

Title:  
*From Image Captioning to Visual Storytelling*

MSc Thesis research proposal

Author: *Admitos Rafail Passadakis*  
Email: [r.a.passadakis@students.uu.nl](mailto:r.a.passadakis@students.uu.nl)

March 2024

Daily Supervisor: *Yingjin Song*

1st examiner: *Albert Gatt*

2nd examiner: *Denis Paperno*

Department of *Artificial Intelligence*

Faculty of *Science*



**Universiteit  
Utrecht**

# Contents

<b>Abstract</b>	<b>3</b>
<b>1 Introduction</b>	<b>4</b>
1.1 Background	4
1.2 Introduction to Visual Storytelling	4
1.3 Motivation	4
1.4 Objective and Research Questions	5
1.5 Structure of the Work	6
<b>2 Literature Review</b>	<b>6</b>
2.1 V&L Tasks	6
2.1.1 Visual Question Answering	6
2.1.2 Image Captioning	7
2.1.3 Video Captioning	8
2.2 Generalized V&L Models	9
2.2.1 Former stages & the BERT family	9
2.2.2 Joint image-text training	9
2.2.3 The CLIP family	9
2.2.4 Other types of V&L models	10
2.3 Visual Storytelling	11
2.3.1 Preliminary Levels & Birth of Visual Storytelling	11
2.3.2 Post VIST era	11
2.3.3 Latest advancements	12
2.4 Text to Text Generation	13
2.4.1 RNN-based modules	13
2.4.2 Transformer-based modules	14
<b>3 Methods</b>	<b>14</b>
3.1 Dataset	14
3.2 Framework	15
3.2.1 The Vision-to-Caption architecture	16
3.2.2 The Caption-to-Story architecture	17
<b>4 Research Timeline</b>	<b>19</b>
<b>References</b>	<b>20</b>

# Abstract

Over the past decade, the intersection of Computer Vision and Natural Language Processing has given rise to transformative advancements in various tasks, including Image Captioning and Visual Storytelling. Contrastingly, to most of the current bibliography, which emphasizes that these two are separated tasks, this proposal embarks on a novel exploration, aiming to bridge the gap between isolated image captions and coherent narrative storytelling. In particular, we give a solid background of what exactly Visual Storytelling is, outlining our main objective and research questions. Subsequently, we provide an extensive review of the current Literature, covering all the surrounding aspects that this project may involve including the majority Vision & Language domain and some specific cases of Text-to-Text generation. Lastly, we give a concrete imprint of the framework that we will utilize during our exploration, underlining all the intermediate steps during the journey of transforming “sequence of images” to “narrative stories”.

# 1 Introduction

## 1.1 Background

In the unfolding landscape of contemporary research, the intersection of Computer Vision (CV) and Natural Language Processing (NLP) has ignited a profound interest in the art of generating textual narratives from images and videos. This interdisciplinary pursuit has given rise to a plethora of consequential tasks in both of these fields such as image labeling, image and video description (or captioning) and visual question answering. Before the development of tasks that flow out from both of these fields, CV and NLP have to show important achievements over the last years individually. In particular in the domain of CV, prominent results have been achieved in image description and classification with various deep neural network architectures such as the Convolutional Neural Networks (CNNs) [32, 57, 88, 92]. Simultaneously, on the NLP side, several tasks such as machine translation or text generation have become easier with new and pioneering models like Encoder-Decoder [15, 91] and most recently with Transformer architectures [11, 18, 79, 81, 98].

Yet, the urgency to forge more seamless connections between these twin pillars of Deep Learning has become more pronounced than ever before. To that end, the need of generating more narrative texts from images which will reflect temporal and sequential coherence, rather than just listing objects and their attributes or generating plain text without any incentive, has given rise to some novel challenges such as *Visual Storytelling* [42]. Herein, the narrative unfolds beyond a superficial representation, delving deep into the realms of experiences and feelings and providing a canvas for a profound exploration of a symbiosis between visual and textual narratives.

## 1.2 Introduction to Visual Storytelling

Visual storytelling [42], as an evolving field at the intersection of Computer Vision and Natural Language Processing, aspires to imbue machines with the ability to go beyond mere descriptive captions. Aiming to bridge the semantic gap between visual data and textual comprehension, this task transcends traditional image descriptions by delving into the nuanced and expressional realms of creating cohesive stories. It is usually considered as the descendant of the traditional image captioning, but unlike the latter, where the focus is often on simply detailing objects and their attributes, visual storytelling incorporates a narrative fashion by introducing coherence and a sense of temporal progression in the generated output.

For this precise reason we can trace the origins of visual storytelling back to the merging of image captioning and sequential image processing, where the goal now is not only to describe visual elements but to thread them together into a meaningful narrative. As technology evolves, so does the demand for multimedia storytelling, interactive interfaces, and immersive experiences. This entails the development of models capable of understanding the sequential relationships between images and crafting engaging and contextualized narratives that unfold over a series of visual inputs. Therefore, not only the interpretability of Artificial Intelligence (AI) systems is enriched but also doors to applications in areas such as content creation, multimedia understanding, and human-computer interaction are opened, fostering at the same time a deeper connection between AI and human expression [1].

A paradigm of a simple visual story that unfolds over 5 correlated images (a common feature is the dog) is shown in Fig. 1. Along with it, five isolated captions for these images are also provided. It is obvious that storytelling engages emotions and situations that the protagonists of the image are part of, instead of simple depictions of those.

## 1.3 Motivation

In this landscape, our research embarks on the challenging task of Visual Storytelling. We aim to navigate in the intersection of CV and NLP, leveraging the foundational achievements in both of these fields. Our focus extends beyond individual language or vision tasks, striving to build a framework that will deal with the complicated quest of Visual Storytelling. However, unlike most of the current work on this field, which deals with Visual Storytelling as an undivided task, we will attempt to split it into two separate levels.



*A small dog is running through the grass.    A small dog is walking through the leaves.    A dog is standing on the side of a road.    A woman and her dog on the edge of a bridge    A dog standing in the grass with it's tongue out*

*The black dog was ready to leave.    He had a great time on the hike.    And was very happy to be in the field.    His mom was so proud of him.    It was a beautiful day for him.*

**Figure 1.** An example of a sequential vision-to-language story, given also the isolated descriptions. The first line of captions, shows some isolated descriptions for these photos, whilst in the second line presents a story unfolded in five respective sentences.

In the first level, we want to keep the simplicity of image captioning that will give us the core of what’s going on in a series of images. Besides, we know already that image captioning is a task in which significant achievements have been made and over the last years major researches have showcased that we have the tools to generate appropriate descriptions for a variety of images. Furthermore, with the advancement of Deep Learning in NLP, many tasks including *text* or *story generation* have been simplified. To that end, our goal in the second level is to discover a suitable architecture that will transform these descriptions, blank of narrative meaning, to a sequential narrative story which will both reflect to visual context but also to temporal actions. Through this exploration, we aim to contribute to the evolving narrative AI systems that not only observe and understand but also recount an articulate story close to human storytelling and thus bridging the semantic gap between vision and language.

## 1.4 Objective and Research Questions

Despite the widespread belief that Visual Storytelling and Image Captioning are two distinct tasks, the objective of this Thesis project is to explore if we can work on these two simultaneously by considering image captioning as a total subset of visual storytelling. To do so, we intend firstly to use a well known transformer architecture, named Clip-Cap [71] to produce isolated captions for images sequentially correlated (i.e. the images will have a degree of correlation for example they can be from the same album). Following this, we target to deploy another transformer architecture, this time text-to-text to reformulate this bare captions to a more narrative and coherent stories which will express the dynamic interplay between visual elements. Ultimately, we are going to evaluate the generated stories compared to the original ones coming from the VIST dataset [42]. This evaluation contains comparison in terms of some well-known automatic metrics for image captioning like METEOR [9], BLEU [74], SPICE [5] and CIDEr [99]. However, lately a lot of criticism has been raised around how credible and efficient are these automatic metrics on evaluating machine-generated descriptions [39, 104, 105]. For this reason, we aim to make our evaluation procedure more robust by using human evaluation and judge on the generated results.

