Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

From Pixels to Sentence: A Comprehensive Study of Transformers-Based Models for Image Captioning

Haider Jaber Samawi [a,*], Ayad Rodhan Abbas [b]

[a] *Department of Computer Science, Faculty of Computer Science and Mathematics, University of Kufa, Najaf, Iraq. Email: haiderj.alshimmary@student.uokufa.edu.iq.*
[b] *Department of Computer Science, University of Technology, Baghdad, Iraq. Email: ayad.r.abbas@uotechnology.edu.iq.*

ARTICLEINFO

Article history:
Received: 03 /11/2025
Rrevised form: 06 /12/2025
Accepted : 14 /12/2025
Available online: 30 /03/2026

Keywords: Image Captioning, Transformers, Vision-Language Models, Multimodal Learning, Deep Learning.

ABSTRACT

The task of image captioning, which involves generating descriptive textual content from visual input, is a pivotal challenge in multimodal learning. This research delves into the advancements in image captioning facilitated by Transformer-based models, comparing their performance, architectures, and innovations across various tasks. Traditional models, such as CNNs paired with RNNs, were initially used to extract visual features and generate corresponding captions. However, the introduction of Transformer architectures has significantly enhanced the performance of image captioning systems, allowing for more coherent, context-aware, and grammatically correct captions. This paper explores the evolution of Transformer-based models, with a particular focus on the Encoder-Decoder, Vision-Language Fusion, and End-to-End Transformers models. By analyzing state-of-the-art architectures such as ViT, GPT, BLIP, and CoCa, the study demonstrates how these models address long-range dependencies, utilize self-attention mechanisms, and seamlessly integrate vision and language for improved caption generation. Furthermore, the paper evaluates the strengths, challenges, and limitations of these approaches, including issues related to computational complexity, dataset biases, and caption diversity. Ultimately, this study presents a comprehensive comparison of these models, offering insights into future research directions in the field of image captioning.

MSC..

https://doi.org/10.29304/jqcsm.2026.18.12477

1. Introduction

In the current era of information proliferation, the number of images generated, caught, and spread daily is still significantly increasing. Most of the images used on social media sharing sites have relevant text descriptions. The automated generation of such descriptions has considerable meritorious applications in a wide range of applications such as web image indexing and retrieval, content-based image search engines, support for the visually impaired, and others [1]. Text description from a given image is a task that is often called image captioning. Image captioning has an enormous potential for the development of human-centered artificial intelligence systems (AI) and the introduction of an additional modality - i.e. comprehension of visual content [2]. Nevertheless, the task proves to be a formidable challenge, as it often includes two intricate subtasks, the interpretation of the content of the input image, and the generation of grammatically correct, semantically meaningful, and contextually relevant sentences that describe the content of the image in detail [3]. This paper presents a complete review of Transformer-based models that combat these issues, from the underlying basic architectures to modern multimodal models.

∗Corresponding author: Haider Jaber Samawi

Email addresses: haiderj.alshimmary@student.uokufa.edu.iq

Communicated by ‘sub etitor’

2. General Architecture of Transformer

The Transformer architecture is a revolutionary concept in the field of deep learning architecture, introduced in the seminal work "Attention Is All You Need" by Vaswani et al. in 2017, which introduces the concept of attention mechanism and increases the performance of sequence-to-sequence tasks like machine translation and text summarization. The key to the success of the Transformer is the ability to process input sequences in parallel, as opposed to the sequential processing of traditional recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) [4]. This parallelism not only makes the training process faster but also makes it easier for the model to learn long-range dependencies in the data. While the basic building blocks of the standard Transformer namely, the encoding stack, decoding stack and multi-attention mechanisms as shown in figure 1, are well laid out, their application in image captioning requires specific architectural modifications [5], [6]. Specifically, the ability of self-attention mechanisms to model relationships between visual patches and textual tokens has become the cornerstone of modern multimodal learning [7], [8]. Therefore, rather than reiterating the basic mathematical formulations of the Transformer, this study focuses on how these components are leveraged and modified within image captioning frameworks.

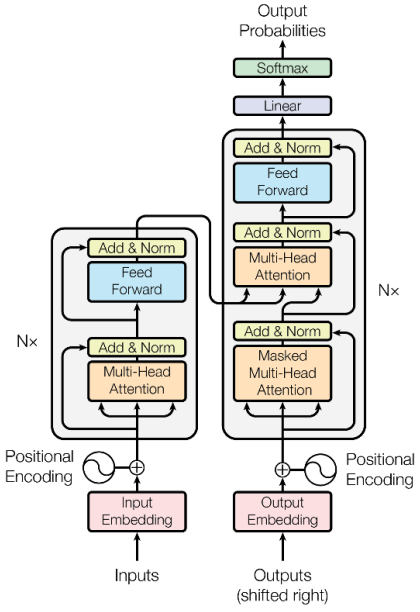

Figure 1. Transformer Architecture [4].

3. Basic Categories in Image Captioning

The field of image captioning transformed significantly after Transformer-based architectural designs became available. Before Transformers replaced the traditional image captioning systems which used convolutional neural networks coupled with recurrent neural networks. Image captioning systems became more effective when Transformer models entered the field. The capability of Transformers to identify extensive contextual relationships within visual and textual data leads to improved caption coherence and relevant contextual understanding. [9]. Furthermore, Image captioning now uses Contrastive Learning as one of its most promising solution approaches. Models under the CLIP name employ contrastive loss functions to establish picture and textual relations. The training process of these models achieves better relevant caption generation by reducing positive pairs' distance while increasing negative pairs' separation. Contrastive learning methods enable models to understand image salient points in order to generate captions that match image semantic content better [10], [11]. To provide a structured analysis, we classify these approaches into distinct architectures, starting with the fundamental Encoder-Decoder framework.

3.1. Encoder-Decoder Architecture

Encoder-Decoder models are one of the most fundamental architectures for image captioning. These models are based on the idea of mapping an input image to a coherent caption by two different components. The Encoder takes the input image and extracts high-level visual information, generally a Convolutional Neural Network (CNN)

or a Vision Transformer (ViT) is used to convert the raw pixel data to a feature vector [12]. Once this, the Decoder takes these encoded features in the visual and produces a textual description. A Transformer-based model like GPT or BERT, takes the visual features as input and turns them into a sequence of words to create the generated caption which makes the text generated consistent with the features extracted during the training [13], [14].

3.2. Encoder-Decoder Models

The following section is mainly devoted to ViT (Vision Transformer) and GPT (Generative Pre-trained Transformer), which are the key models of Encoder-Decoder architecture [15]. The analysis will focus on these two models but also include short contributions to SimVLM and GIT as these models combine vision and language functions.

3.2.1. ViT + GPT

In the ViT + GPT paradigm, the image will first be processed with a Vision Transformer (ViT), which will divide the image into patches and encode them into feature vectors. The GPT model - an autoregressive, Transformer-based language model - is then utilized to generate the image caption from these visual features. This method utilizes the strong power of Vision Transformers for image understanding and GPT for language generation [16].

