Visual Story-Telling

Vini Vincent Kandankulathil
Computer Engineering DepartmentSt. Francis Institute of TechnologyBorivali, Mumbai
vinivincent2001@gmail.com

James Fernandes
Computer Engineering DepartmentSt. Francis Institute of Technology
Borivali, Mumbai jamesfernandes0701@gmail.com

Cyrus Fernandes
Computer Engineering DepartmentSt. Francis Institute of Technology
Borivali, Mumbai fernandescyrus4@gmail.com

Christy Chittilappilly
Computer Engineering DepartmentSt. Francis Institute of Technology
Borivali, Mumbai christythomas239@gmail.com

Mr. Shamsuddin Khan
Computer Engineering DepartmentSt. Francis Institute of TechnologyBorivali, Mumbai
shamsuddinkhan@sfit.ac.in

ABSTRACT

Because stories are varied and extremely individualised, there is a wide range of potential tale outputs. Because they are restricted to the vocab- ulary and knowledge in a single training dataset, existing end-to-end techniques result in repetitive stories.Vision-Language Pre-training (VLP) has advanced the performance for many vision-language tasks. However, most existing pre-trained models only excel in either understanding-based tasks or generation-based tasks. Furthermore, performance improvement has been largely achieved by scaling up the dataset with noisy image-text pairs collected from the web, which is a suboptimal source of supervi- sion. In this paper, we propose BLIP, a new VLP framework which transfers flexibly to both vision- language understanding and generation tasks. BLIP effectively utilizes the noisy web data by bootstrap- ping the captions, where a captioner generates syn- thetic captions and a filter removes the noisy ones. We achieve state-of-the-art results on a wide range of vision-language tasks, such as image-text retrieval , image captioning , and VQA.

Date of Submission: 08-07-2023 Date of acceptance: 25-07-2023

I. Introduction

Currently, there are limitations with existing vision- language pre-training methods that hinder their effec- tiveness for certain downstream tasks. These limitations are related to the model architecture and the data used for pre-training. Most methods use either an encoder-based or an encoder-decoder model, but these have their own drawbacks for different tasks. Additionally, most meth- ods pre-train on image-text pairs collected from the web, which can be noisy and suboptimal for vision-languagelearning.

To address these limitations, a new vision-language pre-training framework called BLIP (Bootstrapping Language-Image Pre-training) has been proposed. BLIP introduces a new model architecture called Multimodal Mixture of Encoder-Decoder (MED), which can operate as an encoder or decoder for text or images, allowing for effective multi-task pre-training and flexible transfer learning. BLIP also introduces a new data bootstrapping method called Captioning and Filtering (CapFilt), which finetunes a pre-trained MED into two modules: a cap- tioner to produce synthetic captions for web images, and a filter to remove noisy captions from both the originalweb texts and the synthetic texts.

Experiments and analysis show that the captioner and filter work together to achieve substantial performance improvements on various downstream tasks by boot- strapping the captions, and that more diverse captions yield larger gains. BLIP achieves state-of-the-art per- formance on a wide range of vision-language tasks, in- cluding image-text retrieval, image captioning, visual question answering, visual reasoning, and visual dialog. It also achieves state-of-the-art zero-shot performance when directly transferring models to two video-language tasks: text-to-video retrieval and videoQA.

II. Problem Formulation

Investigate how text and video versions of four differ- ent stories affect people's interest and affect, emotional engagement, memory of information, ability to sum- marise main points, and judgements. Visual storytelling can facilitate organisational learning, increase informa- tion generation and collaborative engagement and track the development trajectory in cognitively demanding re- search disciplines. Their findings demonstrate that the aural and visual elements of stories conveyed through video interact to increase interest, affect, and emotional engagement as well as the ability to summarise the main points, make judgements about the quality of the story, and form opinions about the subject matter. Im- ages are processed, interpreted, and synthesised by the human brain, which occupies a sizable section of the brain.According to research in cognitive science, individuals still learn, retain, and comprehend information better when it is presented in a multimedia manner (i.e., aural and three visual) as opposed to a purely prose format, independent of numerous context-based considerations. Information presented in modules or chunks and in a multimedia manner also encourages deep processing. We therefore argue for the use of visual story- telling as a complement to the conventional text-based academic article in the information systems (IS) disci- pline, building on the cognitive load theory (CLT), the cognitive theory of multimedia learning (CTML), and the research on deep processing. We contend that, compared to standard text-based articles, multisensory video storytelling would encourage deeper cognitive process-ing, which would enhance study understanding, dissem-ination, acceptability, citation, and influence.