We can now summarize our described objective more formally in the following research question:

**RQ.** *Can we use a transformer based framework where we firstly generate isolated captions for images and then we reformulate them to extract a cohesive and narrative story?*

We can split the above research question in 2 new ones where we handle the caption generation and then the story generation, separately. Therefore, we have the following 2 derived sub-research questions:

**Sub-RQ 1.** *How accurate is Clip-Cap in image captioning on VIST dataset?*

**Sub-RQ 2.** *Can we use text-to-text, transformer-based model, to efficiently produce narrative stories from the captions generated by Clip-Cap?*

## 1.5 Structure of the Work

In this part, we present the structure of our exploration steps through several key sections. A focused Literature Review, is given in section 2, encompassing “V&L Tasks”, “Generalized V&L Models”, “Visual Storytelling” and “Text to Text Generation”, where we provide insights about the tasks and the models that have been applied to Visual Storytelling and other relatives domains such as Image Captioning. Additionally, we give a brief outline of the relevant work on the text generation task, that we want succeed. Following this, Section 3 details our Methods, with subsections “Dataset” and “Framework”, offering a glimpse into the technical foundations of our study such as the general architecture of the model that will be applied and giving the necessary background about the data that will be used. Finally, section 4 imprints the timeline of our future work, setting the general frame of how the upcoming project will unfold.

## 2 Literature Review

As previously mentioned over the last decade major advancements have been done in the both the fields of Computer Vision (CV) and Natural Language Processing (NLP). The confluence of these two has brought birth to many new tasks such as Visual Question Answering (VQA) [7], Vision-Language Navigation (VLN) [6, 12, 47]<sup>1</sup>, Image Captioning [13], Video Captioning [100] and of course Visual Storytelling [42]. At the same time, these tasks necessitated the development of models which will be able to operate in both Vision and Language (V&L) domains. V&L models are designed for dealing with these tasks and their architecture is inspired by the Encoder-Decoder [91] structure originally proposed for Machine Translation. Therefore, in their adaption to V&L tasks, these models consist of a visual encoder (instead of textual), a text encoder and sometimes, a cross-modal interaction module which maps the visual features to semantic embeddings. In the following sections, we will dive into the most significant V&L Tasks, explore some V&L Models that applied in these tasks and in subsection 2.3 we will give the latest progressions in the field of V&L exploring the state-of-the-art task of Visual Storytelling. Finally, in subsection 2.4, we will become acquainted with the technique of generating text (story) given some other text as input (caption).

### 2.1 V&L Tasks

Beginning our journey, we will first delve into the landscape of V&L tasks, exploring their diversity and significance in bridging the gap between the two principal pillars of Deep Learning namely, CV and NLP.

#### 2.1.1 Visual Question Answering

The task of Visual Question Answering (VQA) is to provide the answer given an image and a related question. In VQA, an algorithm is presented with an image, and users pose questions related to the content of that image expecting by the model to provide answers. The VQA task requires a robust understanding of both image and language representations, making it a challenging problem in the intersection of computer vision and natural language understanding. Various methods have been introduced [24, 25, 49] working on a plethora of datasets such as VQA dataset [7], Visual Genome [56] and others [70, 96]. Some of the techniques leverage deep learning architectures, including Convolutional Neural Networks (CNNs) for image processing and Recurrent Neural Networks (RNNs) for processing textual information. On top of this, an Attention Mechanism [8] can be used to enhance the models performance both in comprehension and generation [44, 111]. Lately, Transformer models, like LXMERT [93] or UNITER [14] have been deployed boosting even further the machine understanding over visual and textual forms, achieving new state-of-the-art (sota) results on various benchmarks of VQA. The ultimate goal is to develop architectures that can perform reasoning and inference across different modalities, demonstrating a broader comprehension of visual scenes and matching the corresponding human ability of understanding.

---

<sup>1</sup>A clarification need to be addressed here: Vision and Language tasks can be either Vision to Language or Language to Vision. The present work engages with the first category and since VLN is not clearly a Vision to Language task but it can also be a Language to Vision task, we are not going to analyze it here.



### 2.1.2 Image Captioning

Image Captioning, is closely related to VQA and it aims at generating a natural language description of an image. Open-domain captioning is a very challenging task, as it requires a fine-grained understanding of the global and the local entities in an image, as well as their attributes and relationships. The objective is to equip machines with the ability to understand visual content and express it in a human-like manner.

The origins of image captioning can be traced back to the pre-deep learning era when conventional methods based on manual engineering of features and on linguistic rules attempted to generate captions [23]. Some later approaches to image captioning relied on handcrafted features in conjunction with traditional machine learning algorithms [19, 54]. Others, produced image descriptions by using visual dependency representations that capture existing relationships between image objects [21]. However, the landscape transformed with the advent of deep learning, especially Convolutional Neural Networks (CNNs) [57] for image feature extraction and Recurrent Neural Networks (RNNs) [86] for sequential text generation. Once again, the rise of Transformers altered dramatically the scene on Image Captioning over the last few years.

The point zero for the evolution of Image Captioning but also for other V&L tasks, is considered to be the introduction of the Microsoft COCO (MSCOCO) dataset in 2014 [64], which provided a diverse and large-scale collection of images with associated captions, fostering standardized evaluation metrics and enabling fair comparisons between models. It was one of the first datasets that were not directly intended to image classification or recognition but was made for the broader context of scene understanding. Till today, numerous researches have worked on MSCOCO dataset and many advancements on all the abovementioned V&L tasks have been accomplished on this benchmark [6, 13, 84].

However, image captioning was actualized in practice in the paper of *Vinyals et al.*'s [102]. In this work, the authors apply for the first time deep recurrent architectures for Image Captioning. The model comprises of a Convolutional Neural Network (CNN) [57] to encode the image and a Long Short-term Memory (LSTM) [35] network to generate descriptive captions. The work was pivotal in popularizing the use of neural networks for image captioning, and it demonstrated state-of-the-art performance on the MSCOCO dataset. This marked a departure from the utilization of handcrafted features, to the usage of neural models for image captioning. Similar kind of architecture was deployed by *Karpathy et al.*'s in [51] (this time they used a bidirectional RNN as a decoder), where they successfully generated image descriptions on images from several datasets such as Flickr8K [36], Flickr30K [116] and MSCOCO. A combination of a CNN-LSTM approach was used once more in [20] for producing accurate visual descriptions and in [50] where the authors introduced the dense captioning task, a generalization of object detection and image captioning.

The incorporation of attention mechanisms into image captioning models enhanced their capability to focus on relevant regions of an image while generating captions. Models like *Xu et al.*'s [108] demonstrated improved performance by learning a latent alignment from scratch when generating corresponding words. Later, *You et al.* [115] utilized high-level concepts and injected them into a neural-based approach as semantic attention to enhance image captioning and *Lu et al.*, [68] introduced an adaptive attention mechanism with a visual sentinel to determine when to generate and where to attend during captioning. Lastly, *Anderson et al.* [6] proposes a model that combines bottom-up and top-down attention mechanisms that enables it to focus on prominent objects of the image and other salient regions.

Making one step further from the attention mechanism, there are several outstanding works on image captioning. One of them is [85], where *Rennie et al.* used reinforcement learning to optimize image captioning systems on MSCOCO. More specifically, a new system was built with an optimization approach that is called self-critical sequence training (SCST) which is a form of the REINFORCE algorithm [107]. Another one, is *Yao et al.* [114], who developed a model which consists of a semantic and a spatial scene graph with the purpose to detect objects in the image, based on their spatial and semantic connections. The important features of each node in the scene graph are refined by leveraging Graph Convolutional Networks (GCN) and then are feed into the attention-LSTM decoder for sentence generation. Similarly, *Yang et al.* [110] proposed a Scene Graph Auto-Encoder (SGAE) that incorporates the language inductive bias into the encoder-decoder image captioning framework. In a nutshell, the authors encode the graph structure of the sentence to learn a dictionary and then the semantic scene graph is encoded using the learnt dictionary. The last two approaches (*Yao et al.*, *Yang et al.*) are considered to be graph-based methods for image captioning.

Transitioning to the Transformers era, the first attempt to adapt this kind of models on image captioning was done by *Huang et al.* [41], when they introduced the Attention-on-Attention Network (AoANet) which extends the traditional attention mechanisms to determine the relevance between attention results and queries. Using *gated linear layers* [17], AoA generates an *information vector* and an *attention gate* using the simple attention result and the query and then applies element-wise multiplication to obtain the *attended information*. Another interesting result described by *Herdade et al.* [34] is that the simple position encoding (as proposed in the original transformer [98]), did not improve image captioning performance and thus a 2D encoding of position and size between detected objects is necessitated. Moreover, the entangled transformer [59] features a dual parallel transformer that encodes and refines visual and semantic information in the image, which is fused through a gated bilateral controller and eventually solves the semantic gap between vision and language that attention mechanisms and RNNs highly possess. *He et al.* [33] aimed to combine Transformers and graph-based methods by employing the spatial relations between detected regions inside an image. In their proposed model, each Transformer layer implements multiple sub-transformers to encode relations between regions. This encoding method combines a visual semantic and a spatial graph.