A. ViT (Vision Transformer)

Dosovitskiy et al. proposed Vision Transformer (ViT) in 2020 as a deep learning framework to perform image classification using Transformers as a variant of the traditional Convolutional Neural Network (CNN) [17].

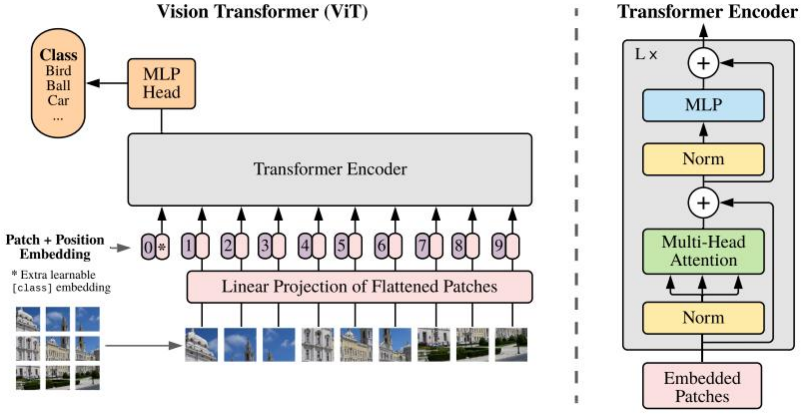

Figure 2. ViT Architecture [16].

Operatively, ViT is a modification of the Transformer architecture to image analysis. ViT initiates its workflow through image patch division, where an image (e.g., 224x224 pixels) becomes divided into patches (e.g., 16x16 segments). Linear embeddings are computed from the flatten one dimensional vectors of these patches. Given that Transformers inherently lack spatial awareness, ViT adds positional encodings to each processed patch, so that the model can understand the relative positions of patches. Each encoder layer has self-attention mechanisms, that help find the relationship between patches, thereby capturing the correlation between images in both local and global regions. The resultant output consists of a series of embeddings that summarize the entire of the contents of the image, and hence supply GPT with the necessary information to produce a suitable caption [18], [19].

Analysis of ViT:

1. Global Attention: The self-attention mechanism in ViT helps in detecting remote relationships in the image, which is not possible in the traditional CNN systems [20].

2. Scalability: ViT shows very good performance when trained in large datasets, outperforming CNN - based solutions on large-scale data [21].

3. Limitations: The quality of the performance of ViT is highly dependent on access to large annotated data [22]. Moreover, self-attention architecture poses a high computational demand especially when processing high-resolution images [23].

B. GPT (Generative Pretrained Transformer)

The Generative Pre-Trained Transformer (GPT) is one of the top natural language processing (NLP) models that performs well in text generation tasks. GPT uses a Transformer architecture with a decoder architecture to predict the next token by taking in the existing context [24]. The GPT mechanism involves the processing of tokens in a sequential manner, tokenization is the process of breaking down content into discrete tokens, and the model processes the tokens sequentially to build contextual understanding [25]. Through pre-training on large corpora using unsupervised learning, GPT learns to predict future tokens and is well generalized to a range of tasks. During inference, GPT uses its self-attention mechanism to discover dependencies between the tokens, which will prove to work well when combined with the ViT image encoder, which associates visual features with linguistic tokens [26]. The signature feature of GPT is its autoregressive generation capability, which generates the next token in its sequence after considering the previous tokens one by one. Finally, supervised fine-tuning helps the model to produce accurate captions with correct correspondence to the content of new images [27].

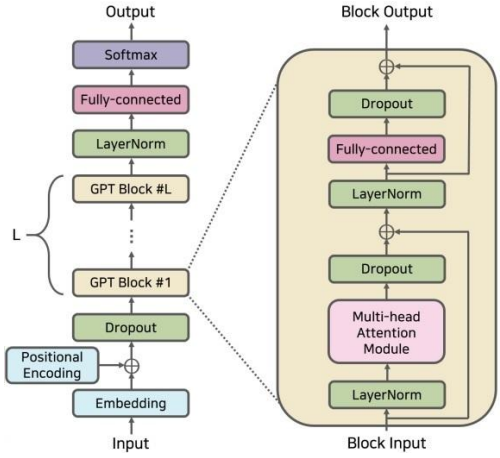

Figure 3. GPT Architecture [28].

Analysis of GPT:

1. High-Quality Text Generation GPT is able to generate text with seamless fluency and context-sensitive relevance [29].

2. Flexibility: It is easy to extend GPT to multimodal processing by training it with visual encoders like ViT [30].

3. Limitations Since GPT is trained on internet derived corona, (it may internalize and reproduce pre-existing biases). Further, natural language processing (NLP) skills are not built into GPT; instead, its power comes solely from the feature provided by the visual encoder [31].

3.2.2. SimVLM

The SimVLM model (Simple Visual Language Model) makes the process more simple and easy by using a joint representation space for both vision and language. It is a combination of a Transformer-based vision model like ViT and a language model like GPT that allows the model to learn how to map images directly to captions without the need for separate feature extraction and caption generation steps [32].

3.2.3. GIT (Generative Image-to-Text Transformer)

The GIT model works as an encoder decoder model that uses a Transformer component that analyses images and generates a textual output. The system processes images and their captions in a single unified workflow, eliminating separate feature extraction protocols. The integrated visual and textual understanding of this unified process gives huge benefits as per this model [33], [34].

3.3. Vision-Language Fusion

In the development of multimodal artificial intelligence systems, Vision-Language Fusion stands as the main research focus because it aims to build models which bridge visual and textual data effectively. Unlike Encoder-Decoder models that process modalities sequentially, fusion models start from separate image and text data which integrates information between modalities into a combined feature representation space [35]. Vision-language fusion directly models intramodality connections to produce more precise captions that align with context. Joint embeddings represent a method to achieve visual integration with text information by projecting the two modality spaces into one shared dimensional structure that improves relationship learning potential The growth in size and quality of these datasets, like MS COCO and Visual Genome, has allowed these models to better identify fine-grained image- text relationships and therefore achieve improved performance on these tasks [36]. We consider three well-known models which define this category, namely Oscar, BLIP and VinVL.

3.3.1. Oscar (Object-Semantics Aligned Pre-training)

Oscar is a vision-language model that is used to strengthen the relationship between visuals and text data by the implementation of object-syntactical tags. The model uses a system based on object detection tags as semantic tokens to implement semantic pre-training [37]. These special object tags serve as key elements for matching up graphical data and its textual descriptions: The model becomes more skilled in producing detailed and contextually accurate captions if it associates identified objects with linguistic elements [38].

