E0-399 Research in Computer Science
Indian Institute of Science - Bangalore

A survey on Visual Storytelling

Mohit Soni[1], **I.B.V.Lakshmana Raju**[2]
Department Of Computer Science & Automation
mohitsoni@iisc.ac.in, bhargavav@iisc.ac.in

Abstract

Visual storytelling aims to automatically generate a human-like short story given an image stream.The main difficulty is how to generate image specific sentences within the context of overall images. In this work, we survey relevant work to date, and conduct a thorough analysis of different approaches to visual storytelling.

1.Introduction

Artificial intelligence continues to evolve, making it increasingly plausible to develop models that interpret vision and language in a humanlike 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 refer to this task as *visual storytelling*: the generation of a cohesive, sequential set of natural-language descriptions across multiple images [1]. Visual storytelling is distinct from image captioning in that the text generated is oftentimes subjective, hinges on contextual image order, and typically employs more abstract and dynamic terms. We illustrate the dichotomy between the two more concretely in terms of possible sets of sentences

Sentence Set 1: (1) *A woman looking at a collection of tribal masks on the wall.* (2) *Three skulls of varying sizes ordered from largest to smallest.* (3) *A top view of a book about mythical creatures.* (4) *Three people standing in a store looking at the products.* (5) *An old traveling wagon that is on display.*

Sentence Set 2: (1) *I went to the natural history museum today.* (2) *Their evolution display was very interesting.* (3) *They had an area for cryptozoology.* (4) *They also have a gift shop.* (5) *My favorite was this real covered wagon from 200 years ago.*

The first is a set of traditional image captions, whereas the latter represents a visual story. Note that the former presents factual descriptions of the images in isolation from one another. The latter also describes the images, but places stronger emphasis on the development of a cohesive narrative underlying the image sequence.

High-performing visual storytelling approaches will enable growth for a variety of applications, many of which

Figure 1: A sequence of images from the VIST dataset.

are associated with language understanding tasks. They may also hold promise as a tool for assistive technology. For instance, it is relatively common for users to upload large photo albums to social media platforms without including any image descriptions at all, making these albums inaccessible to those with sight impairments. Visual storytelling could bridge this gap by automatically generating descriptive narratives for these albums.

Despite recent interest in visual storytelling, this research area is still quite nascent. Such an analysis is necessary to spur additional research and recommend directions for future work. Here, we fill this void, making the following contributions:

- We catalogue existing models for visual storytelling, comparing and contrasting them with one another.
- We provide a performance comparison based on the original results (when publicly available).

We discuss relevant prior work in Section 2, and describe the dataset used for visual storytelling tasks in Section 3. In Section 4 we present an overview of the models included in our analysis, and in Section 5 we explain how these models were evaluated. We compared the corresponding results in Section 6. We summarize these sections and report our final conclusions in Section 7.

2.Related Work

Vision+Language has been an active area of research for many years, addressing tasks such as image/video caption-

ing, paragraph generation, and visual question answering. We briefly review those related areas in the following.

2.1 Image Captioning

Barnard et al. [26] first explored annotating images with text. Since then, image/video captioning has seen a surge of research activity. Initial work utilized pre-trained image embeddings from a CNN network. The success of attention mechanisms for language translation quickly transferred to image captioning as well [27]. Later work leveraged advances in object detection and proposed a bottomup/top-down attention approach to attend to specific objects in the image instead of fixed spatial regions [28]. Further improvements include modeling spatial and semantic interactions with graph neural networks and transformers [29, 30, 31]. A key factor for improving model performance is pre-training on large amounts of data and fine-tuning on curated supervised datasets [32, 33, 34]. Recently, multimodal pre-training models like CLIP are enabling large-scale pre-trained LMs to be guided in a zero-shot fashion [35]. As opposed to captioning images, visual storytelling aims to ground a visual story from multiple cues. This means that only objects related to the narrative should be highlighted, which is the focus of our work.

2.2 Story Generation from images

In the first work for image cued sentence generation (Farhadi et al., 2010), the triplet - $< object, action, scene >$ was predicted for an input image using MRF, and used for searching or generating with templates. In the deep learning era, Jain et al. (2017) utilized the VIST dataset to translate description-to-story without images. Liu et al. (2017) developed semantic embedding of the image features on the bi-directional recurrent architecture to generate a relevant story to the pictures.

3.Data

Most visual storytelling work to date has been trained and evaluated using the VIST Dataset [1]. VIST is the first publicly available dataset for sequential vision-to-language tasks, and consists of sequences or "albums" of images wherein each image is paired with two types of captions; namely, descriptions of images in isolation (DII), and stories of images in sequence (SIS). The images were originally downloaded from Flickr (https://www.flickr.com/). In total, the dataset comprises 10,117 Flickr albums containing 210,819 unique photos.