### 2.1.3 Video Captioning

Around the same period with the emerging of Image Captioning, the highly related task of Video Captioning was also nascent. However, unlike images that are static, working with videos requires modeling their dynamic temporal structure and then properly integrating that information for producing text in natural language. To that end, many works have come up with architectures that deal with both the spatial and temporal structure of videos and simultaneously produce descriptions.

The first time that Video Captioning and Deep Learning co-existed, was in the works of *Venugopalan et al.* [100, 101]. The principal of those works is the usage of stacked CNNs and stacked LSTMs for feature extraction from RGB frames of the video and description generation respectively. While the first study ([101]) uses *mean pooling* as the connection link between the two parts, the second ([100]) uses a more sequence-to-sequence approach exploiting also the optical flow of the images. Embodied with the an attention mechanism the model of *Yao et al.*'s [112] consists of a 3-D Convnet that incorporates spatio-temporal motion features of videos. They also extract dense trajectory features like HoG [16] and concatenate them with the input alongside with attention mechanisms which learn to weight the frame features non-uniformly conditioned on the previous input-words. Another innovative work is by *Wang et al.* [103], who propose a network with a novel encoder-decoder-reconstructor architecture (RecNet). In essence, they leverage both the forward optical flow as the backward one, for accomplishing both video captioning and reconstruction.

With the development of Transformers, video captioning met also radical changes. One of those is *Iashin et al.*'s [43], who presented a new dense video captioning approach that is able to utilize multiple number of modalities for event description from videos. They formulated the captioning task as a machine translation problem by utilizing the Transformer architecture and they showed that audio and speech modalities may improve a dense video captioning model when using automatic speech recognition (ASR) system. Last but not least, models of the BERT family were adjusted to video captioning task. Two important works are from *Sun et al.* and *Zhu et al.* [90, 122] who both altered the BERT architecture in order to learn joint embedding representations of video and language by applying pre-training and to focus on both global context and local details for video-text understanding.

### Similarities with Visual Storytelling:

At this point, we should underline that Video Captioning is probably the closest V&L task to Visual Storytelling which is the end objective of this project. On the commons side, they are both multi-modal tasks since they engage images (or frames) and text, they both engage natural language understanding and in both cases the output is given sequentially. On the other hand, Visual Storytelling focuses on a sequence of isolated images while Video Captioning deals with video data which consists of many continual frames. Additionally, Visual Storytelling involves narrative structure that connects semantically all of the input images forming a story, whilst Video Captioning aims on describing specific events or actions within the video without necessarily adhering to a broader narrative structure.



## 2.2 Generalized V&L Models

Having established some major Vision and Language tasks and most profoundly Image Captioning, the one of two main pillars of this work, now we turn our interest to some of the latest V&L models that have been developed the last few years, especially in the region around Image captioning and Visual Storytelling. However, it should be mentioned that with the establishment of Transformers on many tasks on NLP and CV, many of these models became multi-tasking in the sense that they can complete a plethora of different tasks, given the appropriate pre-training.

### 2.2.1 Former stages & the BERT family

To begin with, we will explore some such models for Image Captioning and VQA. As already seen, UNITER [14] and LXMERT [93] are two models like those. In addition, Zhou *et al.* [121] presents a unified Vision-Language Pre-training (VLP) transformer model, which can be fine-tuned for either vision-language generation (image captioning) or understanding (VQA), while at the same time uses both unified Encoder and Decoder, unlike the majority of previous models where these parts were separated. Some other important researches, concern the involve of the BERT family [18] in image captioning and generally in vision and language tasks. One of those, is named ViLBERT in [67], which is an extension of the classical BERT with the addition of a multi-modal two-stream model which can process both visual and textual inputs. ViLBERT is pre-trained through two proxy tasks and then transferred into multiple other vision-and-language tasks such as VQA and Image Captioning. Another alteration of BERT is VisualBERT [61], which feeds both text inputs and image regions into BERT aiming to discover the internal alignment between images and text with the self-attention mechanism. On a similar note, ImageBERT [77] is another transformer model, pre-trained on a large-scale dataset containing weakly supervised image-text pairs (this time the pretraining included four tasks), which exhibited it's effectiveness on various vision-language tasks, including image captioning.

### 2.2.2 Joint image-text training

Very close to VisualBERT which was trained on image-text pairs, is SimVLM [106], a simple V&L model which reduces the training complexity by exploiting large-scale weak supervision and is trained end-to-end with *prefix language modeling* [62]. The prefix tokens consist of the patched encoded images that are processed by bidirectional attention such that the model can consume visual information and then to generate the associated text in an autoregressive manner. SimVLM was trained on image-text pairs from ALIGN [48] and text-only data from C4 dataset [81]. CM3 [2] is another autoregressive model which combines causal and masked language modeling and is particularly known for producing hypertext and thus it has been used on designing HTML web pages of noted sites like CC-NEWS and Wikipedia.

### 2.2.3 The CLIP family

On January 2021 OPEN AI (Radford *et al.*) published a paper entitled “Learning Transferable Visual Models From Natural Language Supervision” [78] and introduced a new type of V&L architecture, named CLIP (Contrastive Language-Image Pretraining). CLIP is a multimodal vision-language model that is trained using contrastive learning, which means that it associates similar image-text pairs on it's input and differentiate between the dissimilar pairs. This enables the model to understand the relationships between images and their respective textual descriptions. It is designed to understand images and text in a unified manner, allowing it to perform a variety of tasks without task-specific training data. Given a dataset that contains image-text pairs CLIP is able to learn from it and then transfer this knowledge to various downstream tasks. In essence, CLIP links vision and language modalities into a unified embedding space, yielding tremendous potential for all V&L tasks including Visual Storytelling.

Since the introduction of CLIP, many works have been based on this and tried to utilize it to improve the contextual results on various V&L tasks. One of the first deployments of CLIP in these tasks was made by Shen *et al.* [87]. Epigrammatically, the authors explore the advantages that the usage of CLIP, as a visual encoder (with and without pre-training), can bring in many V&L tasks such as VLN, VQA and Image

Captioning. Likewise, *Portillo et al.* [76] deployed CLIP on Video Retrieval (Captioning) tasks obtaining state-of-the-art results on several respective benchmarks. An additional CLIP-based model, targeted for image captioning is *Luo et al.*'s VC-GPT [69], which avoids to place an object detector prior to the captioning model by connecting the visual encoder of CLIP to the used language model (GPT-2 [80]) using a self-ensemble cross-modal attention mechanism that handles both single- and cross-modal inputs. CoCa [117], is yet another model that uses contrastive learning and it is designed as pre-trained image-text encoder-decoder architecture which combines two kind of losses during training: contrastive loss and generation-caption loss. Interestingly, CoCa accomplished very high zero-shot transfer knowledge, on a variety of multi-modal evaluation tasks.

However, probably the most notable work that utilized CLIP specifically for image captioning is [71]. In this groundbreaking approach, *Mokady et al.* address the task of image captioning by using CLIP encoding as a prefix to captions [62]. This procedure involves employing a simple mapping network in order to connect aptly the CLIP encoder to the language decoder as well as fine-tuning the language model in case a simple mapping network is used. The authors show that without additional annotations or pre-training, CLIP-Cap (as they name the model) efficiently generates meaningful captions for large-scale datasets while simultaneously remains extremely light-weight for training.

Furthermore, CLIP-Cap became influential for some subsequent works such as *Tewel et al.*'s [94] and *Ramos et al.*'s [82]. In the former, the authors propose an entirely unsupervised approach (no training or tuning, named ZeroCap), by essentially using CLIP-Cap's architecture to perform zero-shot image captioning. They observe the capability of a CLIP-based model to create reasonable captions, beyond the prefix prompted zero-shot learning that [71, 78] propose. In the latter, the CLIP-Cap's structure is exploited once again with the goal of reducing its trainable parameters even more. Their key contribution is that this model (SMALLCAP as they name it) uses only interleaved cross-attention layers between the CLIP encoder and the language decoder, thus reducing the training time, and that it generates captions conditioned on the input image and some related captions retrieved from a datastore.