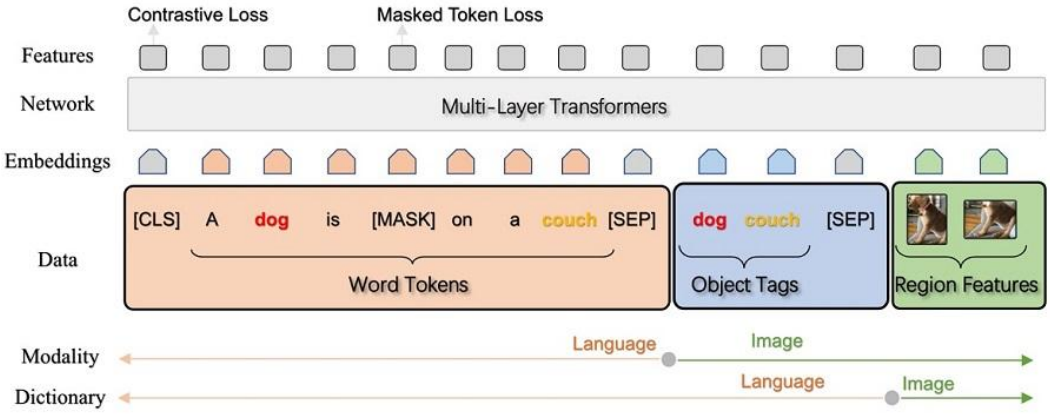

Figure 4. Architecture of OSCAR Model [39].

Operationally, the model combines visual features with text tokens. The visual features are converted into vectors before the Transformer system is combined with linguistic descriptions [40]. The process of training enables the model to predict the captions from the combination of the features of the images and the object tags that reinforce the text of the image description [41]. As a result, the model produces captions describing the objects in the image and the relationships between them, which increases the accuracy and complexity [42].

Analysis of Oscar:

1. Strengths: The main strength of Oscar is in providing the ability to improve the interpretation of image contents by direct matching between semantic objects and captions. By associating these captions, the model recognizes specific objects in images to give more descriptive and detailed summaries [43].
2. Limitations: The model works on object detection and this approach introduces some challenges into using them. During pre-training, there can be detrimental effects of incorrect object detection by the detection

system to the accuracy of caption [44]. In addition, images that contain abstract art, overlapping objects, or subtle contextual cues could pose challenges for the object detection models, which may result in less accurate outcomes [45].

3.3.2. BLIP (Bootstrapped Language-Image Pre-training)

The use of BLIP is a significant achievement in the field of vision and language modeling and employs a new type of bootstrapping to solve vision and language integration issues. BLIP follows a different approach to training since it employs unsupervised bootstrapping methods to build the system iteratively after abandoning supervised object detection and other language-image alignment approaches used in traditional systems [9]. The model conducts dynamic learning through successive improvement of noisy captions to enhance its ability to understand images together with textual content [46].

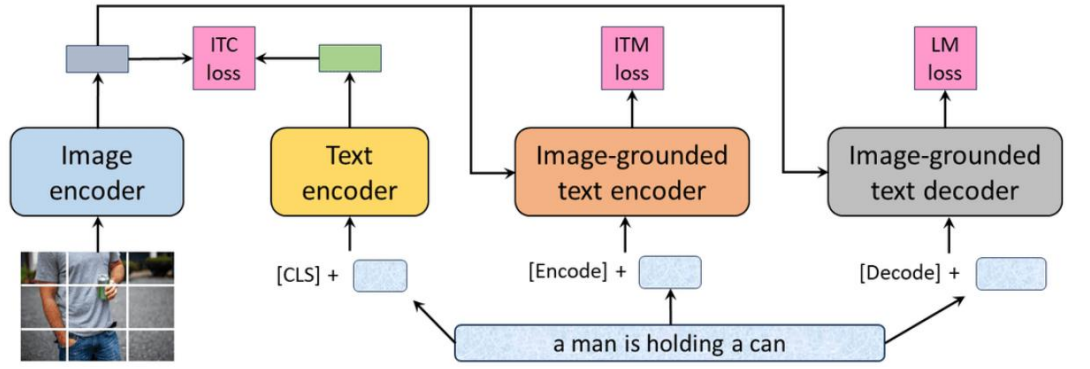

Figure 5. Architecture of BLIP model [47].

BLIP is characterized by its iterative refinement process. The foundation of BLIP's system contains two essential elements that include the Vision Transformer (ViT) to process visual elements alongside the Transformer-based language model that generates textual outputs. The traditional CNN-based approaches are replaced by ViT to enable a model that grasps images at a global scale. Bootstrapping together with joint learning allows BLIP to consistently improve its output which results in superior caption generation for images the model has not encountered before. [48], [49]. The initial output captions from BLIP contain specific types of language errors when dealing with images during training; however, the model receives captions which are subsequently compared to reference captions for refinement purposes before gaining better prediction accuracy through iterative processes [50].

Analysis of BLIP:

1. Performance: Benchmarks show that BLIP shows exceptional performance (its successful caption generation for COCO and Flickr30k test platforms) compared to other models [51].
2. Limitations: While BLIP shows great power in understanding context, it has its limitations due to its requirement for heavy computational power and large-scale annotated caption datasets. The production of high quality captioning takes some time during its bootstrapping procedure and requires a lot of time and resources for well performing results [52].

3.3.3. VinVL (Enhanced Visual Representations for Vision-Language Models)

The VinVL model (Visual in Vision- Language) uses design improvements in visual representations to more effectively improve performance in vision-language tasks, specifically image captioning [53]. VinVL is different from the earlier methods that combine image and text features as it focuses on enhancing visual features before combining them with textual information. The improved feature-processing mechanism allows VinVL to have better visual detail detection capability to improve the performance of diverse multimodal tasks [54], [55].

The architecture difference between VinVL is its visual backbone. The core of VinVL has a deeply trained ResNet architecture, the result of processing a large dataset, which gives the advantage of better visual feature extraction compared to standard frameworks. The pre-trained ResNet is used to extract high-quality visual features that are

then combined with the text information through a Transformer-based structure in a common feature space. VinVL brings two robust representations, visual features and language representations, to produce caption outputs that better represent the content of images and context information [56].

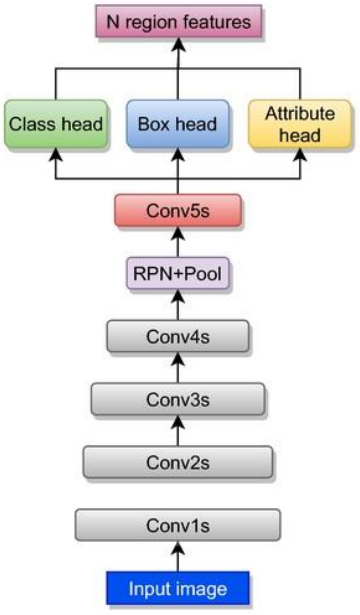

Figure 6. Architecture of VinVL model [57].

Analysis of VinVL:

1. Strengths The greatest advantage of using VinVL stems from the fact that it is effective in the treatment of images where there are complex elements, including fine details. VinVL enhances the image captioning performance by achieving appropriate representation even for pictures that do not have clear object tags via its bootstrapping-free approaches [58].
2. Limitations: Visual feature extraction at scale is tough for the model because it requires a lot of computational power especially when it has to process high-resolution images. The performance of this model may be reduced in two specific conditions; limited textual annotations and images without prominent visual elements [59].