III. Review of Literature

The paper [2] explores the advantages of using convolu- tional networks for machine translation and conditional image generation. The success of these networks has inspired the development of a convolutional image cap- tioning technique [8, 18, 19], which is demonstrated to be effective on the challenging MSCOCO dataset. The technique achieves performance on par with the tradi- tional LSTM baseline, but has a faster training time per number of parameters. The paper also presents a detailed analysis, highlighting the benefits of using convolutional language generation approaches. Building on this suc- cess, the paper presents a convolutional image caption- ing technique, which achieves comparable performance to the traditional LSTM baseline while having a faster training time per number of parameters. The paper also provides a detailed analysis of the benefits of using con- volutional language generation approaches, highlighting their ability to capture local structure and patterns in the text. Overall, the use of convolutional networks for ma- chine translation and conditional image generation has shown great promise, and this paper adds to the growing body of research demonstrating their efficacy in these areas.

In The next paper [15], a new approach for gener- ating image captions that uses CLIP encoding and a fine-tuned language model (GPT2) in combination with a simple mapping network is proposed. The use of CLIP encoding provides rich semantic features that have been trained with textual context, making them highly suitable for the task of vision-language perception. By com- bining the CLIP model with GPT2, the proposed approach is able to achieve a broad understanding of both visual and textual data, allowing it to generate meaningful captions for large-scale and diverse datasets with relatively quick training and no additional annotations or pre-training.[1, 3, 6, 13, 16, 17] One of the notable features of the proposed approach is its ability to gener- ate captions that are comparable to those generated by state-of-the-art methods, even when only the mapping network is trained, while the CLIP and language mod- els remain frozen. This leads to a lighter architecture with fewer trainable parameters, which can help to re- duce the computational cost of training the model. The approach is evaluated on challenging datasets like Con- ceptual Captions and nocaps, where it performs well while being simpler, faster, and lighter than competing methods. The results demonstrate the effectiveness of the proposed approach in generating meaningful image captions without the need for additional pre-training or annotations. Overall, the paper provides a novel and ef- fective approach to image captioning that leverages the strengths of CLIP encoding and GPT2 language model to achieve high-quality results on challenging datasets. The proposed approach has the potential to advance the state-of-the-art in image captioning, and could be valuable in a range of applications, including computer visionand natural language processing.

In this work [9], the paper presents a novel attention mechanism called ”Attention on Attention” (AoA) that extends the conventional attention mechanisms in neu- ral networks. The authors propose a new module that determines the relevance between attention results and queries, which helps the network to better focus on the most important features in the input data. The proposed AoA mechanism generates an ”information vector” andan

”attention gate” using the attention result and the current context. The information vector contains the information about the attended features, while the at- tention gate acts as a filter to control the information flow. The authors then apply element-wise multiplica- tion to the information vector and attention gate, which produces the ”attended information” - the expected use- ful knowledge. The authors apply the proposed AoA mechanism to both the encoder and decoder of their im- age captioning model, which they name AoANet. The experimental results show that AoANet outperforms all previously published methods on the challenging MS COCO dataset, achieving a new state-of-the-art perfor- mance of 129.8 CIDEr-D score on the ”Karpathy” offline test split and 129.6 CIDEr-D (C40) score on the official online testing server. Overall, the proposed Attention on Attention mechanism is a valuable contribution to the field of attention-based neural networks. The authors show that the mechanism can improve the performance of image captioning models, and achieve state-of-the-art results on challenging datasets. The results demonstrate the effectiveness of the proposed method in attending to important features and generating more accurate andmeaningful image captions. [4, 7, 10, 14, 21]

In the last paper [20], the authors introduce a new net- work architecture called the Transformer, which is based on attention mechanisms alone and does not use recur- rence or convolutions. The Transformer is more par- allelizable and requires significantly less time to train, while still achieving superior quality in machine trans- lation tasks. The authors demonstrate the effectiveness of the Transformer by achieving state-of-the-art results on the WMT 2014 English-to-German and English-to- French translation tasks. The model establishes a new single-model state-of-the-art BLEU score of 41.8 on the latter task after training for only 3.5 days on eight GPUs, which is significantly less training time than the best models from the literature. The authors also show that the Transformer generalizes well to other tasks, such as English constituency parsing, both with large and limited training data.

