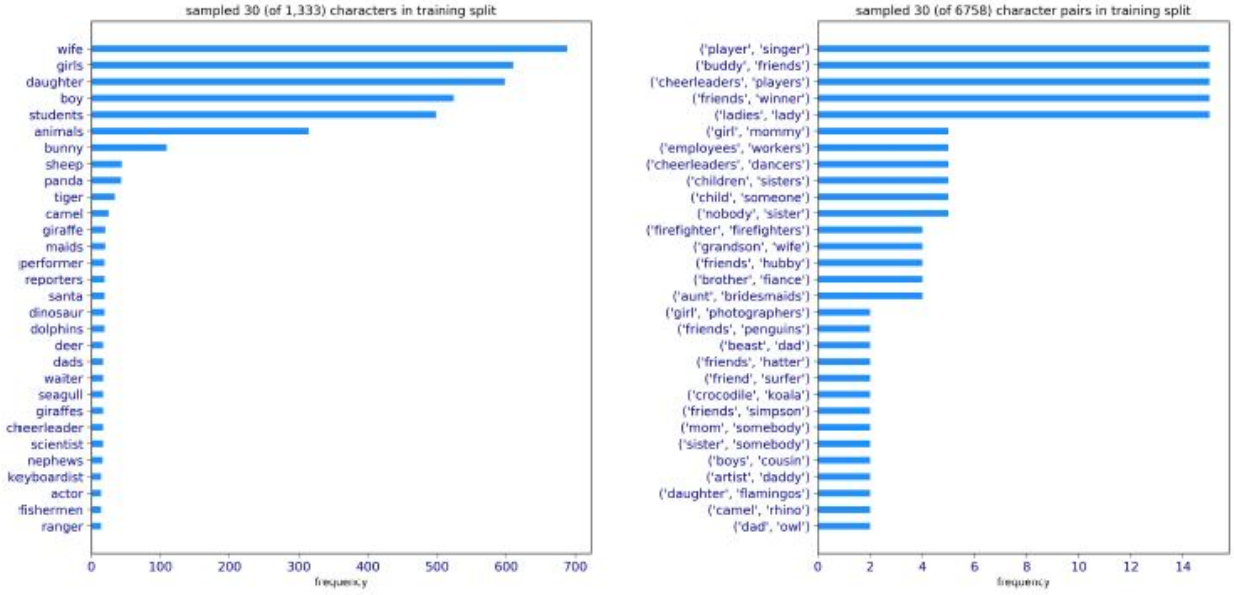


Figure 4.2.1: Training split character frequencies (left) and characters co-occurrence frequencies (right). This shows that the VIST dataset has high character pairing bias like "dad-mom" , "friend-friend" [27].

4.2.1 Solution

Create a new dataset, by asking people to give a series of events and annotate them by themselves rather than asking a third person to annotate for them.

Advantages

1. We can add vivid range of emotions and diversity to our data-base.
2. Egocentric approach is used where people self annotate their image sequences, hence, we can incorporate emotions easily [15].
3. We can train other state-of-the-art models on our new dataset and compare it with the VIST results.

Chapter 5

Future Work

In current work, we are trying to understand various state-of-the-arts models for visual story telling and find out where can we do better to gain advantage. We found, VIST dataset is a common link in all the papers and the dataset is biased [22]. In, next 4 months period we are going to create a new dataset and try to train the same state-of-the-art models on our dataset and analysis how much performance boost we can bring. Our plan is to do a comprehensive comparison between VIST and our new dataset.

References

[1] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. In *J. Mach. Learn. Res.*, 2000.

[2] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.

[3] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[5] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell. Language models for image captioning: The quirks and what works, 2015.

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks, 2014.

[7] F. Guo, J. Yang, Y. Chen, and B. Yao. Research on image detection and matching based on sift features. In *2018 3rd International Conference on Control and Robotics Engineering (ICCRE)*, pages 130–134, 2018.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.

[9] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997.

[10] X. Hong, R. Shetty, A. Sayeed, K. Mehra, V. Demberg, and B. Schiele. Diverse and relevant visual storytelling with scene graph embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 420–430, Online, Nov. 2020. Association for Computational Linguistics.

[11] H. Kato and T. Harada. Image reconstruction from bag-of-visual-words. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[12] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models, 2014.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

[14] A. Lavie and A. Agarwal. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[15] N. Li, B. Liu, Z. Han, Y.-S. Liu, and J. Fu. Emotion reinforced visual storytelling. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, ICMR '19, page 297–305, New York, NY, USA, 2019. Association for Computing Machinery.

[16] S. Li, Z. Tao, K. Li, and Y. R. Fu. Visual to text: Survey of image and video captioning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 3:297–312, 2019.

[17] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft coco: Common objects in context, 2015.

[18] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[19] M. Mirza and S. Osindero. Conditional generative adversarial nets, 2014.

[20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[21] C. C. Park and G. Kim. Expressing an image stream with a sequence of natural sentences. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[22] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning, 2017.

[23] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning, 2017.

[24] S. Sehgal, H. Singh, M. Agarwal, V. Bhasker, and Shantanu. Data analysis using principal component analysis. In *2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom)*, pages 45–48, 2014.

[25] X. Shen. A survey of object classification and detection based on 2d/3d data, 2019.

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[27] A. Surikuchi. Visual Storytelling: Captioning of Image Sequences. Master's thesis, Aalto University. School of Science, 2019.

[28] Ting-Hao, Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, C. L. Zitnick, D. Parikh, L. Vanderwende, M. Galley, and M. Mitchell. Visual storytelling, 2016.

[29] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation, 2015.

[30] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator, 2015.

[31] Q. Wu, C. Shen, A. van den Hengel, P. Wang, and A. Dick. Image captioning and visual question answering based on attributes and external knowledge, 2016.

[32] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention, 2016.