Amazon Mechanical Turk (AMT) workers selected subsets of five images per album about which to write sequential, cohesive stories. The dataset contains 50,200 story sequences overall; these are divided into subsets of 40,155 training, 4,990 validation and 5,055 testing stories. Five written stories were collected per album. Three standalone descriptions per image (DII, first defined above) were also collected separately using the image captioning interface used to build the COCO image caption dataset [2]. In both the stories and descriptions, all people names were replaced with generic MALE/FEMALE tokens, and all named entities were replaced with their entity type (e.g., location). A small number of broken images were filtered from VIST by most research groups. For concrete examples of DII and SIS from VIST, we refer readers to Figure 1, where Sentence Sets 1 and 2 (see Section 1) are from the DII and SIS subsets, respectively.

4.Methods

We analyze different approaches for Visual Storytelling : AREL [3], GLACNet [4] ,Contextualize, Show and Tell [5] ,Storytelling from an Image Stream Using Scene Graphs [6] , Plot and Rework:Modeling Storylines for Visual Storytelling [7] , Ordered Attention for Coherent Visual Storytelling [8] , Latent Memory-augmented Graph Transformer for Visual Storytelling [9] , Knowledge-enriched Attention Network with Group-wise Semantic for Visual Storytelling [10] and Associative Learning Network for Coherent Visual Storytelling [11] . We selected these approaches as the focus of our work for reason were very recent models, representing the current state of the art in visual storytelling. We summarize every approaches in coming sections and refer readers to the original papers for fuller detail.

4.1 Adversarial Reward Learning (AREL)

AREL [3] is an adversarial reinforcement learning approach that makes use of two models: a policy model, followed by a reward model. The policy model is an encoder-decoder model utilizing a CNN-recurrent neural network (RNN) architecture, used to generate new stories. Specifically, a pre-trained CNN is fed a sequence of 5 images as input to extract high-level image features. These features are passed forward and further encoded as visual context vectors using bidirectional GRUs. The outputs of the encoder are then fed into a GRU-RNN decoder to generate sub-stories for the image sequence in parallel. The sub-stories are concatenated to form a single full story. The CNN-based reward model is applied to every sub-story to compute its partial reward, and from the input sequence embeddings, n-gram features are extracted using convolution kernels of different sizes and passed through pooling layers. Image features are concatenated with these sentence representations and passed through a fully connected layer to obtain the final reward. To perform adversarial reward learning, the models were alternately optimized using stochastic gradient descent. The objective of the story generation policy was to maximize the similarity between a Reward Boltzmann distribution and itself. The first model optimized the policy to minimize the KL divergence [12] between itself and the Boltzmann Distribution, and the second model attempted to (a) minimize the KL divergence with the empirical distribution, and (b) maximize the KL divergence with the approximated policy distribution, with the objective of distinguishing between human and machine generated stories.

Wang et al. [3] demonstrated that AREL outperforms a generative adversarial network (GAN) model, a cross-entropy model, and other baselines and achieves state-of-the-art results across both automated and human metrics. The human metrics considered included both a Turing test

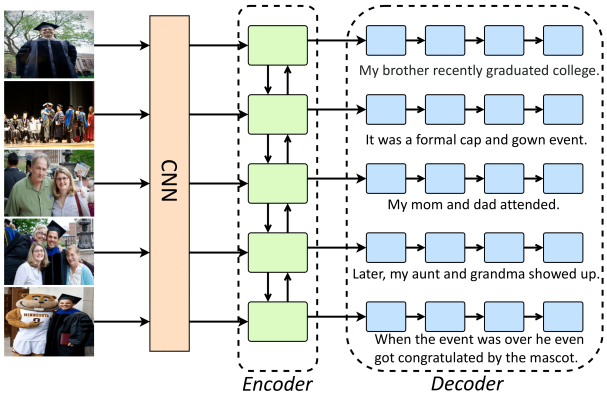

Figure 2: Overview of the policy model. The visual encoder is a bidirectional GRU, which encodes the high-level visual features extracted from the input images. Its outputs are then fed into the RNN decoders to generate sentences in parallel. Finally, we concatenate all the generated sentences as a full story. Note that the five decoders share the same weights.

(in which annotators attempted to guess which of two stories was written by a human) and pairwise comparisons measuring relevance, expressiveness, and concreteness.

4.2 GLocal Attention Cascading Networks (GLACNet)

GLACNet [4] also uses an encoder-decoder architecture, but it adds a hard attention mechanism which stresses feeding both the local image features and the overall context to the decoder as input. The image-specific features are extracted using a 152-layer residual network [13]. Those features are fed sequentially into a bidirectional LSTM, which then produces the global context vectors. The global context and local image features are combined to form *glocal* vectors and passed through fully connected layers. The output is concatenated with word tokens and fed to the decoder (LSTM) as input. Thus, five glocal vectors for each image are fed into the decoder one after another, creating a cascading mechanism by passing the hidden state of one sentence generator as the initial hidden state of the next sentence generator.