IV. Methodology

In this paper we used the BLIP model proposed by [12]The BLIP model is a neural network designed for image captioning that learns a joint representation of images and text using the Transformer architecture. The model leverages the Vision Transformer (ViT)[5, 20], a neural network that processes images as a sequence of patches and passes them through a series of Transformer lay- ers to learn spatial relationships between the patches.In addition to the ViT, the BLIP model includes a lan- guage model based on the Transformer architecture. The language model takes as input both the image features extracted by the ViT and a partial caption and generates the next word in the caption until the entire caption is generated. The BLIP model is pre-trained on a large dataset of image-caption pairs using various tasks such as image-text matching, masked language modeling, and caption generation. This pre-training allows the model to learn a robust joint representation of images and text, which is then fine-tuned on a specific image caption- ing task using a smaller dataset of image-caption pairs. Overall, the BLIMP model combines the strengths of computer vision and natural language processing to gen-erate accurate and descriptive captions for a wide range of images. Its pre-training and fine-tuning approach al-low for efficient learning and adaptation to specific tasks,making it a powerful tool for image captioning.

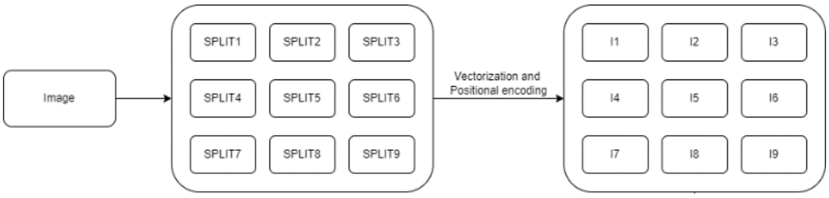


Figure 1: ViT Architecture

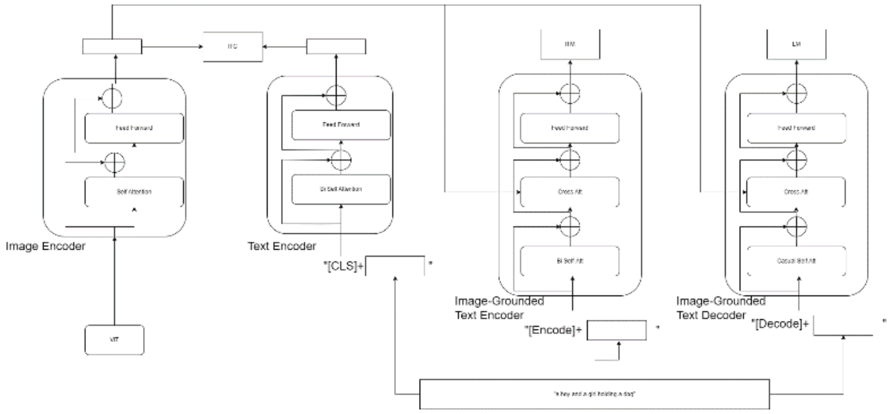


Figure 2: BLIP Architecture

Figure2. Pre-training model architecture and objec- tives of BLIP (same parameters have the same color). We propose multimodal mixture of encoder-decoder, a unified vision-language model which can operate in one of the three functionalities: (1) Unimodal encoder is trained with an image-text contrastive (ITC) loss to align the vision and language representations. (2) Image- grounded text encoder uses additional cross-attention layers to model vision-language interactions, and is trained with a image-text matching (ITM) loss to distin- guish between positive and negative image-text pairs. (3) Image-grounded text decoder replaces the bi-directional self-attention layers with causal self-attention layers, and shares the same cross-attention layers and feed forward networks as the encoder. The decoder is trained with a language modeling (LM) loss to generate captions given images.

V. Implementation Details

The multimodal mixture of encoder-decoder (MED) is a proposed multi-task model designed for pre-training a unified model with both understanding and generation capabilities. The MED model can operate in one of three functionalities.