3.4. End-to-End Transformers

The End-to-End Transformer system combines two critical image captioning processes in a single unified system, unlike standard practices, which use different structural elements for encoding and decoding. With End-to-End Transformers, images processing and text processing are simultaneously done in one Transformer model, thus increasing efficiency in the caption generation process [60], [61]. The main advantage of this architecture is that it enables visual features to link directly with created text through a unified model structure without the necessity for intermediate feature extraction and that allows the model to discover correlations between visual and textual elements as a unified learning process [62]. As a result, these models have the benefit of sophisticated attention mechanisms that allow it to pay attention to parts of the image that are associated with certain caption parts, which can be useful for image descriptions of images with complex structures [63]. The system is a direct processing module that monitors visual content and linguistic details simultaneously and thus allows for optimal performance in both areas. When applied to caption generation, End-to-End Transformers have proven superior to traditional systems thanks to the ability to integrate images as well as text descriptions in an integrated fashion [64], [65].

3.4.1. CoCa: Contrastive Captioner Transformer

CoCa uses contrastive learning techniques to support its vision language model and thus produce accurate image captions. The contrastive method employed by CoCa differs from the classical captioning models, which are based on generative methods, in reducing the distances between mismatched image-caption pairs rather than reducing the distances between matched image-caption pairs [66].

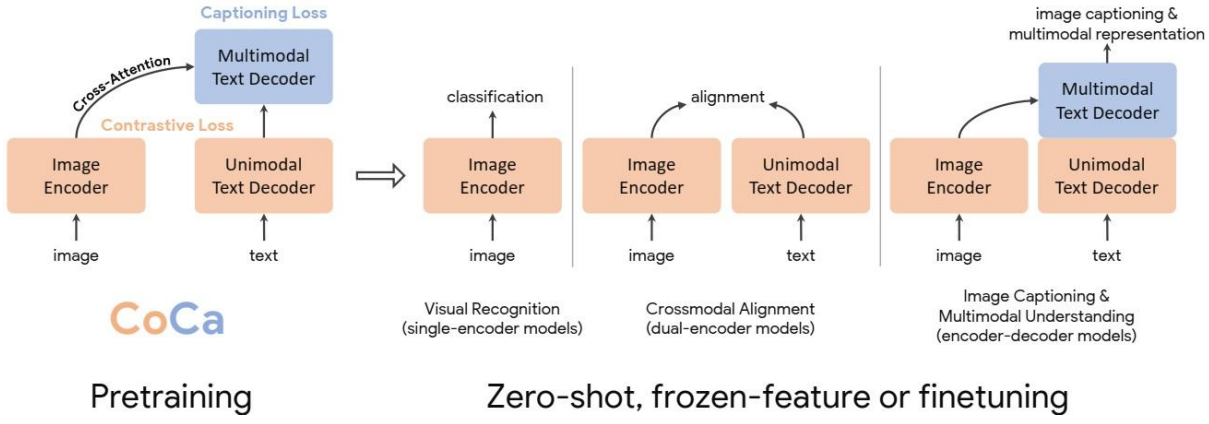

Figure 7. Architecture of CoCa Model [67].

Architecturally, the CoCa system consists of two main parts: a Vision Transformer (ViT) as the image processing part, and another transformer-based language model for caption generation. During training CoCa maximizes similarities between visual features and their corresponding captions by using a contrastive loss function [68]. This loss helps in creating high-quality captions by making the model learn the exact relationships between the image and is description, which improves its discriminative power between the positive and negative image-description relationships [69].

Analysis of CoCa:

1. Strengths: The major strength of CoCa is that it can simultaneously learn from correctly and incorrectly labeled training examples. CoCa achieves enhanced discriminative power which leads to better image captioning capabilities [70].
2. Limitations: CoCa encounters difficulties mainly because it demands extensive datasets to deliver optimal contrastive learning while simultaneously requiring substantial computing power for training its significant models [71], [72].

3.5. Contrastive Learning

The machine learning technique Contrastive Learning serves diverse applications within Vision-Language models to learn by comparing opposite and distinct pairs [73]. Image captioning training with contrastive learning teaches models to recognize appropriate captions for pictures alongside performing separation of matching and non-matching pairs. Similar images with matching captions should exist within close proximity in the representation space and dissimilar images with mismatched captions should exist at distant positions [74]. The effectiveness of contrastive learning grows when labeled data is scarce because it becomes a powerful solution for image captioning combined with visual question answering [75].

3.5.1. CLIP (Contrastive Language-Image Pretraining)

OpenAI built CLIP as a strong model to apply contrastive learning methods for uniting visual information with linguistic data. CLIP undergoes training with massive image datasets containing their matching text descriptions in order to develop its ability to match pictures with textual content representations. A dual-encoder system comprises CLIP by implementing ViT as a Vision Transformer and a text encoder functioning from the Transformer family of

models. The algorithm places the vectors originating from image data together with text data features into one common space to enable relationship analysis between these elements [76], [77].

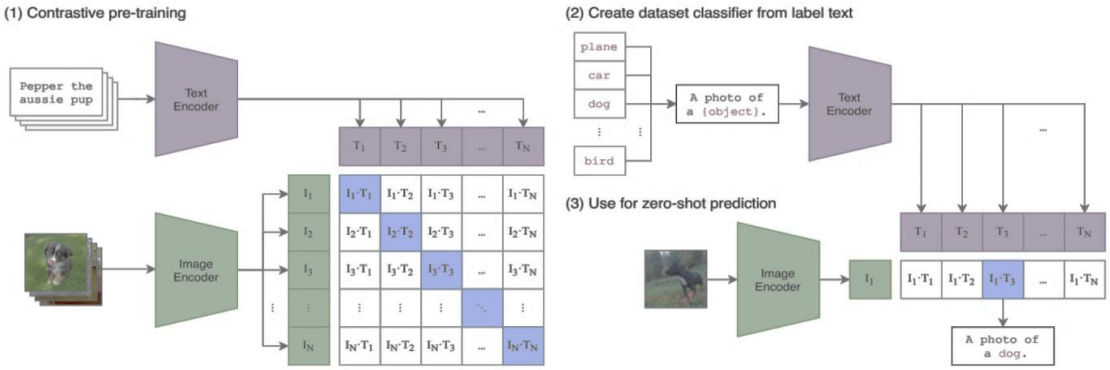

Figure 8. Architecture of CLIP Model [78].

Analysis of CLIP:

1. Strengths: The major benefit of CLIP lies in its capability to apply to diverse vision tasks while avoiding the need for specific task-based training procedures [79]. Visual along with textual information processing enables CLIP to produce image captions efficiently through highly flexible methods [80].
2. Limitations: The performance of CLIP suffers when it encounters abstract or highly complex images. The model performs effective generalization across most datasets but its quality of output depends critically on the quality of training textual data [81].