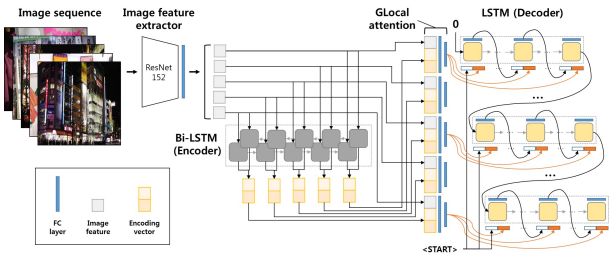

Figure 3: The global-local attention cascading (GLAC) network model for visual story generation. Note: activation function (ReLU), dropout, batch normalization, and softmax layer are omitted for readability.

To validate that all components of the GLACNet architecture contributed to the model's performance, Kim et al. [4] conducted an ablation study in which the cascading, global attention, local attention, and post-processing routines were removed one at a time, comparing perplexity and METEOR [14] scores between conditions as well as with a standalone LSTM sequence-to-sequence (Seq2Seq) model and the full GLACNet model. The full GLACNet model exhibited the best performance, and the other GLACNet-based models exhibited better performance than the LSTM Seq2Seq model, thereby verifying the utility of this approach.

4.3 Contextualize, Show and Tell

Contextualize, Show and Tell [5] won the 2018 Visual Storytelling Challenge. The model uses an encoder LSTM to read in the image representations one by one for every image in a sequence. The image representations are generated using Inception V3 [15]. Five decoders, again LSTMs, then read in the image embedding as input. The first hidden state of each decoder is initialized using the last hidden state of the encoder to provide the model with global context. Gonzalez-Rico and Pineda [5] obtained the final story by concatenating the outputs of the model's five decoders.

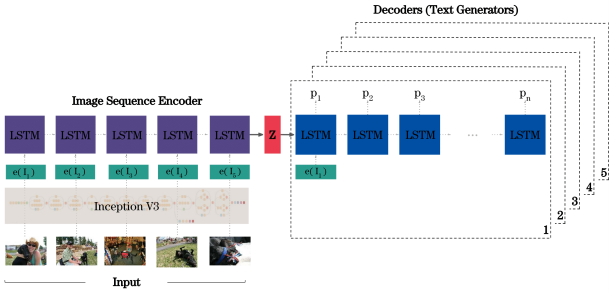

Figure 4: Proposed sequence to sequence architecture.

As part of the Visual Storytelling Challenge, the model was evaluated on public and hidden test sets using both human evaluation and an automated metric (METEOR). METEOR scores of 30.88 and 31 were obtained on the public and hidden test sets, respectively.Human evaluation scores were collected via Amazon Mechanical Turk. Crowd workers evaluated six aspects of each story using a Likert scale. Each worker was asked to indicate the degree to which: 1) the story was focused, 2) the story had good structure and coherence, 3) the worker would share the story, 4) the worker thought the story was written by a human, 5) the story was visually grounded, and 6) the story was detailed. In summing the average scores received for each criterion, Gonzalez-Rico and Pineda's [5] model achieved a score of 18.498, whereas human-generated stories achieved a score of 23.596.

4.4 Storytelling from an Image Stream Using Scene Graphs

It is a novel graph based architecture named SGVST for visual storytelling, which first translates each image into a graph-based semantic representation, i.e., scene graph, and then models the relationship on within-image level and cross-images level.Specifically, inspired by the success of

scene graph generation (Xu et al. 2017 [36]; Li et al. 2018 [37]; Zellers et al. 2018 [38]), a scene graph parser, consisting of Faster R-CNN [16] and relationship detector, is firstly implemented to parse images into scene graphs. In each scene graph, vertexes represent different regions and directed edges denote relationships between them, which can be represented as tuples $< subject - predicate - object >$, e.g., $< man - holding - girl >$, explicitly encoding the objects and relationships detected within an image.

Then for processing the scene graphs to enrich region representations, we employ Graph Convolution Network (GCN) which passes the information along graph edges. After processing the local region representations for each image, we further utilize Temporal Convolution Network (TCN) (Bai, Kolter, and Koltun 2018)[39] to process the region representations along the temporal dimension, which models relationships on cross-images level. To this end, the relation-aware representations are integrated with the information on both within-image level and cross-images level. In order to make full use of image information, we use a bidirectional-GRU (Chung et al. 2014) (biGRU) [40] to encode the feature maps obtained from Faster R-CNN as high-level visual features, and then fuse them with the relation-aware representations to get new representations. Finally, the obtained new relation-aware representations are fed into the hierarchical decoder to conduct the story generation.

The main contributions can be summarized as follows:

• They first propose to translate images into graph-based semantic representations called the scene graphs to benefit representing images and high-quality story generation.

• They propose a framework based on scene graphs to realize enriching fine-grained representations by modeling the visual relationships through GCN on the within-image level and through TCN on the cross-images level.