#### 2.2.4 Other types of V&L models

Outside of the limits of CLIP, there are other models that have been well established in the general area of V&L tasks in the last few years. Starting with, is OSCAR (Object-Semantics Aligned Pre-training) [63], a model which uses object tags detected in images as points of notification that can be used as an alignment between visual and textual features. During pre-training OSCAR collocates *Contrastive Learning* along with *Mask Language Modeling* [18] since it incorporates both the contrastive loss and the mask language loss given by the input image-text pairs which in this case are represented as a triple (word tokens, object tags, region features). *Zhang et al.*'s work [119] is considered to be another major object-centric model in which OSCAR is integrated as well. The main focus of the paper is to improve visual representations for V&L tasks by utilizing a much larger object-detector which contributes to richer semantics and visual concepts which eventually will generate more accurate responses.

Very similar to CLIP-Cap, *Tsimpoukelli et al.*'s model (Frozen) [97], extends the notion of static prefix and prompt tuning [58, 62] by making a dynamic prefix, in the sense that it is not a constant bias added to the text embeddings, but an input-conditional extension produced by a neural network. In this whole process, the language model remains totally frozen. The resulting structure is a multimodal few-shot learner, with the ability of learning a variety of relevant tasks. Working on a different way, *Li et al.* [60], involved bootstrapping in an attempt to create a system that jointly deals with understanding-based tasks and generation-based tasks. More precisely, a captioner is used for refining the training data by generating synthetic captions with a filtering step to mitigate the noise present in the given web-collected captions. Ultimately, Flamingo by *Alayrac et al.* [4] is another model for few-shot learning which interleaves cross-attention fusing layers into the text decoder and interposes a specific transformer-based schema called *Perceiver* [45], between the latter and the visual encoder. It is worth mentioning, that the training procedure in Flamingo uses *NLL loss* in an autoregressive manner. Due to these facts, Flamingo outperformed many task-specific models on numerous benchmarks, such as COCO [64] or VQAv2 [30] datasets.

## 2.3 Visual Storytelling

Thus far, we have explored various Vision and Language tasks and models, yet we have not delved into the primary objective of this study, which is Visual Storytelling. In the following paragraphs we will break down some techniques that have been applied for Visual Storytelling over the last couple of years.

### 2.3.1 Preliminary Levels & Birth of Visual Storytelling

As already mentioned Video Captioning can be placed quite close to Visual Storytelling as a task. To that end, some works have approached Visual storytelling via very similar tasks such as video summarization [27] by using supervised probabilistic models. Others, like *Park and Kim* [75], came even closer to the traditional definition of Visual Storytelling, by applying a multimodal architecture called *Coherent Recurrent Convolutional Network (CRCN)* in order to get a sequence of natural sentences for a stream of images.

Nevertheless, Visual Storytelling was properly introduced in 2016 with the construction of a pioneering dataset named the VIsual StoryTelling (VIST) [42]. This work created the first dataset for sequential vision-to-language exploration. Initially the dataset included 81,743 unique photos in 20,211 sequences, aligned to both mere language descriptions (captions) and story language. In particular, the authors dub the captions as “Description of Images in Isolation - DII” and the stories as “Stories of Images in Sequence - SIS”. Moreover, for producing both DII and SIS, the authors applied crowdsourcing, resulting all the annotations to be human-written. The main purpose of Visual Storytelling is to generate a sequence of sentences that collectively form a coherent story, a task that requires not only understanding individual images but also composing a narrative structure. With that said, the authors clearly differentiate Visual Storytelling from Image Captioning. Last but not least, the paper proposes new evaluation metrics for assessing the quality of generated stories and based on those it gives some baselines on this task by deploying neural networks.

With the introduction of VIST dataset, new opportunities arose for researchers to develop models and evaluate their performance specifically on Visual Storytelling tasks. This landmark dataset paved the way for innovative exploration in the common region of CV and NLP, sparking new avenues of research.

### 2.3.2 Post VIST era

#### Neural Networks Deployment:

Among the first works that experimented on VIST dataset was by *Agrawal et al.* [3], who retrieved jumbled images and their respective captions from the foresaid dataset and proposed a method of sorting them into a coherent story (something quite similar to our goal), by using two different type of networks: 1) A language-alone unary model that uses a Gated Recurrent Unit (GRU) [15] and 2) A language-vision binary model that embeds both the caption and the image using also CNNs. What is more, *Liu et al.* [66] also explored the generation of structured paragraphs from photo streams of the VIST dataset, but unlike [75], their model encompassed an additional attention mechanism to combat the large visual variance between the photo collection and to preserve the long-term language coherence among a multiple sentence text.

*Gonzalez-Rico and Fuentes-Pineda* in [28] proposed a neural visual storyteller that creates short stories from image sequences, extending the capability of the model by *Vinyals et al.* [102], which generated mere image descriptions. The extension relies on a series of encoder LSTMs (instead of only one in [102]), to compute a context vector of each story from the input sequence of images. Then, this context vector serves as the initial state for several separate decoder LSTMs, each of which generates the story segment for the corresponding image in the input sequence.

*Kim et al.*'s work [53] stands as a major turning point in the history of Visual Storytelling. This paper suggests a deep learning network model, designated as GLAC-Net (Global-Local Attention Cascading Network). This network is very alike to many previous models (CNN feature extractor, LSTM generators) but it adds two variations. Firstly, as its name indicates, it uses two types of attention mechanisms, one local (image feature level) and one global (the overall storyline encoding level). Secondly, it exploits a cascading mechanism which means that during generation the hidden state (context) of the previous sentence is conveyed to the next sentence. As a result, GLAC-Net contributed to very competitive results on the VIST dataset and was instituted as a robust storytelling model.

### Deep Hierarchical Approaches:

After the release of VIST dataset, two contemporary works proposed deep hierarchical approaches to Visual Storytelling which were by *Krause et al* [55] and by *Yu et al* [118]. These studies, extend the idea of image captioning and deal with the complicated task of visual storytelling by developing deep recurrent architectures with special modalities. More specifically, in [55] the model composes of an object-detector that projects image features (via pooling) to the space of two RNNs accountable for the coherence of the generated text, in word and sentence level respectively. Concurrently, the model of [118], consists of an image-features extractor (ResNet101 [32]) and three different RNNs, which encode the given album (as a whole), isolate the most relevant photos to the matched descriptions and generate the final story correspondingly.

A further hierarchical approach to our task, was given by *Nahian et al.* [72] who controversially used the sole image descriptions themselves (along with the images of course) of the VIST dataset, so to generate stories that maintain their context and coherence. Model-wise their architecture comprises of different levels having multiple encoders and decoders for both isolated and sequential images and/or descriptions.

### Application of Reinforcement Learning:

*Wang et al's* work [105], was regarded as another significant milestone for the Visual Storytelling field. To begin with, the authors highlight the limitations of conventional metrics such as BLEU and METEOR in assessing the quality of generated stories and afterwords, to address this problem they propose an adversarial reward learning (AREL) framework, which includes a policy and a reward model. In a nutshell, the former takes an image sequence as input and attempts to form a narrative story while the latter, is optimized adversarially and aims to bring the prediction closer to the respective human annotations. Using both automatic metrics and human evaluation this system established state-of-the-art results in VIST dataset and approached more human-like stories. Like in the case of GLAC-Net, AREL is also considered a mainstay for many models developed for Visual Storytelling tasks, even till today.

Similarly to [105], *Hu et al* [40] proposed another storytelling framework that applies reinforcement learning by exploiting some specific reward functions. Working also on the VIST dataset, the paper conducts a study on what are the characteristics of a coherent story, emphasizing that there are three such key-criteria: *relevance, coherence and expressiveness*. Subsequently, the authors use the aforementioned model trying to capture the essence of these three characteristics when generating stories.

### 2.3.3 Latest advancements

#### Transformer-based Approaches with external Knowledge Graphs:

With the new decade, noteworthy progress continued to emerge in the field of Visual Storytelling, some of which were marked by the integration of Transformer-based models. Perhaps one of the first works that implemented a versatile architecture consisting of multiple levels of transformer-based sub-parts, was by *Hsu et al.* [37]. In this work, the authors presented how the utilization of external *Knowledge Graphs*, can aid for story generation. To that end, a three-stage framework, named *KG-Story*, is developed. Collectively, this model applies three type of actions dubbed as *distill-enrich-generate*. Initially, the prompted images are distilled to word-terms in order to get descriptive representations. Following this, these word-terms are enriched to a linked conceptual path using the external graphs and finally, a simple transformer model [98], enhanced with positional encoding converts these paths to stories. In like manner, authors in [89] combine BERT [18] and hierarchical LSTMs to accomplish the generation of semantically complete stories. Conforming another hierarchical approach (from [55]), *Su et al.* attempted to approach Visual Storytelling, with a framework that models word-level and sentence-level semantics separately. Except of using a CNN image-feature extractor (VGG16 [88]) the paper deploys BERT to obtain textual embeddings from sentences/words and feeds them (along with image-features) to the hierarchical LSTM decoders.