1. Unimodal Encoder: The first functionality is the unimodal encoder, which separately encodes image and text inputs using a BERT text encoder with a [CLS] token appended to the beginning of the text input to summarize the sentence.
2. Image-Grounded Text Encoder: The second func- tionality is the image-grounded text encoder, which injects visual information by inserting an addi- tional cross-attention (CA) layer between the self- attention (SA) layer and the feed forward network (FFN) for each transformer block of the text en- coder. A task-specific [Encode] token is appended to the text input, and the output embedding of [En- code] is used as the multimodal representation of the image-text pair
3. Image-Grounded Text Decoder: The third func- tionality is the image-grounded text decoder, which replaces the bi-directional self-attention layers in the image-grounded text encoder with causal self- attention layers. A [Decode] token is used to sig- nal the beginning of a sequence, and an end-of- sequence token is used to signal its end.

Overall, the MED model provides a flexible approach for pre-training a multi-task model with various func- tionalities for encoding and decoding multimodal inputs.

During pre-training, the multimodal mixture of encoder-decoder (MED) model jointly optimizes three objectives, consisting of two understanding-based objec- tives and one generation-based objective. Each image-text pair only requires one forward pass through the computationally heavier visual transformer, and three forward passes through the text transformer, where dif- ferent functionalities are activated to compute the three losses.

1. Image-Text Contrastive Loss (ITC): The Image-Text Contrastive Loss (ITC) is activated using the uni- modal encoder. Its goal is to align the feature space of the visual transformer and the text transformer by encouraging positive image-text pairs to have simi- lar representations in contrast to the negative pairs. The ITC loss has been shown to be an effective objective for improving vision and language under- standing in previous studies (Radford et al., 2021; Li et al., 2021a). The model follows the ITC loss proposed by Li et al. (2021a), which introduces a momentum encoder to produce features and creates soft labels from the momentum encoder as training targets to account for the potential positives in thenegative pairs.
2. Image-Text Matching Loss (ITM): During pre-training, the BLIMP model optimizes three objec- tives: Image-Text Contrastive Loss (ITC), Image- Text Matching Loss (ITM), and Language Model- ing Loss (LM). Each objective requires a forward pass through different functionalities of the model. The ITC objective uses the unimodal encoder to align the feature space of the visual transformer and text transformer by encouraging positive image- text pairs to have similar representations. The ITM objective uses the image-grounded text encoder to learn a multimodal representation that captures the alignment between vision and language, where the model predicts whether an image-text pair is posi- tive or negative using an ITM head. The LM objec- tive uses the image-grounded text decoder to gen-erate textual descriptions given an image.
3. Language Modeling Loss (LM): To perform effi- cient pre-training while leveraging multi-task learn- ing, the text encoder and text decoder share all pa- rameters except for the self-attention (SA) layers because the SA layers capture the differences be- tween the encoding and decoding tasks. The en- coder employs bi-directional self-attention to build representations for the current input tokens, while the decoder employs causal self-attention to pre- dict next tokens. Sharing the embedding layers, cross-attention (CA) layers, and feed-forward net- work (FFN) can improve training efficiency whilebenefiting from multi-task learning.

VI. Results

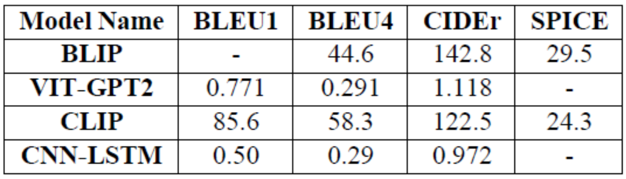
6.1 A. *Evaluation Metrics*

1. **BLEU** - BLEU is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. Quality is considered to be the correspondence between a ma- chine's output and that of a human: ”the closer a machine translation is to a professional human translation, the better it is”. This is the central idea behind BLEU. It was one of the first metrics to claim a high correlation with human judgements of qual-ity and remains one of the most popular automatedand inexpensive metrics.