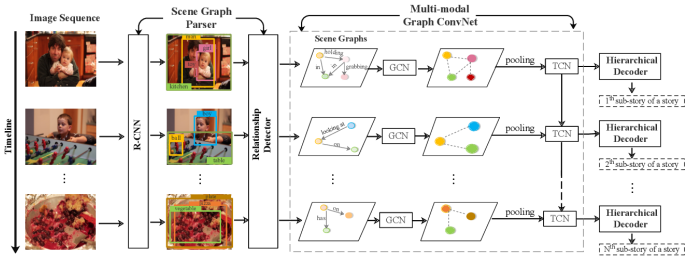

Figure 5: An overview of SGVST model (better viewed in color).

The overall architecture of the proposed model is shown in Figure 5. Here we have an image stream $I = \{I_1, ..., I_N\}$, we aim to output a story $y = \{y_1, ..., y_N\}$, where N is the number of images in the image stream and sentence $\{y_n = \{w_1, ..., w_T\}$ consisting of T words in the vocabulary V_s of all output words. We argue that modeling relationships on within-image and cross-images levels would help for understanding and describing images. To this end, we propose a graph-based architecture. First, scene graphs $G = \{G_1, ..., G_N\}$ are first generated by a pre-trained scene graph parser, where the vertex (object) represents each region and the edge denotes the visual relationship between them. Then the scene graphs are passed through Multi-modal Graph ConvNet to obtain the relation-aware representations $\bar{v} = \{\bar{v_1}, ..., \bar{v_N}\}$, which integrate both withinimage and cross-images levels information. In the story generation state, we feed the relation-aware representations $\bar{v}$ into a hierarchical decoder to generate the story.

4.5 Plot and Rework:Modeling Story lines for Visual Storytelling

The above mentioned approaches are often optimised towards predefined aspects such as image relevancy or topic coherence, which do not necessarily lead to engaging stories from a human perspective. For these challenges PR-VIST model is introduced. PR-VIST operates by constructing a graph that represents the relations between elements in an input image sequence. It then identifies the optimal path in the graph, which corresponds to the best storyline. This selected path is used to generate a coherent and meaningful story. It is done in three stages namely : Preparation(Stage 0), Story Plotting(Stage 1) and Story Reworking(Stage 2).

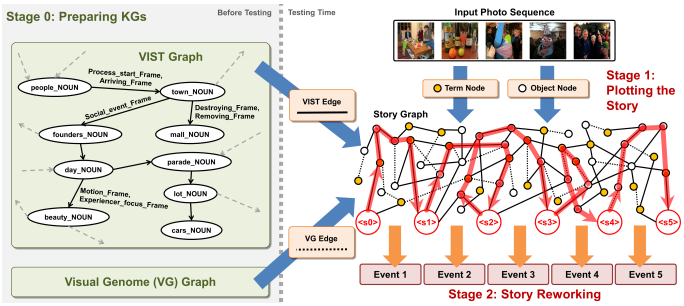

Figure 6: Overview of PR-VIST. In Stage 1 (Story Plotting), PR-VIST first constructs a graph that captures the relations between all the elements in the input image sequence and finds the optimal path in the graph that forms the best storyline. In Stage 2 (Story Reworking), PR-VIST uses the found path to generate the story. PR-VIST uses a story generator and a story evaluator to realize the "rework" process. In Stage 0 (Preparation), a set of knowledge graphs that encode relations between elements should be prepared for the uses in Stage 1.

In Stage 0 (Preparation), a set of knowledge graphs that encode relations between elements are prepared for use in Stage 1. To prepare for story plotting, it collects information from the images and knowledge from the knowledge graph. To extract information from the images, two extraction methods are used to extract image-oriented and story-oriented story elements: objects and terms, respectively representing image and story intuition. Objects can be detected by current object detection models, for which it uses a pre-trained object detection model—Faster-RCNN [16]. Terms are the story-like nouns such as events, time, and locations, which current object detection models are unable to extract. Therefore, it further uses a Transformer-GRU [17] to predict story-like terms. From this information it prepares two knowledge graphs: a VIST graph and a visual genome (VG)

graph. They construct the VIST graph based on the VIST dataset, representing in- domain knowledge; the VG graph is an existing resource [18], representing generic knowledge.

In Stage 1, PR-VIST uses a storyline predictor to find what it deems the best path in the story graph as the storyline and then pass this to Stage 2. It uses HR-BiLSTM [19] as the scoring model. This model incorporates word embeddings generated from GloVe to represent the objects. Furthermore, relation embeddings are decomposed into graphical and textual embeddings, which are then combined into a unified representation. This storyline predictor uses UHop [20], a non-exhaustive relation extraction framework. During the training process of UHop, the goal is to learn to construct a storyline by finding the best path in the golden storyline. It starts with an initial noun token entity "<s0>" and learns to select the correct relation from a list of candidate relations. In testing, UHop finds a storyline in the story graph Gstory by iteratively selecting relations based on the predicted entities, until a termination decision is reached. In Stage 2, the framework consists of two components: the story generator and the story evaluator. The story generator generates a story according to the storyline. A storyline consists of a set of events e1...eL that are input to the story generator, which is based on the Transformer [21]. The story evaluator is a discriminator trained on the MTurk human ranking data to classify good and bad stories—outputs a story quality score and modifies the loss functions.