The fragmentation, of Visual Storytelling as task, continued at *Hsu et al's* work [38], which introduces PR-VIST, a multi-stage model that generates human-relevant stories. Much like their previous work in [37], this model also takes advantage of external representation graphs. More specifically, PR-VIST depicts the input image as a visual graph in which the elements of the image stand as nodes and ultimately finds the optimal path that forms the best storyline. Subsequently, in the second stage of the model, this storyline is



reworked in a similar fashion that an image is generated by Generative Adversarial Networks (GANs) [29]. In particular, a story generator takes the storyline and produces the actual story, back-propagating the loss which is given by the evaluator during the classification of the generated story as good or bad.

### **VIST Dataset Expansion & Criticism:**

Taking into consideration the importance of GLAC-Net and AREL frameworks, *Hsu et al.* [39] utilized these two models in their process to expand the VIST dataset by adding human edits on machine-generated visual stories. The new dataset, *VIST-Edit*, includes 14,905 human-edited versions of 2,981 stories generated by AREL and GLAC-Net and underscores the weak correlation between automatic evaluation metrics and human ratings. Operating also in the dataset-level *Ravi et al.* [83], proposed *AESOP* (Abstract Encoding of Stories, Objects, and Pictures). This new dataset contains different themes and its stories are highly narrative, coherent and follow a clear causal arc having even a moral at the end. Thus, *AESOP* fosters a more creative and causal reasoning taking Visual storytelling to its next level. Lastly, the author set some baselines for this novel dataset, creating ample opportunities for further exploration and feature research.

An additional study that criticized the traditional natural language generation metrics which are based on  $n$ -gram matching, such as BLEU or CIDEr, is by *Wang et al.* [104]. Claiming that these scores have weak correlation with corresponding human evaluation/judgment, the paper proposes and mathematically formulates three new scores: 1) visual grounding, 2) coherence, and 3) non-redundancy, which are eventually measured in reliability on the VIST dataset. Ultimately, *Liu and Keller* [65] inspired by [83], extended the notion of Visual Storytelling to a more character-centric point of view. Underlining, that a coherent story does not necessarily imply a narrative with protagonists and action, the paper introduced the *VIST-Character* dataset which provides rich character-centric annotations, including visual and textual co-reference chains and importance ratings for the characters.

## **2.4 Text to Text Generation**

The purpose of this work is to examine, unlike most of the current literature review [42], the step of transitioning from Image Captioning to Visual Storytelling. As explained in 1.3 and 1.4, in order to do so, we will need to deploy a novel architecture which will be capable of reformulating captions to coherent and narrative storylines. For this purpose, it would be valuable to acquaint ourselves on how previous studies have dealt with the problem of text-to-text reformulation (generation). The idea of storytelling is not new and even with the dawn of the new millennium, works developed multi-agent frameworks in virtual environments [26, 95] with the ability of generating natural language and undoubtedly till today, there are plenty of papers dedicated to story generation most of which concern pre-trained (Large) Language Models (LLMs) [1, 11, 18, 52, 81] or their variations which they are able of completing a large number of tasks including text-reformulation and text-generation. Nonetheless, the works that use specifically prompted sentences or keywords to construct a story are more limited and precisely, to the best of our knowledge, the methodology of “evolving” from Image Captioning to Visual Storytelling has not been applied yet.

### **2.4.1 RNN-based modules**

Quite close to our purpose, *Jain et al.* [46] propose a sequence-to-sequence RNN [91] to address the task of coherent story generation from independent descriptions. In addition to that, they also deal with the task of text-generation as a Statistical Machine Translation (SMT) problem and they use two popular methods: Phrase-based-SMT and Syntax-based-SMT, which they compare it head-to-head with the deep learning-based approach. *Yao et al.* [113] propounded a *plan-and-write* hierarchical generation framework that given a title as input, first plans a storyline and then generates a story based on the storyline. The hierarchical approach is composed by two levels: the dynamic and the static parts. The former (composed by a *Gated Recurrent Unit* - *GRU* [15]) is accountable for generating the next word in the storyline as well as the next sentence in the story and at each timestep uses the previously generated text as context (initialized with the title), whilst the latter (composed by an LSTM) is not delighted with such flexibility and generates a whole storyline straight away, giving though a general realization of the story, enhancing its coherence.

## 2.4.2 Transformer-based modules

Moving at the same pattern, *Guan et al.* [31] came up with a transformer-based architecture that firstly, utilizes commonsense insights from external knowledge bases and secondly, uses multi-task learning with the objective of distinguishing true and fake stories during fine-tuning, with the final aim of generating reasonable stories given the central context of the line. The model outperformed other state-of-the-art language models such as GPT-2 [80] on different evaluation metrics like Perplexity (PPL) and BLEU. Inspired by these previous works and based on [52], *Xu et al.* [109] proposed *MEGATRON-CNTRL*, a novel large-scale language model, that adds control to text generation by incorporating once more, external knowledge sources. This architecture is comprised of four main elements: a keyword predictor, a knowledge retriever, a contextual knowledge ranker and a conditional text generator which are respectively responsible for: setting the general theme of the story, receiving the external information, ranking this information and promoting only the useful parts and generating the final story. It’s worth emphasizing, that *MEGATRON-CNTRL* showcases remarkable controllability and simultaneously immense fluency, consistency and coherence in it’s stories under both automatic metrics and human evaluation.

Likewise, *Fang et al.* [22] worked also in controllable story generation by devising another transformer-based framework which heavily depends on text-prompts. Specifically, the authors integrate latent representation learning on a pre-trained transformer model so to construct a Conditional - Variational Autoencoder (VAE-CVAE) [10, 120]. In order to incorporate correctly the latent representation vectors in CVAE, they use three alternative injection mechanisms based on: 1) A simple addition of the input token to the positional embeddings, 2) Projection of encoded representation to a larger space through “pseudo-attention layers” and 3) *Softmax* operation<sup>2</sup>. Eventually, this study achieves both controllability and state-of-the-art efficiency, over story generation under automatic metrics such as PPL and ROGUE (along with it’s variations).

## 3 Methods

Having established the necessary background in Vision & Language tasks and models, especially in Image Captioning and Visual Storytelling, and having also a slight acquaintanceship with text-to-text story-generation models, now we will give a brief overview of the data that will be used and the methods that will be deployed/implemented throughout this project.

### 3.1 Dataset






Similarly to the majority of Visual Storytelling studies that were previously mentioned, this work will also make use of the VIST dataset [42]. As already mentioned on 2.3.1, VIST dataset is the first collection of sequential images with corresponding isolated descriptions and human-made stories and it introduces the task of *Visual Storytelling*. The dataset initially included 81,743 unique photos in 20,211 sequences and is divided in albums, for each of which, crowd-workers were selected to form five-image sequel by completing a linguistic story, consisting of five sentences as well. In total, for each of the stories, the dataset contains aligned three tiers of information. 1) *Descriptions of images-in-isolation (DII)*; 2) *Descriptions of images-in-sequence (DIS)*;<sup>3</sup> and 3) *Stories for images-in-sequence (SIS)*. An example, of a five-images story along with all of *DII*, *DIS* and *SIS* is shown in Fig. 2.

To obtain more insights about the complicated structure of VIST data, we conducted a preliminary research, from which we retrieved the exact statistics of the dataset. Regarding albums the dataset consists of 8,031 albums, in the training set, 998 albums in the validation and 1,011 albums in the test set. In summary, from those albums, 50,200 stories are constructed. In particular, 40,155 stories exist in the training set, 4,990 stories are in the validation set and 5,055 stories are in the test set. At the same time, regarding the images, the dataset contains 154,430 unique images in the training set, 21,048 images in the validation set and lastly 30,000 images in the test set.

<sup>2</sup>However, it is found that this method practically can’t work

<sup>3</sup>Note that in the release of the dataset, DIS annotations were not published.



DII					
	A black frisbee is sitting on top of a roof.	A man playing soccer outside of a white house with a red door.	The boy is throwing a soccer ball by the red door.	A soccer ball is over a roof by a frisbee in a rain gutter.	Two balls and a frisbee are on top of a roof.
	A roof top with a black frisbee laying on the top of the edge of it.	A man is standing in the grass in front of the house kicking a soccer ball.	A man is in the front of the house throwing a soccer ball up	A blue and white soccer ball and black Frisbee are on the edge of the roof top.	Two soccer balls and a Frisbee are sitting on top of the roof top.
SIS	A discus got stuck up on the roof.	Why not try getting it down with a soccer ball?	Up the soccer ball goes.	It didn't work so we tried a volley ball.	Now the discus, soccer ball, and volleyball are all stuck on the roof.