3.5.2. ALIGN: A Large-Scale Image-Text Model

Google Research developed ALIGN (A Large-scale ImaGe and Noisy-text model) as a Vision-Language model which effectively conducts image-text alignment. The model uses a contrastive learning setup like CLIP to position images with their textual descriptions in one vector space. However, the training process of ALIGN differs from CLIP because it operates on a massive web-scraped data collection of noisy text with unstructured formatting [82]. This massive training process allows ALIGN to extract better data representations from imperfect datasets [83].

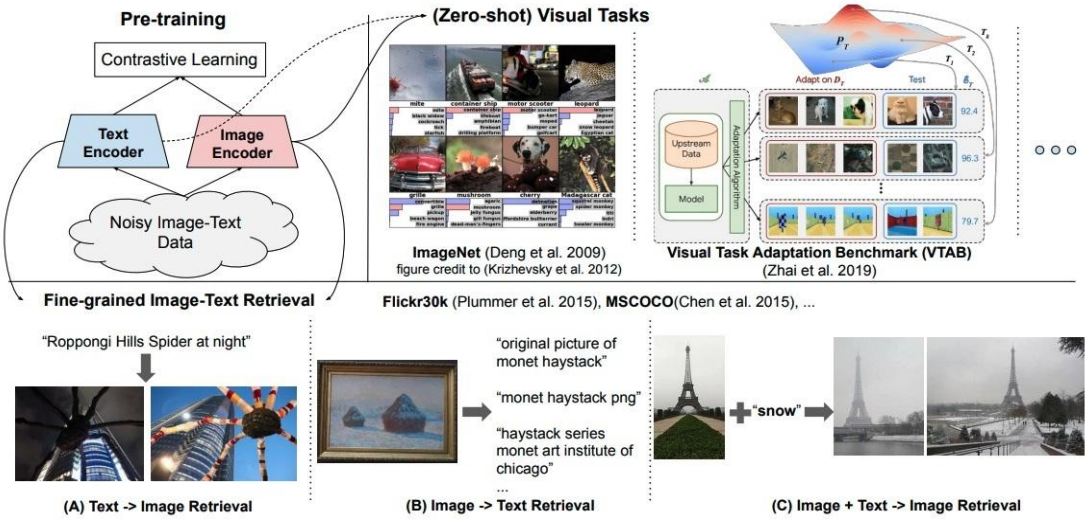

Figure 9. Summary of ALIGN Architecture [84].

Analysis of ALIGN:

1. Strengths: The main strength of the ALIGN system is its capability to process enormous datasets. ALIGN learns adaptable representations with generalization abilities by processing extensive noisy text obtained from online sources, demonstrating an advantage over alternative models since it absorbs knowledge from messy unorganized textual data [85].
2. Limitation: Training system is limited by processing noise textual data, which can lead to mismatch between images and captions. In the generation, the model may generate descriptions that do not accurately reflect the content and context of the image [86]. In addition, the need for large data processing resources prevents the implementation of ALIGN in resource-limited computing environments [87].

4. Popular Data Sets

High quality datasets form the basis of efficient image captioning models, in that they establish various visual and text pairs required to train on. The process of training the image encoders and decoder driven caption generation essentially requires large datasets with many images paired with their corresponding expressions [88]. The highly popular image captioning dataset provides robust stability and large scale content while extending its functionality by multiple versions that support dense captioning and image paragraph recognition [89].

4.1. Flickr8K

Flickr8k provides 8,000 images with free-form descriptions linked to each of them. The dataset includes five descriptions for each image which were produced by human annotators [90]. The displayed images within the dataset showcase numerous subjects which appear throughout different environmental settings such as wildlife scenes as well as landscapes and activities involving human participants [91]. Because of its compact size the dataset delivers fast experimental feedback thus allowing quick beneficial analysis experiments. Using Flickr8k in research enables fundamental model testing to help evaluate these configurations for simple image captioning systems [92], [93].

Figure 10. Images with their captions from Flickr8K dataset [94].

4.2. Flickr30k

Flickr30k is a significant development in image captioning datasets because it lays a firm groundwork after the launch of limited datasets related to OCR. The creators of this dataset based Flickr30k on the previous Flickr8k whereas they added 31,000 Flickr images supported by five crowd-sourced captions distributed equally among them [95]. The dataset demands detailed understanding of the depicted scenes because its captions present complex descriptions beyond basic verbalizations that smaller datasets provide [96]. The wide accessibility of Flickr30k makes it suitable for various challenges including VQA and beauty and artist identification beyond its core role in image captioning [97].

Figure 11. Images with their captions from Flickr30k dataset [98].

4.3. MS COCO

MS COCO or Microsoft COCO represents a widely utilized recent large-scale benchmark dataset specifically designed for image captioning purposes since 2015. The collection within MS COCO includes 328,350 images yet the validation and test sets together comprise 100,000 images for training purposes. A main advantage of MS COCO is its dense annotation system with extensive details among its labeling structure [99], [100]. The collection consists of 414,113 captions which correspond to 5 descriptive sentences for each image. The pictures show a minimum of one dominant object with accompanying contextual elements from different scene categories. The database images together with their captions have been extensively validated through widespread research utilization [101].

Figure 12. Images with their captions from the MSCOCO dataset [102].

4.4. Visual Genome

The Visual Genome dataset is a comprehensive collection of images annotated with dense captions describing individual objects, backgrounds, and the relationships among the objects. Each image may include multiple such descriptions [103].

Figure 13. Images with their captions from visual genome dataset [106]

Across the entire dataset, there are over 16 million captions and over 8 million relationships for every pair of objects. Unlike standard captioning datasets, Visual Genome tightly integrates image representations with structured information, creating a scene graph representation capable of describing an image and its part-whole relationships [104]. Having data potentially annotated by multiple humans, each with their expertise and perspective, also provides varied and diverse views of images [105]. Visual Genome continues to expand in scale in terms of segments of the object descriptions and relationships, as well as incorporating new types of annotations such as image attributes [106], [107].

5. Evaluation Metrics in Image Captioning

Effective comparison of image captioning models requires robust quantitative metrics. The assessment of automatic caption output consists of two steps to verify caption relevance to images while checking for proper syntactical organization. Multiple tools have been established to assess generators of text descriptions from images, concentrating their evaluation on quality-based aspects [96].

5.1. BLEU Score

BLEU serves as an established metric which applies n-gram precision methods to compare the automatic output with reference captions. The n-gram precision functions as separate limits throughout the assessment process which extends from unigram to four-grams [108]. BLEU treats equivalent words identically while evaluating n-gram precision because it does not differentiate between synonyms. However, a known limitation is that BLEU ranking depends on the count of n-grams during calculation steps, making it sensitive to random variations [109].

5.2. METEOR Score

Currently the METEOR score operates as a universal measurement tool for evaluating automated caption performance through its assessment of both precision and recall. Image captioning evaluation uses the METEOR score which recognizes word-level exact matches while also considering synonyms and those words reduced to their stem basis [110]. The measurement technique determines image captioning scenario quality by analyzing how well a generated caption matches the reference text semantically. The METEOR score uses the harmonic mean of precision and recall values and gives more weight to recall than precision to produce its final score [111].