4.6 Ordered Attention for Coherent Visual Storytelling

The goal of visual storytelling is to generate a story, composed of N *ordered* sentences $\{y_s|1 \leq s \leq N\}$, given an *ordered* sequence of images $I = \{I_s|1 \leq s \leq N\}$. Each sentence $y_s = (y_{s,0}, \ldots, y_{s,t}, \ldots)$ is composed of words $y_{s,t} \in$ from vocabulary .

The order in which the images are given is essential as it defines the plot line of the story. The story should be focused, , each sentence should be related to the remainder of the story. Importantly, the sentences should form a coherent body of text describing the set of images, and not only a set of related information. For instance, the story *"The church was beautiful. The bride and groom walk down the aisle. The cake was amazing."* is less coherent than: *"We went to the church for the wedding today. The bride and groom were excited for the day. Both cut the cake together."*

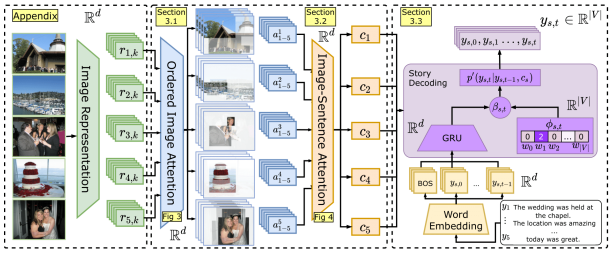

Figure 7: Architecture for Visual Storytelling synthesis

Overview: To address this challenge, They develop the model illustrated in fig:7. It infers conditional probabilities $p'(y_{s,t}|y_{s,t-1}, c_s)$ for the t-th word $y_{s,t} \in Y$ in sentence y_s given the previous word $y_{s,t-1}$ and the context embedding c_s for sentence s. The context embedding c_s summarizes region representations $r_{i,k}$ of all K object regions across all N images I_i ($i \in [1, N]$, $k \in [1, K]$) via Ordered Image Attention (OIA) and Image-Sentence Attention (ISA) . Specifically, when generating sentence s, OIA computes an attended image representation a_i^s for every image I_i by attending to the K region representations $r_{i,k}$. These attended image representations a_i^s are subsequently summarized into the context embedding c_s via an image-sentence attention .

Two types of Attention are given below :

Ordered Image Attention form a structure across ordered images and to 2) select the relevant objects per image. For this we model preceding and subsequent interactions separately using different attention factors. THey calibrate each factor's importance with trainable scalars, which forms a graph of dependencies between the images. For each sequence of N images, the model infers a total of N^2 attention maps, one per image for each sentence.

Image-Sentence Attention (ISA) : In a next step they summarize the attended image representations produced by OIA to compute the context embedding for the sentence that we wish to generate. For this they use the Image-Sentence Attention (ISA) unit. It picks the relevant image context for generating the specific sentence.

4.7 Latent Memory-augmented Graph Transformer for Visual Storytelling

Most existing works utilise either scene-level or object-level representations, neglecting the interaction among objects in each image and the sequential dependency between consecutive images. To meet these requirements in the context of the Transformer and inspired by the human way of telling a story, a novel method Latent Memory-augmented Graph Transformer (LMGT) for visual storytelling is proposed. LMGT shares the basic architecture with the Transformer, but enhances it with two carefully designed components, i.e., a graph encoding module and a latent memory unit.

Specifically, the graph encoding module encodes visual embeddings simultaneously integrated with structured semantic relationships among various image regions by constructing a scene graph. GEM consists of a graph attention layer, a feed-forward layer, as well as residual connection and layer normalisation. In this module, it inputs the node representations extracted from the object detector, and then enriches the new relation-aware node embeddings by aggregating the neighbourhood information from the scene graph. Most importantly, the module can learn to assign high attention weights to the critical semantic relations in each image, which is essential to compose an informative story. In addition, the latent memory unit is introduced to capture latent contextual clues as the story line, as well as record the previous history of images and generated sentences to preserve topic consistency and inter-sentence coherence, wherein the memory state can be updated based on the current input and previous memory.

Finally, we integrate the feature embeddings learned from both the graph encoding module and the latent memory unit,