2. **CIDEr** - The CIDEr (Consensus-based Image De- scription Evaluation) score is a metric used to evaluate the quality of image captions generated by automatic image captioning systems. It is based on the concept of consensus between the generated captions and human-generated captions. The CIDEr score measures the degree of agreement between the n-gram distributions of the candidate captions and the reference captions provided by human eval- uators. Specifically, it computes the cosine similar- ity between the two distributions, after weighting each n-gram by its term frequency-inverse docu- ment frequency (TF-IDF) score. The CIDEr score is often used in conjunction with other evaluation metrics such as BLEU (Bilingual Evaluation Under- study) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) to provide a more com- prehensive evaluation of the performance of imagecaptioning systems.

3. **SPICE** - SPICE (Semantic Propositional Image Caption Evaluation) is a metric used to evaluate the quality of image captions generated by computer models. It measures how well the model-generated captions match the human-generated captions based on the semantic propositions (i.e., the relationships between objects, attributes, and actions) in the cap- tions.The SPICE score ranges from 0 to 100, with higher scores indicating better performance. It is based on the F1 score, which balances precision and recall, and measures how many semantic proposi- tions in the model-generated caption match those in the human-generated caption. The score is then modified to account for how many unique seman- tic propositions are in the human-generated caption and how many are covered by the model-generated caption. SPICE is considered to be a more com- prehensive evaluation metric than some other cap- tioning metrics, such as BLEU, METEOR, and ROUGE, as it takes into account the semantic con-tent of the captions rather than just the surface-level similarity.

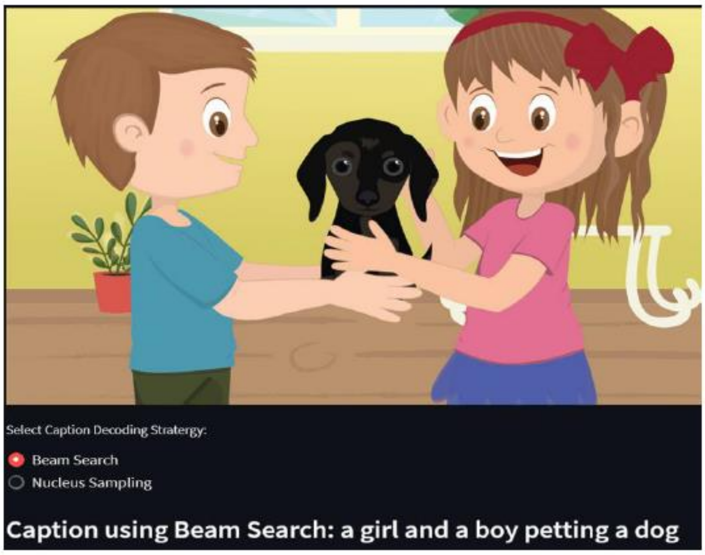
Table 1: Evaluation Score of the Image Captioning Mod- els.



Model Name	BLEU1	BLEU4	CIDEr	SPICE
BLIP	-	44.6	142.8	29.5
VIT-GPT2	0.771	0.291	1.118	-
CLIP	85.6	58.3	122.5	24.3
CNN-LSTM	0.50	0.29	0.972	-

6.2 B. *Results*

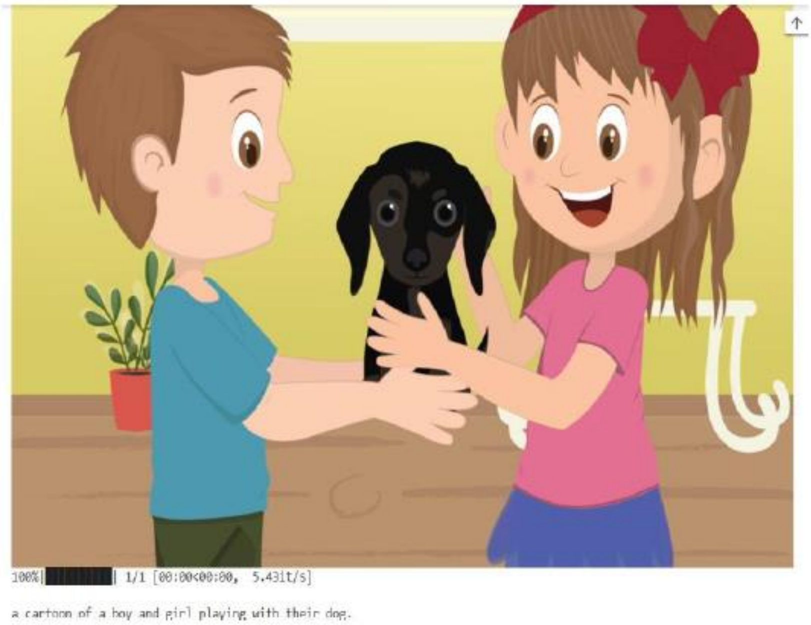
1) BLIP model:BLIP (Bootstrapping Language-Image Pre-training) is a model that is employed in research on computer vision (CV) and natural language processing (NLP). It is a pre-training method designed to boost lan- guage and vision models' performance on downstream tasks like picture captioning, answering visual questions, and visual grounding. To learn from noisy image-text pairs, we suggest BLIP, a unified VLP framework. This section outlines CapFilt for dataset bootstrapping after first introducing our new model architecture MED andits pre-training goals.