**Figure 2.** Example of a five-images story along with the descriptions of images in isolation (DII); descriptions of images in sequence (DIS); and stories of images in sequence (SIS) [42]

However, it should be underlined that due to crowdsourcing the number of annotations in DII and SIS is not the same<sup>4</sup>. Recalling also the fact that our work focalizes on how we could use the image descriptions (captions) to produce an interesting story, it becomes plausible that we should turn our attention only on SIS annotations that are present on DII annotations and as well as they form a five-length story. Therefore, this procedure reduces the total number of training stories to 26,959, validation stories to 3,354 and test stories to 3,385. In Table 1, we depict the complete statistics of the VIST dataset with regard to the initial and subsequent annotations and stories before and after the disposal process. It also, shows the total number of albums, from where the Stories-in-Sequence were derived as well as the total number of images in each set.

VIST Dataset Statistics			
Data Type\Set	Train	Validation	Test
SIS Albums	8,031	998	1,011
Images	167,528	21,048	21,075
DII Annotations	120,465	14,970	15,165
SIS Annotations	200,775	24,950	25,275
SIS Stories (of length 5) <sup>5</sup>	40,155	4,990	5,055
SIS Annotations with photo_id in DII	159,899	19,977	20,080
DII-SIS Annotations repeated 5 times	134,795	16,770	16,925
DII-SIS Stories (of length 5)	26,959	3,354	3,385

**Table 1.** Analytical statistics of the VIST dataset regarding the number of annotations, stories, albums and images

## 3.2 Framework

In this stage we are going to explore and familiarize ourselves more with the framework that is we are going to use in the journey of generating isolated captions (initially) and then transforming them to a cohesive narrative. As previously underlined in sections 1.3 and 1.4, we will deal with these two steps

<sup>4</sup>There are more annotations in SIS, because crowd-workers could repeatedly pick the same images from the given albums.

<sup>5</sup>Here we length of 5 we mean that the generated stories contain five images with five story descriptions. This is obvious for the initial SIS stories but after the truncation due to the demand of correspondence with the DII, all stories are not of length 5.

separately and thus, we will utilize two different architectures to achieve that. Firstly, comes the Vision-to-Caption architecture which practically consists of the known to us Clip-Cap model [71] and then is the Caption-to-Story architecture which will comprised of a transformer-based model, intended to be built from scratch following some general guidelines as we saw them earlier in the Literature review.

### 3.2.1 The Vision-to-Caption architecture

In this section, we will gonna describe the main architecture that we are going to use which is Clip-Cap [71] by *Mokady et al.* In a nutshell, Clip-Cap is comprised by three sub-parts; the CLIP model, a language generator model and a mapping network that projects the CLIP embeddings in a way that are feedable to the language model. In the original paper, the authors created two variants of the Clip-Cap model depending on whether the language generator was fine-tuned or not. It's worth mentioning that in both of these versions, the CLIP model was frozen (not fine-tuned). In the following paragraphs, we will elaborate more on these three sub-parts of Clip-Cap.

#### Image Encoder:

The image encoder is comprised by CLIP (Contrastive Language-Image Pretraining), which (as we saw) is a multimodal vision-language model that is trained using a contrastive learning approach. It is designed to understand images and text and judge if these can fit together in the form of captioning. During training, it tries to maximize the “cosine similarity” between correct image-caption vector pairs, and minimize the similarity scores between all incorrect pairs. During testing, it calculates the similarity scores between the vector of a single image with a bunch of possible caption vectors, and picks the caption with the highest cosine similarity. It should be underlined that CLIP is not a caption generation model; rather, it's primary capability lies in determining the compatibility between existing textual captions and corresponding images. The basic architecture of CLIP is shown in Fig. 3 and as you can observe it comprises of both image and text encoders for projecting these two modalities in the same embedding space (a 2–dimensional matrix).

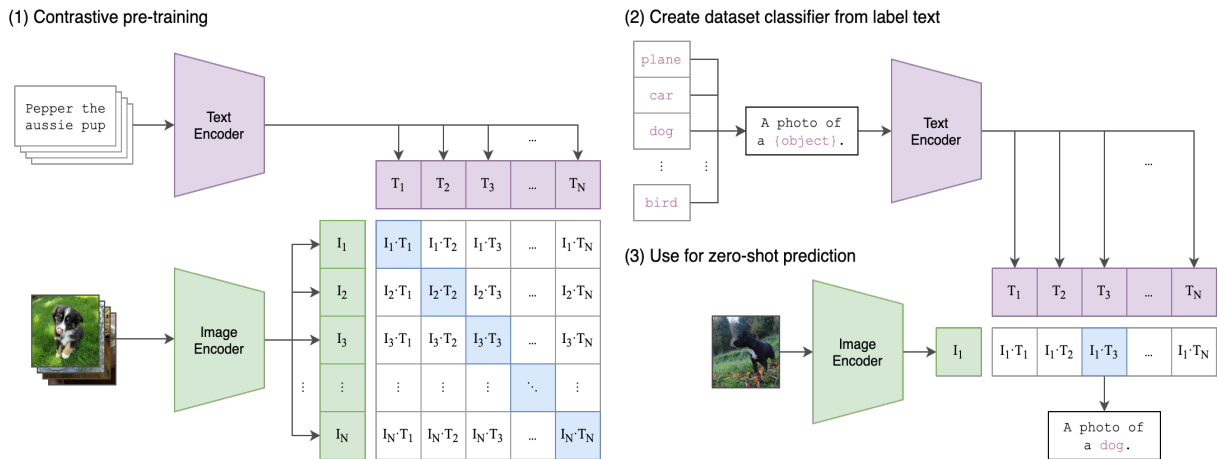


Figure 3. Architecture of the CLIP model [78]

#### Language Decoder (Generator):

The text generator model can be any language model that is capable of generating text. In the original work the authors used the Generative Pretrained Transformer-2 or GPT-2 [80]. GPT-2 is a large language model by OPEN AI and is the second in their GPT series of models. This model is pre-trained on BookCorpus, a dataset of over 7,000 self-published fiction books from a variety of genres. After this, GPT-2 was trained also on a dataset of 8 million web pages. In its final and largest form the model contains about 1.5 billion parameters. In short, the model processes input sequences in parallel through layers of self-attention mechanisms, capturing contextual relationships between tokens (words). In addition, it uses a mask-mechanism to make sure that the prediction for the token  $i$ , only uses the inputs from 1 to  $i$  but not the future tokens. During training, GPT-2 learns to predict the next token in a sequence based on the preceding context. In generation tasks, the model utilizes its learned knowledge to generate coherent language autoregressively.

### Mapping Network:

The main functionality of the mapping network, is to translate the embeddings generated by CLIP and a learned constant to the GPT-2 input space. As previously noted according to [71], the 2 versions of Clip-Cap depend on fine-tuning or not of the language model (GPT-2). As a results the authors used a different mapping network when GPT-2 was frozen and when it was not. In the first case, they deployed a simple Multi-Layer Perceptron (MLP), whilst in the second they utilized a more expressive transformer [98] architecture. The transformer enables global and self attention between input tokens while reducing the number of parameters for long sequences. In their work, the transformer is fed with two inputs, the visual encoding of CLIP and a learned constant input. This will enable the model to retrieve meaningful information from CLIP embeddings as well as to learn to adjust the fixed language model to the new data. An illustration of the mapping network, as a black-box, embedded on the whole of Clip-Cap’s architecture is shown in Fig. 4, where we can see both CLIP (encoder) and GPT-2 (decoder/generator).