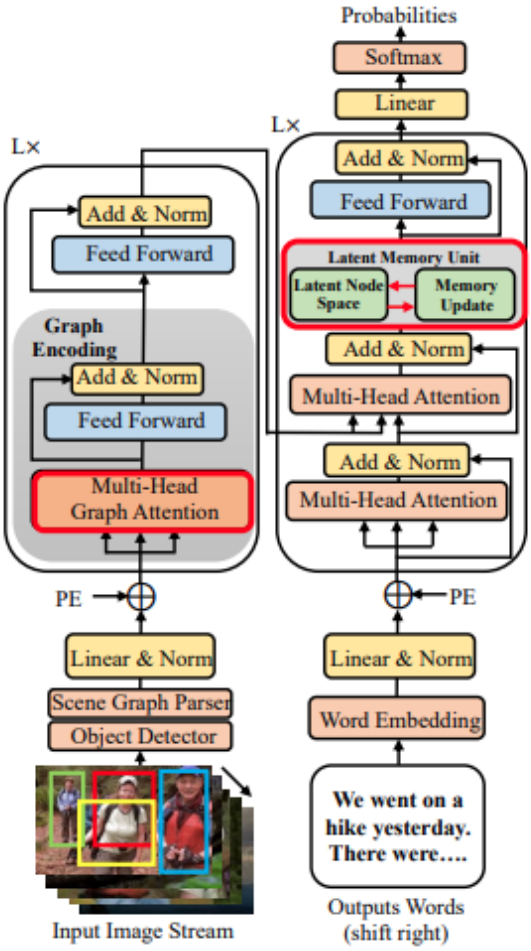

Figure 8: Overview of the proposed Latent Memoryaugmented Graph Transformer, which mainly consists of two carefully designed components marked with the red borders: a graph encoding module to obtain implicit semantic relational embeddings of input image regions based on scene graphs, and a latent memory unit to help the Transformer record the important contextual and historic information as latent memory. "PE" denotes Positional Encoding.

and then decode them into a human-like, coherent and informative story for the given image stream. The story decoder in the proposed LMGT takes the output of the encoder as input, and outputs each word of the sentence in a story. It shares a similar architecture with the encoder, containing identical self-attention blocks and an augmented latent memory unit, followed by residual connection and layer normalization. Finally, the output is achieved after applying the feed forward layer, and then fed it to the classifier to predict the next word based on the pre-defined vocabulary.

4.8 Knowledge-enriched Attention Network with Group-wise Semantic for Visual Storytelling

Above approaches encounter problems in their optimization because of memory dilution along the longer feature

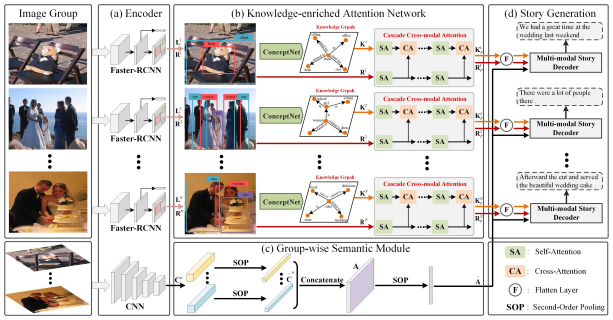

Figure 9: Pipeline of the proposed KAGS for visual storytelling. The framework contains four key components: (a) a Faster-RCNN network and a ResNet backbone to extract regional features and high-level convolutional features; (b) the proposed KAN to obtain the attentive heterogeneous representations by exploiting the intra- and inter- interactions of visual and knowledge concepts; (c) the proposed GSM to explore the global guided aggregation with a set of hierarchical second-order pooling algorithms in a convolutional feature group; and (d) a story generation that fuses the multi-modal information in a decoder to produce the final predicted sentences.

sequence, failing to generate the topic-aware information of an image stream. Nevertheless, the storyline containing long- range dependencies is crucial to output the coherent multi- sentences. Furthermore, the most serious problem among above approaches is that they are incapable of establishing a unified framework to simultaneously capture sufficient regional features and topic-aware global features for visual storytelling. To address these challenges a knowledge- enriched attention network with group-wise seman- tic (KAGS) model is proposed.

The KAGS model uses a CNN [22] and Faster-RCNN [16] as encoders to extract features from an image stream. These features include convolutional features, semantic labels, and regional object features. The model then employs a knowledge-enriched attention network (KAN) to process the semantic labels with ConceptNet and the regional features with a cascade cross-modal attention module.

Furthermore, the KAGS model introduces a group-wise semantic module (GSM) with second-order pooling (SOP) to capture long-range dependencies in the sequential convolutional features and transform them into a global guided vector. SOP actually reduces the channel dimension, computes a covariance matrix of pairwise vector similarities, and highlights meaningful feature channels.

Finally, the optimised visual and textual features, combined with the global semantic vector, are used by a multimodal story decoder to generate the story. Multi-model story decoder flattens the regional-visual and knowledge indicator vectors to obtain relevant information and incorporates them into the decoding process. Then it explores the contextual representation to generate coherent and reasonable sentences.

4.9 Associative Learning Network for Coherent Visual Storytelling

Although visual storytelling methods have made promising improvement in recent years, existing methods pay little attention to the association ability and divergent thinking of the model, which are essential for humanistic stories. To overcome these challenges, a novel method Associative Learning Network for Coherent Visual Storytelling is introduced to explore the model's association ability while telling a new story.