Caption using Beam Search: a girl and a boy petting a dog

Figure 3: BLIP model

2) CLIP model:New approach for image captioning that does not require additional object annotation and can be applied to any data. The CLIP model, which gen- erates semantic encodings for arbitrary images without additional supervision, as a prefix to textual captions, and fine-tune a pretrained language model to generate captions. The authors achieve close to state-of-the-art results on the Conceptual Captions dataset, and their method is faster than similar approaches.

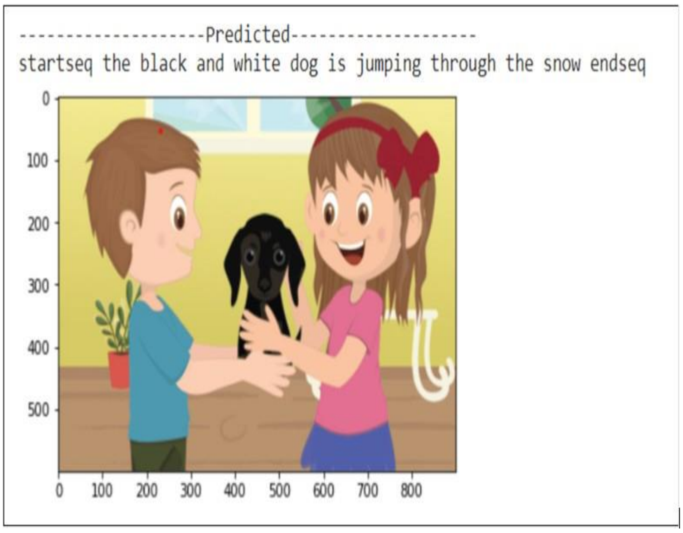


```
100%|██████████| 1/1 [00:00<00:00,  5.431t/s]

a cartoon of a boy and girl playing with their dog.
```

Figure 4: CLIP model

3) LSTM+CNN model:LSTM (Long Short-Term Memory) is a type of recurrent neural network (RNN) architecture that is commonly used for natural language processing (NLP) tasks. So to answer your question, LSTM is not a model in the sense that it is not a com- plete system or algorithm, but rather it is a specific type of neural network layer that can be incorporated into a larger model.



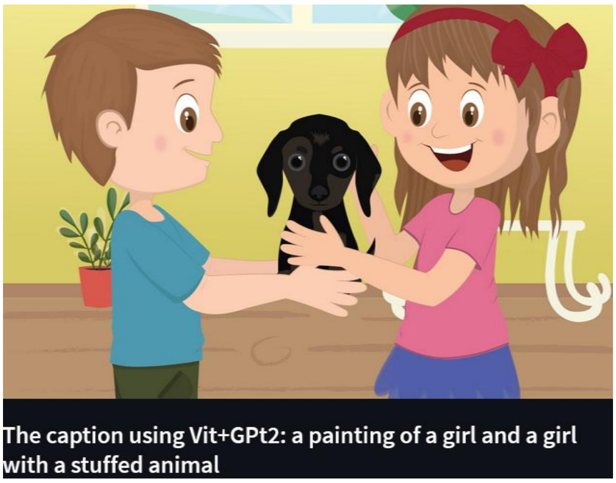
```
--------------------Predicted--------------------
startseq the black and white dog is jumping through the snow endseq
```

Figure 5: LSTM+CNN model

4) Vit-GPT2 model:The ViT/GPT-2 model for im- age captioning is a two-stage approach that combines two powerful pre-trained models[11]: the Vision Trans- former (ViT) and the Generative Pre-trained Transformer 2 (GPT-2). In the first stage of the model, an input image is fed into the ViT model, which extracts a set of feature vectors representing different parts of the image. The ViT model is trained on a large dataset of images and has been shown to be very effective at encoding visual information into a set of features. In the second stage, the feature vectors extracted by the ViT model are fed into the GPT-2 model, which generates a natural language caption for the image. The GPT-2 model is a language model that has been pre-trained on a large corpus of text data and is capable of generating high-quality natural language text.