5.3. CIDEr Score

The Consensus-based Image Description Evaluation (CIDEr) established itself as a consensus evaluation method when it entered the image captioning field in 2015 [112]. Designed specifically for image captioning, the CIDEr evaluation method gathers human consensus to evaluate captions while it enforces diversity through redundancy penalty structures. The algorithm of CIDEr relies on consensus to calculate similarities by measuring the cosine distance between machine output and human captions. Evaluation requires many human references for each of the generated caption so that a full evaluation is possible [113], [114].

5.4. ROUGE-L Score

ROUGE-L is a general evaluation benchmark for multiple text generation tasks. The ROUGE metrics use very special mathematical formulations to measure specific properties during the evaluation process. The ROUGE-L metric introduces the LCS (longest common subsequence) concept which represents the sequence of words that subsists across two sentences in order [115], [116]. The main goal of ROUGE-L is to assess the performance of words and the order of the sequence in the response that is produced. The conversational nature of sentence structure that is inherent to ROUGE-L makes it the algorithm of choice for quality evaluation of tasks that are reliant on sentence structure [117]. Using both recall and precision in ROUGE-L achieves a more balanced performance measurement by reducing the bias present in using either metric separately [118].

5.5. SPICE Score

The novel evaluation approach SPICE (Semantic Propositional Image Caption Evaluation) evaluates image captions by direct proposition evaluation. Diverging from the n-gram based metrics, the evaluation system changes

its focus from random events to the creation of captions with sequential semantic propositions [119]. SPICE offers a methodological framework that considers the relationship between visuals in an image and the accompanying textual content, and thus allows a better understanding of the meaning of images beyond a lexical comparison. A caption is considered to be accurate and relevant if its semantics precisely represent the interrelations between elements of an image [120], [121].

6. Related Works

To provide an overview of the development of the field, Table 1 summarizes key research on transformer-based image captioning. This chronological compilation places a focus on the transition of model architectures from simple encoder-decoder systems to sophisticated architectures for vision-language fusion and end-to-end systems. It further specifies the particular datasets used and evaluation metrics that were used in each particular study and can be used as a reference for following the development of state-of-the-art solutions and to look for the current trends in the literature.

Table 1. Related works using transformers-based models for image captioning.

No	Reference	Year	Model	Architecture	Dataset Used	Measures
1.	Ref. [34]	2019	Meshed-Memory Transformer	Encoder-Decoder	MS COCO	B4: 39.1, M: 29.2, C: 131.2, S: 22.6
2.	Ref. [122]	2020	X-LAN	Encoder-Decoder	MS COCO	B4: 39.5, M: 29.5 C: 132.0, S: 23.4
3.	Ref. [39]	2020	OSCAR	Vision-Language Fusion	MS COCO, Flickr, Visual Genome	B4: 40.5, M: 29.7 C: 137.6, S: 22.8
4.	Ref. [123]	2021	VinVL	Vision-Language Fusion	MS COCO, Flickr, Visual Genome	B4: 40.9, M: 30.9 C: 140.9, S: 25.2
5.	Ref. [9]	2021	BLIP	Vision-Language Fusion	MS COCO, Visual Genome	B4: 40.4, M: - C: 136.7, S: -
6.	Ref. [124]	2021	ER-SAN	Encoder-Decoder	MS COCO	B4: 40.5, M: 29.8 C: 135.3
7.	Ref [125]	2021	CLIP	Contrastive Learning	WebImageText	Retrieval R1: 59.9
8.	Ref [84]	2021	ALIGN	Contrastive Learning	ALIGN Data	Retrieval R@1: 77.0
9.	Ref. [126]	2021	VIVO	Contrastive Learning	MS COCO, OpenImages	B4: 39.0, C: 127.2
10.	Ref. [127]	2022	GIT	Encoder-Decoder	MS COCO Visual Genome	B4: 42.7, M: 32.2 C: 144.8, S: 25.4
11.	Ref. [128]	2022	SimVLM	Encoder-Decoder	MS COCO, C4	B4: 40.6, M: 33.7, C: 143.3, S: 25.4
12.	Ref. [129]	2022	CoCa	End-to-End Transformers	MS COCO	B4: 40.9, M: 33.9 C: 143.6, S: 24.7
13.	Ref. [130]	2022	DLCT	Encoder-Decoder	MS COCO	B4: 39.8, M: 29.5 C: 133.8, S: 23.0
14.	Ref. [131]	2022	GRIT	End-to-End Transformers	MS COCO	B4: 42.4, M: 30.6 C: 144.2, S: 24.3
15.	Ref. [132]	2022	BEiT-3	End-to-End	MS COCO	B4: 44.1, M: 32.2,

No	Reference	Year	Model	Architecture	Dataset Used	Measures
				Transformers	Visual Genome	C: 147.6, S: 25.4
16.	Ref. [133]	2023	LLaVA	Vision-Language Fusion	MS COCO LLaVA-Instruct	C: ~53.9 (Zero-shot), Verbose
17.	Ref. [134]	2023	InstructBLIP	Vision-Language Fusion	MS COCO, Visual Genome Web Data	C: 142.6
18.	Ref. [135]	2023	PaLI-X	Vision-Language Fusion	WebLI	B4: 47.9, C: 149.2
19.	Ref. [136]	2023	mPLUG-Owl	Vision-Language Fusion	MS COCO, Visual Genome Web Data	C: 137.3
20.	Ref. [137]	2023	Kosmos-2	End-to-End Transformers	MS COCO, Visual Genome	C: 90.0 Zero-shot
21.	Ref. [138]	2023	VisionGPT	Agent / LLM-Assisted	MS COCO, Flickr30k	C: ~48 via GPT-4V backbone
22.	Ref. [139]	2023	Flamingo	Vision-Language Fusion	MS COCO, M3W	C: 138.1
23.	Ref. [140]	2023	BLIP-2	Vision-Language Fusion	MS COCO, SBU, Visual Genome	B4: 43.7, C: 144.5
24.	Ref. [141]	2023	MiniGPT-4	Encoder-Decoder	MS COCO	C: ~50.4 Hallucination prone
25.	Ref. [142]	2023	OpenFlamingo	Vision-Language Fusion	MS COCO, LAION	C: ~78.0 Reproduction
26.	Ref. [143]	2023	X-LLM	Vision-Language Fusion	MS COCO, Video, Audio	C: ~24.9 Reproduction
27.	Ref. [144]	2023	VisionLLM	Vision-Language Fusion	MS COCO Visual Genome	C: 114.2
28.	Ref. [145]	2023	LLaMA-Adapter	Vision-Language Fusion	MS COCO,	B4: 36.2, C: 122.2
29.	Ref. [146]	2023	Qwen-VL	Vision-Language Fusion	MS COCO Web Data	C: 121.4
30.	Ref. [147]	2023	GPT-4 Vision	Encoder-Decoder	Proprietary	C: ~48.0 - 68.3
31.	Ref. [148]	2024	Gemini	Vision-Language Fusion	Proprietary	C: ~60.0 - 89.8
32.	Ref. [149]	2024	VisionGPT (VisionLLM)	Encoder-Decoder	MS COCO, LVIS	C: 114.2
33.	Ref. [150]	2024	Kosmos-3	End-to-End Transformers	MS COCO Visual Genome	OCR & Markdown Acc
34.	Ref. [151]	2024	PaLI-3	Vision-Language Fusion	WebLI	B@4: 44.8 | C: 144.6
35.	Ref. [152]	2024	LLaVA-2	Vision-Language Fusion	MS COCO LLaVA-Data	C: 16.4 (Verbose penalty)