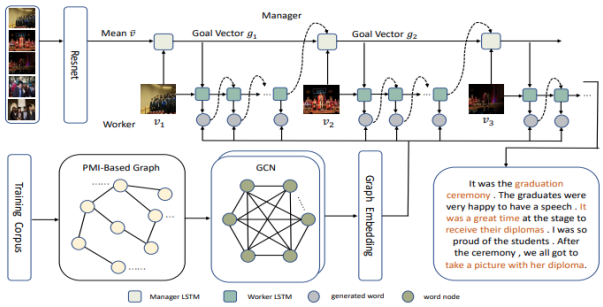

Figure 10: Overview of the proposed model. In the preparation stage, we first leverage the point-wise mutual information to measure the association degree of word pairs and exploit graph convolutional network to aggregate neighbor information. Image features are extracted by pre-trained CNN and we integrate graph embedding to the hierarchical decoder to generate story.

This method firstly builds a word co-occurrence graph based on pointwise mutual information(PMI) to capture the association degree of the word pair. Specifically, the node information in the graph is the number of unique words in the corpus and the edge information is the pointwise mutual information. PMI is basically the word co-occurrence based measure to calculate weight between two word nodes.The assumption behind PMI is that if two words co-occur more than expected under independence there must be some kind of association between them. Secondly, it exploits the graph convolutional network to encode and update the representation of the nodes using the features of neighbours. Through the convolution of GCN, the model possesses the associative ability to acquire information about relationships between events or entities in arbitrary environments. In the Encoder stage, it leverages the ResNet as our image encoder to learn the deep visual features from each image. Then, mean pooling is deployed to generate an image-sequence content vector. Lastly, the model incorporates useful story elements related to visual contents to the hierarchical decoder to generate coherent stories. The language decoder in the model is a two level manager-worker decoder which holds two variants of LSTMs. The manager LSTM aims to capture the consistency of the story at the sentence-level. This module considers three aspects namely : 1) the overall features of the image sequence 2) the image feature of the ith image and the previous decoding output. After the manager LSTM produces a goal vector, the worker LSTM then predicts one word at a time and controls the fluency of one sequence. The worker also takes three kinds of information into account, which are: 1) the image feature of the ith image 2) the goal vector from the manager and the word embedding. Then the worker LSTM exploits a linear layer to approximate the probability to choose the word. Finally, the sentences are assembled to generate the required story.

5. Metrics for Evalution

Common metrics for evaluating visual storytelling models include METEOR [14], BLEU [23], CIDEr [24], and ROUGE-L [25]. METEOR, the primary metric considered in the Visual Storytelling Challenge, calculates the alignment between the machine-generated hypotheses and the reference stories based on the exact, stem, synonym, and paraphrase matches between words and phrases. While AREL was evaluated using METEOR as well as the other metrics, GLACNet was evaluated using only METEOR scores and measures of perplexity. Contextualize, Show and Tell was also evaluated using only METEOR. SGVST, LMGT , KAGS , Ordered Attention for Coherent Visual Storytelling , Associative Learning Network for Coherent Visual Storytelling was evaulated for all metrics presented above.

6. Results

We have made comparison of all the above novel methods discuss in this report on basis already experimented results. The comparison table (Table 1) is shown below. As can be observed, the stories generated by LMGT are more informative, human-like and closer to the ground-truth compared with the results of other models. This is because LMGT successfully maintains the globally coherent story line, so that the generated sentence of each image is more consistent with the main topic as well as more coherently associated with the preceding/following sentences. For BLEU-1 and BLUE-3, it can be seen that KAGS model is having a better results.

7. Conclusions

In our survey report on visual storytelling, we extensively researched various papers on the topic. Our findings indicate that models incorporating important interactions among visual objects in the form of graphs tend to yield superior results compared to models that do not consider such interactions. This suggests that capturing and leveraging the relationships between visual elements significantly contributes to the effectiveness of visual storytelling models.

References

[1] Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Ja- cob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Compu- tational

Model	METEOR	CIDEr	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
AREL-s-50	34.9	9.1	29.4	62.9	38.4	22.7	14.0
GLACNet	30.14	-	-	-	-	-	-
Contextualize, Show and Tell	34.4	5.1	29.2	60.1	36.5	21.1	12.7
SGVST	35.8	9.8	29.9	65.1	40.1	23.8	14.7
PR-VIST	31.6	-	-	-	-	-	7.65
Ordered Attention for Coherent Visual Storytelling	36.8	10.1	30.2	68.4	42.7	**25.2**	15.3
LMGT	**37.2**	**12.9**	**32.8**	67.5	41.6	25.0	**16.7**
KAGS	36.2	11.3	31.4	**70.1**	43.5	**25.2**	14.7
Associative Learning Network for Coherent Visual Storytelling	36.2	12.0	29.8	69.3	**44.0**	24.7	13.7

Table 1: Comparison of all methodologies on above metrics

Linguistics: Human Language Technolo- gies, pages 1233–1239. Association for Computational Linguistics.

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dolla r, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In Computer Vision – ECCV 2014, pages 740–755, Cham. Springer Inter- national Publishing.

[3] Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. 2018. No metrics are perfect: Adversarial reward learning for visual storytelling. In Proceedings of the 56th Annual Meeting of the As- sociation for Computational Linguistics (Volume 1: Long Papers), pages 899–909. Association for Com- putational Linguistics.