VII. Conclusion

After extensive research, we can say that the outcomes of the suggested solution met our needs. The proposed architecture could be modelled and the image-sequence encoder was able to correctly learn the captain for the image. For the uploaded photographs, we were success- ful in telling the required story. The enhanced accuracy of the captions that were created was also a result of this encoder. Every person’s memories are incredibly personal, and with this technique, we hope to summarise these visual moments.And generate accurate story for every images.



The caption using Vit+GPt2: a painting of a girl and a girl with a stuffed animal

Figure 6: Vit-GPT2 model

Future Scope

In the future, we might potentially use our model to help by incorporating text into speech for long stories and allowing users to input images and receive a spoken explanation of the story’s visually developed that can be used as an aid for the blind. On the other hand, if there are any burglaries, it can be used to detect malicious behaviour.

References

[1]. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6077–6086, 2018.

[2]. Jyoti Aneja, Aditya Deshpande, and Alexander G.Schwing. Convolutional image captioning. In Pro-ceedings of the IEEE Conference on Computer Vi-sion and Pattern Recognition (CVPR), June 2018.

[3]. Xinlei Chen and C. Lawrence Zitnick. Learning a recurrent visual representation for image caption generation. CoRR, abs/1411.5654, 2014.

[4]. Yangyu Chen, Shuhui Wang, Weigang Zhang, andQingming Huang. Less is more: Picking in- formative frames for video captioning. CoRR, abs/1803.01457, 2018.

[5]. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, MatthiasMinderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognitionat scale. CoRR, abs/2010.11929, 2020.

[6]. Hao Fang, Saurabh Gupta, Forrest N. Iandola, Ru- pesh Kumar Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From captions to visual concepts and back. CoRR, abs/1411.4952, 2014.

[7]. Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. CoRR, abs/1611.08002, 2016.

[8]. Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convo- lutional sequence to sequence learning. CoRR, abs/1705.03122, 2017.

[9]. Lun Huang, Wenmin Wang, Jie Chen, and Xiao- Yong Wei. Attention on attention for image caption- ing. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019.

[10]. Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, JungWoo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. CoRR, abs/1610.04325, 2016.

[11]. Wing Man Casca Kwok and Kwok. Image caption- ing by vit/gpt-2, 03 2023.

[12]. Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language under- standing and generation. CoRR, abs/2201.12086,2022.

[13]. Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and JianfengGao. Oscar: Object-semantics aligned pre-training for vision-language tasks. CoRR, abs/2004.06165,2020.

[14]. Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via A visual sentinel for image captioning. CoRR, abs/1612.01887, 2016.

[15]. Ron Mokady, Amir Hertz, and Amit H. Bermano.Clipcap: CLIP prefix for image captioning. CoRR,abs/2111.09734, 2021.

[16]. Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, andRita Cucchiara. From show to tell: A survey on image captioning. CoRR, abs/2107.06912, 2021.

[17]. Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from trans- formers. CoRR, abs/1908.07490, 2019.

[18]. Aaron van den Oord, Nal Kalchbrenner, Lasse Es- peholt, koray kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with pixel- cnn decoders. In D. Lee, M. Sugiyama, U. Luxburg, Guyon, and R. Garnett, editors, Advances in Neu- ral Information Processing Systems, volume 29. Curran Associates, Inc., 2016.

[19]. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention isall you need. CoRR, abs/1706.03762, 2017.

[20]. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention isall you need. CoRR, abs/1706.03762, 2017.

[21]. Zhilin Yang, Ye Yuan, Yuexin Wu, Ruslan Salakhutdinov, and William W. Cohen. Encode, review, and decode: Reviewer module for captiongeneration. CoRR, abs/1605.07912, 2016.