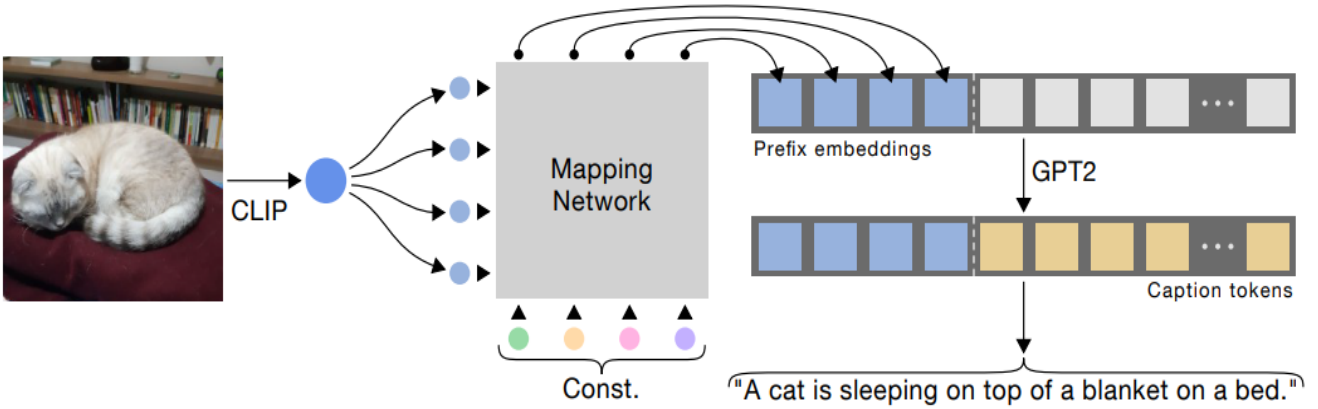


Figure 4. Architecture of the Clip-Cap model [71]

### 3.2.2 The Caption-to-Story architecture

Coming to the second part of our framework we will need a text generation model, that will be able to construct narrative stories from simple, isolated captions [46]<sup>6</sup>. One simple way, would be to reuse a generative model, which will originate from the transformer-class, such as those introduced earlier in sub-section 2.4 (e.g. *MEGATRON-CNTRL* [109]). However, in order to explore in-depth the realm of transformer models and to gain more hands-on experience on how to build a model that can to ponder and create narrative storytellings, our first option will be to build from scratch this type of architecture. There are three main types of transformer architectures that can be used for natural language processing and specifically for language-to-language generation, like in our case:

- **Encoder-Only or Autoencoding Transformer:** This part of the transformer aims to capture meaningful representations of input data and squeeze them on an encoded form. This type of model is particularly strong towards language understanding and classification tasks. The training process of autoencoding models (or simply autoencoders) often occurs on a bidirectional fashion, which means that they consider both the forward and backward context of the input sequence and thus it captures dependencies within the text more effectively. Additionally, autoencoders employ masking techniques to deliberately hide or corrupt certain parts of the input during training. This procedure forces the model to learn robust features, rendering it with the ability of retrieving any missing information. This type of models, include the *BERT family* [18], which became a mainstays in a plethora of classifications and summarization-oriented tasks.
- **Decoder-Only or Autoregressive Transformer:** Autoregressive models, that rely solely on the decoder of the transformer, leverage probabilistic inference to predict the next token iteratively by depending also at one prior token. Unlike sequence-to-sequence models, autoregressive models don't

<sup>6</sup>In contrast to *Jain et al.* [46] we are going to use a transformer-based model instead of sequence-to-sequence RNNs.

require an explicit input sequence and are suitable for text generation tasks. Among the most well-known autoregressive architectures, is the *GPT-series* [79], which have gained immense popularity on tasks such as text generation, sequence-to-sequence modeling or even time series forecasting.

- **Encoder-Decoder Transformer:** Combines some strengths from both of the former categories. The encoder processes the input text, capturing any contextual information and the decoder generates sequential output based on the encoded information. This architecture is suitable for tasks, when understanding the input context and generation of a coherent sequence is needed and thus it's applicable for tasks like text-to-text generation. A very popular Encoder-Decoder transformer is the *T5 model* [81].

Due to the demand of generating a narrative story from a given input text (captions), we will stick with the complete Encoder-Decoder architecture. In particular, in the *encoding phase*, the transformer's encoder will process each caption independently, creating a high-dimensional representation for each of those. Using a *Self-Attention* mechanism we want to weigh the importance of different words in each caption, hoping that the model will capture the relationships between words and phrases. In the *decoding stage*, the purpose is to generate the output sequence (the narrative story) autoregressively, one word at a time. As has been employed in numerous prior works, the decoder should consider both the encoded captions and the words it has generated so far, to produce the next word in the sequence. A high-level illustration of an Encoder-Decoder Transformer that could be deployed for our mission is given in Fig. 5.

The last step of this project, as already mentioned during our research questions (sub-section 1.4), will be the evaluation phase. We will quantify the results of the model via some automatic metrics like METEOR, BLUE, SPICE and CIDEr. Albeit, in order to ensure more coherence, fluency, and relevance, human evaluation will also take place in the form of crowdsourcing.

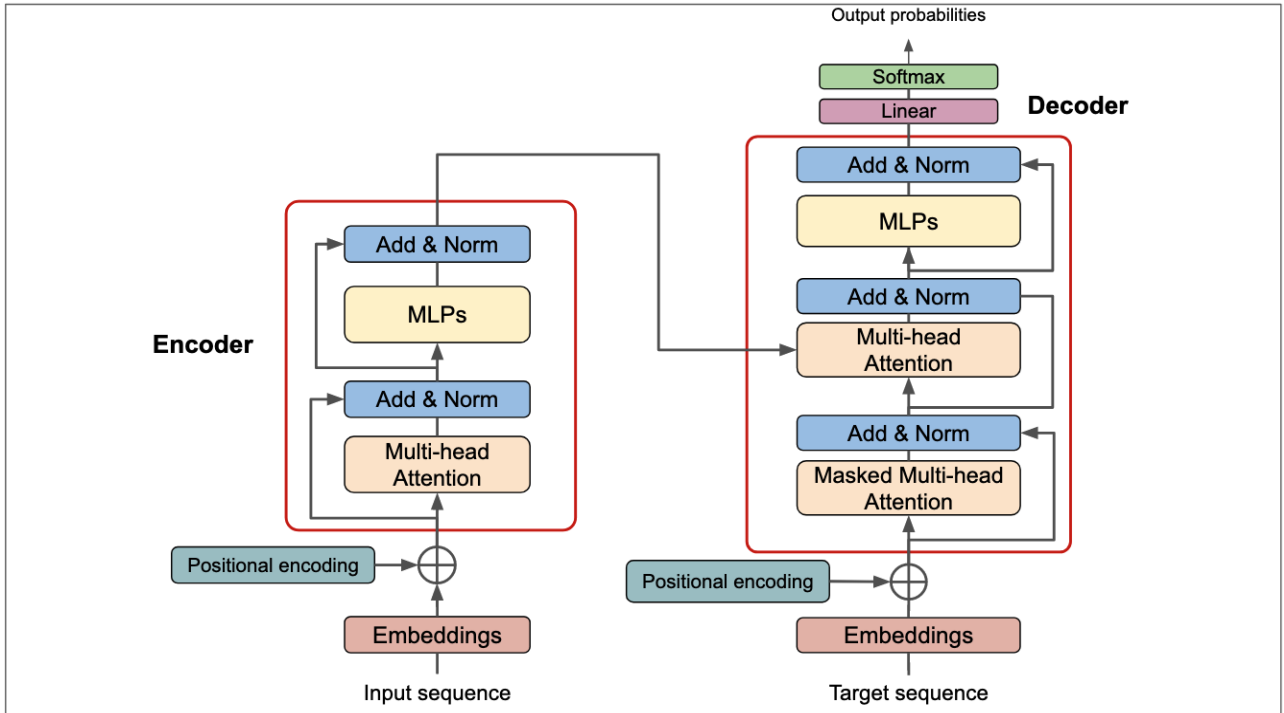


Figure 5. The Transformer Architecture [73] (adapted from [98])

## 4 Research Timeline

In the last section of our proposal we will give a brief and approximate plan (in terms of months) of how the remainder of this project will unfold:

- **March:** Settle down the environment on the SURF server, learn how to navigate with the remote connection on Linux and load the data (images).
- **April:** Load the model (Clip-Cap) to the server, familiarize and make use of it in order to implement the first part of the project, i.e. the producing of captions for the images of VIST dataset.
- **May:** Evaluate Clip-Cap captions on the VIST dataset (Sub-RQ 1). Start implementing from scratch the Encoder-Decoder Transformer architecture so to convert the isolated captions to stories.
- **June:** Complete the Text-to-Text model and start evaluating the generated stories with the automatic metrics (Sub-RQ 2). Come up with a strategy of how human evaluation through crowdsourcing could be applied for our purpose. Start writing the report of the project.
- **July:** Continue with the (mainly the human) evaluation since it demands more time and synchronization. Finalize the project and the report.
- **August:** Submit the Project and Defend the Thesis.