[4] 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. CoRR, abs/1805.10973.

[5] Diana Gonzalez-Rico and Gibran Fuentes Pineda. 2018. Contextualize, show and tell: A neural visual storyteller. CoRR, abs/1806.00738.

[6] Wang, R., Wei, Z., Li, P., Zhang, Q., Huang, X. (2020). Storytelling from an Image Stream Using Scene Graphs. Proceedings of the AAAI Conference on Artificial Intelligence, 34(05), 9185-9192

[7] Hsu, Chi-Yang Chu, Yun-Wei Ting-Hao, Huang, Ku, Lun-Wei. (2021). Plot and Rework: Modeling Storylines for Visual Storytelling.

[8] Braude, Tom Schwartz, Idan Schwing, Alexander Shamir, Ariel. (2021). Towards Coherent Visual Storytelling with Ordered Image Attention.

[9] Qi, Mengshi Qin, Jie Huang, Di Shen, Zhiqiang Yang, Yi Luo, Jiebo. (2021). Latent Memory-augmented Graph Transformer for Visual Storytelling. 4892-4901. 10.1145/3474085.3475236.

[10] Li, Tengpeng Wang, Hanli He, Bin Chen, Chang. (2022). Knowledge-enriched Attention Network with Group-wise Semantic for Visual Storytelling.

[11] X. Li, C. Liu and Y. Ji, "Associative Learning Network for Coherent Visual Storytelling," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10094740.

[12] Kullback and R. A. Leibler. 1951. On information and sufficiency. The Annals of Mathematical Statis- tics, 22(1):79–86.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recog- nition. In 2016 IEEE Conference on Computer Vi- sion and Pattern Recognition, CVPR 2016, Las Ve- gas, NV, USA, June 27-30, 2016, pages 770–778.

[14] Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In Proceed- ings of the ACL Workshop on Intrinsic and Extrin- sic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor. Association for Computational Linguistics.

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. Rethinking the inception archi- tecture for computer vision. In 2016 IEEE Con- ference on Computer Vision and Pattern Recogni- tion (CVPR), pages 2818–2826, Los Alamitos, CA, USA. IEEE Computer Society.

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems, pages 91–99.

[17] Chao-Chun Hsu, Zi-Yuan Chen, Chi-Yang Hsu, Chih-Chia Li, Tzu-Yuan Lin, Ting-Hao (Kenneth) Huang, and Lun-Wei Ku. 2020. Knowledge-enriched visual

storytelling. In Proceedings of Thirty-Fourth AAAI Conference on Artificial Intelligence.

[18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, and et al. 2017. Visual Genome: Connecting language and vision using crowdsourced dense image anno- tations. International Journal of Computer Vision, 123(1):32–73.

[19] Yang Yu, Kazi Saidul Hasan, Mo Yu, Wei Zhang, and Zhiguo Wang. 2018. Knowledge base relation de- tection via multi-view matching. New Trends in Databases and Information Systems, page 286–294.

[20] Zi-Yuan Chen, Chih-Hung Chang, Yi-Pei Chen, Jijnasa Nayak, and Lun-Wei Ku. 2019. UHop: An unrestricted-hop relation extraction framework for knowledge-based question answering.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Pro- cessing Systems, pages 5998–6008.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in CVPR, 2016, pp. 770–778.

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic eval- uation of machine translation. In Proceedings of 40th Annual Meeting of the Association for Com- putational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

[24] R. Vedantam, C. L. Zitnick, and D. Parikh. 2015. Cider: Consensus-based image description evaluation. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4566–4575.

[25] Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality us- ing longest common subsequence and skip-bigram statistics. In Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume, pages 605–612, Barcelona, Spain.

[26] Kobus Barnard, Pinar Duygulu Sahin, David A. Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. 2003. Matching Words and Pictures. JMLR (2003).

[27] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In ICML.

[28] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In CVPR.

[29] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi- aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021).

[30] Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. 2020. Normalized and geometry-aware self-attention network for image captioning. In CVPR.

[31] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-linear attention networks for image captioning. In CVPR.

[32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. NAACL (2019).

[33] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In ECCV.

[34] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In CVPR.

[35] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2022. Zero-Shot Image- to-Text Generation for Visual-Semantic Arithmetic. CVPR (2022).

[36] Xu, D.; Zhu, Y.; Choy, C.; and Fei-Fei, L. 2017. Scene graph generation by iterative message passing. In CVPR.

[37] Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2018. Neural baby talk. In CVPR, 7219–7228.

[38] Zellers, R.; Yatskar, M.; Thomson, S.; and Choi, Y. 2018. Neural motifs: Scene graph parsing with global context. In CVPR.

[39] Bai, S.; Kolter, J. Z.; and Koltun, V. 2018. An empirical evaluation of generic convolutional and recurrent networks.

[40] Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Em- pirical evaluation of gated recurrent neural networks on se- quence modeling. arXiv preprint arXiv:1412.3555.