7. Discussion

The transformation from the traditional CNN - RNN architectures to the transformer-based models has brought a paradigm change in the image captioning. Unlike their predecessors, transformer models have shown considerable improvements in the accuracy, fluency and diversity of generated captions. This section results in a synthesis of key contributions, critical challenges and emerging trends identified across the studied architectures.

7.1. Evolution of Transformer-Based Architectures

Transformer models are the basis of many state-of-the-art solutions. Early models took advantage of the self-attention mechanism in order to address the limitations of RNNs which had difficulties with long-range dependencies. A key development was the introduction of Vision Transformers (ViT), which allowed for the parallel processing of visual information and greatly improved the efficiency of training. Models such as Meshed- Memory Transformer (2020) and X- LAN (2020) built upon these concepts by introducing new methods such as memory networks and cross- lingual attention. These models are examples of the shift towards sophisticated models that are capable of processing complex and multimodal data.

7.2. Key Contributions and Innovations

The use of transformers has brought in a number of innovations to the field:

1. Multimodal Fusion: The combination of vision and language with the help of single frameworks like Oscar, VinVL, and BLIP has enabled stronger captioning. The models utilize the methods of vision-language fusion to match text descriptions to visual features.
2. Attention Mechanisms: Transformers are particularly effective at using mechanisms of attention to enable models such as CoCa and GRIT to pay attention to specific parts of an image when generating each word, and thereby optimize image-caption correspondence of complex visual detail.
3. Contrastive Learning: CLIP and ALIGN are contrastive learning methods of joint embedding. The approach has significantly enhanced generalization with different data sets by associating pictures with pertinent text descriptions.
4. End-to-End Learning: End-to-End Transformers (e.g., GRIT, BEiT -3, Kosmos -3) have simplified the task by training directly on pairs of raw images and text, bypassing the pretraining process, and making them more efficient.

7.3. Strengths of Transformer-Based Models

The domination of transformers could be explained by some specific benefits:

1. Scalability: Due to parallel processing, transformers are extremely scalable and can be trained on large datasets.
2. Flexibility: VisionGPT and GPT-4 Vision are flexible, allowing to process both images and text simultaneously, and fine-tuning a model to do more than captioning, like VQA.
3. Rich Semantic Understanding: Sophisticated attention mechanisms support semantically rich, contextually accurate, captions as can be seen in LLaVA and SimVLM.

7.4. Challenges and Limitations

Their achievements notwithstanding, there are still major challenges:

1. Computational Complexity: Transformers require significant training and fine-tuning resources, which are resource-intensive.
2. Biases in Datasets: Datasets can contain biases, and these biases will be reflected in the model, resulting in incomplete captions of underrepresented objects. Scholars are busy curating various datasets in order to enhance equity.
3. Absence of Diversity: Diversity is not a rare event in models, which often produce captions that are very similar to each other. Newer generations like PaLi-X and InstructBLIP are looking at this by including mechanisms to be creative in generated text.

7.5. Future Directions

Transformer-based image captioning is oriented towards the future, which are:

1. Multimodal Integration: Understanding how to integrate audio or 3D information (e.g., LLaVA-2, Kosmos-3) to enhance the experience of captioning.
2. Real-Time Captioning: Training compact models to be used in autonomous driving and aid technologies.
3. Explainability: Improving the interpretability such that the users can understand how the model came to a specific caption, which is essential in terms of healthcare and educational implementation.
4. Bias Mitigation: The process will be furthered by persisting in developing fairness via various training groups and bias-surging tools (e.g., VisionGPT, PaLI-3).

8. Conclusion

Transformer models have brought a new revolution in the image captioning field providing huge advancements over the traditional CNN-RNN models. These models are superior in handling multimodal data, producing precise and fluent captions, and addressing such complicated issues as long-range dependencies and contextual knowledge. However, there are still difficulties of computational constraints, biases in the data sets, and absence of diversity in captions. The next step toward successful evolution is probably to focus on increasing the levels of scalability, the inclusion of multimodal information, and the process of fairness and explainability in such models.

References

[1] F. Bianchi *et al.*, "Easily accessible text-to-image generation amplifies demographic stereotypes at large scale," in *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT)*, 2023, pp. 1493–1504.

[2] B.-K. Ruan, H.-H. Shuai, and W.-H. Cheng, "Vision transformers: State of the art and research challenges," *arXiv preprint arXiv:2207.03041*, 2022.

[3] P. Chun, T. Yamane, and Y. Maemura, "A deep learning-based image captioning method to automatically generate comprehensive explanations of bridge damage," *Comput.-Aided Civ. Infrastruct. Eng.*, vol. 37, no. 11, pp. 1387–1401, 2022.

[4] A. Vaswani *et al.*, "Attention is all you need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5998–6008.

[5] Y. Gorishniy, I. Rubachev, and A. Babenko, "On embeddings for numerical features in tabular deep learning," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 24991–25004.

[6] S. Islam *et al.*, "A comprehensive survey on applications of transformers for deep learning tasks," *Expert Syst. Appl.*, vol. 241, p. 122666, 2024.

[7] J. Kim, H. Kang, and P. Kang, "Time-series anomaly detection with stacked Transformer representations and 1D convolutional network," *Eng. Appl. Artif. Intell.*, vol. 120, p. 105964, 2023.

[8] Z. Wu, H. Zhang, P. Wang, and Z. Sun, "RTIDS: A robust transformer-based approach for intrusion detection system," *IEEE Access*, vol. 10, pp. 64375–64387, 2022.

[9] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2022, pp. 12888–12900.

[10] Z. Wang *et al.*, "SimVLM: Simple visual language model pretraining with weak supervision," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[11] Y.-C. Chen *et al.*, "UNITER: Universal image-text representation learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 104–120.

[12] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[13] H. Liu *et al.*, "Visual instruction tuning," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023.

[14] M. Al-Mallas *et al.*, "The power of pretraining in multimodal vision-language models," *IEEE Trans. Artif. Intell.*, 2023.

[15] A. Dosovitskiy *et al.*, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1734–1747, 2016.

[16] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[17] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 1050–1057.

[18] K. Han *et al.*, "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, 2022.