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, et al. Cm3: A causal masked multimodal model of the internet. *arXiv preprint arXiv:2201.07520*, 2022.
- [3] Harsh Agrawal, Arjun Chandrasekaran, Dhruv Batra, Devi Parikh, and Mohit Bansal. Sort story: Sorting jumbled images and captions into stories. *arXiv preprint arXiv:1606.07493*, 2016.
- [4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.
- [5] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016.
- [6] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.
- [7] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.
- [8] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [9] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.
- [10] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.
- [11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [12] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snaveley, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [13] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
- [14] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *open review*, 2019.



- [15] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- [16] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005.
- [17] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017.
- [18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [19] Jacob Devlin, Saurabh Gupta, Ross Girshick, Margaret Mitchell, and C Lawrence Zitnick. Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*, 2015.
- [20] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.
- [21] Desmond Elliott and Frank Keller. Image description using visual dependency representations. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1292–1302, 2013.
- [22] Le Fang, Tao Zeng, Chaochun Liu, Liefeng Bo, Wen Dong, and Changyou Chen. Transformer-based conditional variational autoencoder for controllable story generation. *arXiv preprint arXiv:2101.00828*, 2021.
- [23] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision—ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5–11, 2010, Proceedings, Part IV 11*, pages 15–29. Springer, 2010.
- [24] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.
- [25] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6639–6648, 2019.
- [26] Pablo Gervás, Belén Díaz-Agudo, Federico Peinado, and Raquel Hervás. Story plot generation based on cbr. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 33–46. Springer, 2004.
- [27] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. *Advances in neural information processing systems*, 27, 2014.
- [28] Diana Gonzalez-Rico and Gibran Fuentes-Pineda. Contextualize, show and tell: A neural visual storyteller. *arXiv preprint arXiv:1806.00738*, 2018.
- [29] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

- [30] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.
- [31] Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. A knowledge-enhanced pre-training model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108, 2020.
- [32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [33] Sen He, Wentong Liao, Hamed R Tavakoli, Michael Yang, Bodo Rosenhahn, and Nicolas Pugeault. Image captioning through image transformer. In *Proceedings of the Asian conference on computer vision*, 2020.
- [34] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *Advances in neural information processing systems*, 32, 2019.
- [35] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, nov 1997.
- [36] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.
- [37] Chao-Chun Hsu, Zi-Yuan Chen, Chi-Yang Hsu, Chih-Chia Li, Tzu-Yuan Lin, Ting-Hao Huang, and Lun-Wei Ku. Knowledge-enriched visual storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7952–7960, 2020.
- [38] Chi-Yang Hsu, Yun-Wei Chu, Ting-Hao’Kenneth’ Huang, and Lun-Wei Ku. Plot and rework: Modeling storylines for visual storytelling. *arXiv preprint arXiv:2105.06950*, 2021.
- [39] Ting-Yao Hsu, Chieh-Yang Huang, Yen-Chia Hsu, and Ting-Hao’Kenneth’ Huang. Visual story post-editing. *arXiv preprint arXiv:1906.01764*, 2019.
- [40] Junjie Hu, Yu Cheng, Zhe Gan, Jingjing Liu, Jianfeng Gao, and Graham Neubig. What makes a good story? designing composite rewards for visual storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7969–7976, 2020.
- [41] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4634–4643, 2019.
- [42] Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*, 2016.
- [43] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 958–959, 2020.
- [44] Ilija Ilievski, Shuicheng Yan, and Jiashi Feng. A focused dynamic attention model for visual question answering. *arXiv preprint arXiv:1604.01485*, 2016.
- [45] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.

- [46] Parag Jain, Priyanka Agrawal, Abhijit Mishra, Mohak Sukhwani, Anirban Laha, and Karthik Sankaranarayanan. Story generation from sequence of independent short descriptions. *arXiv preprint arXiv:1707.05501*, 2017.
- [47] Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. Stay on the path: Instruction fidelity in vision-and-language navigation. *arXiv preprint arXiv:1905.12255*, 2019.
- [48] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.
- [49] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10267–10276, 2020.
- [50] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Denscap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574, 2016.
- [51] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.
- [52] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019.
- [53] Taehyeong Kim, Min-Oh Heo, Seonil Son, Kyoung-Wha Park, and Byoung-Tak Zhang. Glacnet: Glocal attention cascading networks for multi-image cued story generation. *arXiv preprint arXiv:1805.10973*, 2018.
- [54] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *arXiv preprint arXiv:1411.7399*, 2014.
- [55] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–325, 2017.
- [56] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.
- [57] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [58] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.
- [59] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. Entangled transformer for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8928–8937, 2019.
- [60] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

- [61] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.
- [62] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.
- [63] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020.
- [64] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- [65] Danyang Liu and Frank Keller. Detecting and grounding important characters in visual stories. *arXiv preprint arXiv:2303.17647*, 2023.
- [66] Yu Liu, Jianlong Fu, Tao Mei, and Chang Wen Chen. Let your photos talk: Generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.
- [67] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.
- [68] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383, 2017.
- [69] Ziyang Luo, Yadong Xi, Rongsheng Zhang, and Jing Ma. A frustratingly simple approach for end-to-end image captioning. *arXiv preprint arXiv:2201.12723*, 2022.
- [70] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27, 2014.
- [71] Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: Clip prefix for image captioning, 2021. [2](#).
- [72] Md Sultan Al Nahian, Tasmia Tasrin, Sagar Gandhi, Ryan Gaines, and Brent Harrison. A hierarchical approach for visual storytelling using image description. In *Interactive Storytelling: 12th International Conference on Interactive Digital Storytelling, ICIDS 2019, Little Cottonwood Canyon, UT, USA, November 19–22, 2019, Proceedings 12*, pages 304–317. Springer, 2019.
- [73] Jean Nyandwi. The transformer blueprint, 2023. Accessed: March 1, 2024.
- [74] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing 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, 2002.
- [75] Cesc C Park and Gunhee Kim. Expressing an image stream with a sequence of natural sentences. *Advances in neural information processing systems*, 28, 2015.
- [76] Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. A straightforward framework for video retrieval using clip. In *Mexican Conference on Pattern Recognition*, pages 3–12. Springer, 2021.

- [77] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020.
- [78] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [79] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *mike captain*, 2018.
- [80] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [81] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.
- [82] Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjieva. Smallcap: lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2840–2849, 2023.
- [83] Hareesh Ravi, Kushal Kafle, Scott Cohen, Jonathan Brandt, and Mubbasir Kapadia. Aesop: Abstract encoding of stories, objects, and pictures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2052–2063, 2021.
- [84] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.
- [85] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017.
- [86] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.
- [87] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021.
- [88] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [89] Jing Su, Qingyun Dai, Frank Guerin, and Mian Zhou. Bert-hlstm: Bert and hierarchical lstm for visual storytelling. *Computer Speech & Language*, 67:101169, 2021.
- [90] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019.
- [91] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.
- [92] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

- [93] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.
- [94] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17928, 2022.
- [95] Mariët Theune, Sander Faas, Anton Nijholt, and Dirk Heylen. The virtual storyteller: Story creation by intelligent agents. In *Proceedings of the Technologies for Interactive Digital Storytelling and Entertainment (TIDSE) Conference*, volume 204215, page 116, 2003.
- [96] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.
- [97] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.
- [98] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [99] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.
- [100] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015.
- [101] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014.
- [102] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.
- [103] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7622–7631, 2018.
- [104] Eileen Wang, Caren Han, and Josiah Poon. Rovist: Learning robust metrics for visual storytelling. *arXiv preprint arXiv:2205.03774*, 2022.
- [105] Xin Wang, Wenhui Chen, Yuan-Fang Wang, and William Yang Wang. No metrics are perfect: Adversarial reward learning for visual storytelling. *arXiv preprint arXiv:1804.09160*, 2018.
- [106] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.
- [107] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.
- [108] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.



- [109] Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. Megatron-cntrl: Controllable story generation with external knowledge using large-scale language models. *arXiv preprint arXiv:2010.00840*, 2020.
- [110] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10685–10694, 2019.
- [111] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016.
- [112] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, pages 4507–4515, 2015.
- [113] Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385, 2019.
- [114] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699, 2018.
- [115] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016.
- [116] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.
- [117] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.
- [118] Licheng Yu, Mohit Bansal, and Tamara L Berg. Hierarchically-attentive rnn for album summarization and storytelling. *arXiv preprint arXiv:1708.02977*, 2017.
- [119] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021.
- [120] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*, 2017.
- [121] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13041–13049, 2020.
- [122] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020.