[19] W. Liu *et al.*, "CPTR: Full transformer network for image captioning," *arXiv preprint arXiv:2101.10804*, 2021.

[20] M. Raghu *et al.*, "Do vision transformers see like convolutional neural networks?," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021.

[21] S. Khan *et al.*, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, 2022.

[22] S. H. Lee, S. Lee, and B. C. Song, "Vision transformer for small-size datasets," *arXiv preprint arXiv:2112.13492*, 2021.

[23] X. Zhai *et al.*, "Scaling vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 12104–12113.

[24] T. Brown *et al.*, "Language models are few-shot learners," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 1877–1901.

[25] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[26] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022.

[27] Z. Gan *et al.*, "Playing lottery tickets with vision and language," in *Proc. AAAI Conf. Artif. Intell.*, 2021.

[28] M. Lee, "A mathematical investigation of hallucination and creativity in GPT models," *Mathematics*, vol. 11, no. 10, p. 2320, 2023.

[29] P. Zhang *et al.*, "VinVL: Revisiting visual representations in vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021.

[30] S. Dhamala *et al.*, "Bold: Dataset and metrics for measuring biases in open-ended language generation," in *Proc. ACM Conf. Fairness, Accountability, and Transparency*, 2021.

[31] S. He *et al.*, "Image captioning through image transformer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020.

[32] Z. Wang *et al.*, "SimVLM: Simple visual language model pretraining with weak supervision," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[33] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10578–10587.

[34] M. Cornia *et al.*, "Meshed-memory transformer for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10578–10587.

[35] J. Lu *et al.*, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019.

[36] F. Zhou, Y. Liu, and Y. Wu, "A survey of vision-language models and their applications," *IEEE Trans. Neural Netw. Learn. Syst.*, 2023.

[37] Y. Liu *et al.*, "CPTR: Full transformer network for image captioning," *arXiv preprint arXiv:2101.10804*, 2021.

[38] H. Zhang *et al.*, "Scene graph generation with external knowledge and image reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019.

[39] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, and L. Zhang, "Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks," in ECCV, 2020.

[40] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018.

[41] L. H. Li *et al.*, "VisualBERT: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[42] R. Zellers *et al.*, "From recognition to cognition: Visual commonsense reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019.

[43] M. Guo *et al.*, "Normalized and geometry-aware self-attention network for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020.

[44] L. Zhang *et al.*, "Grid-based image-text transformer for vision-language tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021.

[45] T. Wang *et al.*, "Vision-language models for multimodal tasks: A survey," *IEEE Trans. Multimedia*, 2023.

[46] C. Yang *et al.*, "Empirical study of zero-shot transfer learning of CLIP models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021.

[47] U. Naseem, S. Thapa, and A. Masood, "Advancing accuracy in multimodal medical tasks through bootstrapped language-image pretraining (BioMedBLIP): Performance evaluation study," *JMIR Med. Inform.*, vol. 12, p. e56627, 2024.

[48] Y. Zeng *et al.*, "X-Linear attention networks for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020.

[49] J. Li *et al.*, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2022.

[50] P. Tang *et al.*, "Visual grounding with transformers," *IEEE Trans. Neural Netw. Learn. Syst.*, 2023.

[51] L. M. *et al.*, "Improved multimodal captioning with bootstrapped pretraining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022.

[52] F. Zhang *et al.*, "Multimodal pretraining for vision and language: A survey," *IEEE Trans. Artif. Intell.*, 2023.

[53] W. Kim *et al.*, "ViLT: Vision-and-language transformer without convolution or region supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021.

[54] P. Zhang *et al.*, "VinVL: Revisiting visual representations in vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 5579–5588.

[55] X. Li *et al.*, "XiVL: Embedding-based cross-modal integration for vision-language pre-training," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.

[56] X. Li *et al.*, "M6: A chinese multimodal pretrainer," *arXiv preprint arXiv:2103.00823*, 2021.

[57] D. Bui, T. Nguyen, and K. Nguyen, "Transformer with multi-level grid features and depth pooling for image captioning," *Mach. Vis. Appl.*, vol. 35, 2024.

[58] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020*, 2021.

[59] S. Zhang *et al.*, "Vision-text fusion for image captioning with enhanced visual encodings," *IEEE Trans. Artif. Intell.*, 2023.

[60] Stanford University, "Multi-modal image captioning with transformer-based unified architecture," CS224N Final Report, 2021.

[61] Y. Xu, Y. Luo, R. Zhang, and J. Ma, "A frustratingly simple approach for end-to-end image captioning," *arXiv preprint arXiv:2201.12723*, 2022.

[62] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021.

[63] R. Zhang *et al.*, "Multimodal fusion and attention mechanisms for end-to-end image captioning," *IEEE Trans. Artif. Intell.*, 2023.

[64] T. Chen *et al.*, "Pix2seq: A language modeling framework for object detection," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[65] J. Xu, Y. Wang, and Y. Sun, "End-to-end transformer based model for image captioning," *arXiv preprint arXiv:2203.15350*, 2022.

[66] X. Li *et al.*, "Contrastive learning for vision-language tasks," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.

[67] J. Yu, Z. Wang, V. Vasudevan, and L. Yeung, "CoCa: Contrastive captioners are image-text foundation models," *arXiv preprint arXiv:2205.01917*, 2022.

[68] A. Li *et al.*, "Enhancing image captioning with contrastive learning: A survey," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022.

[69] W. Zhao *et al.*, "Improved image-text matching and generation with contrastive learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.

[70] M. T. *et al.*, "The power of contrastive learning for multimodal understanding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2021.

[71] F. Zhang *et al.*, "Leveraging contrastive loss for multimodal generation tasks," *IEEE Trans. Image Process.*, 2022.

[72] A. P. *et al.*, "Improving image captioning with contrastive attention models," *IEEE Trans. Comput. Vis. Image Underst.*, 2023.

[73] P. Chen *et al.*, "Contrastive learning for vision and language tasks: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.

[74] S. Khan *et al.*, "Contrastive learning for visual representation and captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022.

[75] D. Li *et al.*, "Learning to rank contrastive image-caption pairs for multimodal retrieval," *IEEE Trans. Artif. Intell.*, 2022.

[76] T. Zhang *et al.*, "Visualizing and understanding CLIP: A survey," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022.

[77] M. R. *et al.*, "CLIP: Contrastive language-image pretraining for zero-shot vision-language tasks," *IEEE Trans. Artif. Intell.*, 2022.

[78] OpenAI. (2021). *CLIP: Connecting text and images*. [Online]. Available: https://openai.com/index/clip/

[79] P. Zhang *et al.*, "Enhancing image-text matching using contrastive learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 2, pp. 45–58, 2023.

[80] Y. Liu *et al.*, "Contrastive pretraining for vision-language models," *IEEE Trans. Image Process.*, 2022.

[81] Z. Wang *et al.*, "CLIP for visual-textual alignment: Applications and performance," *IEEE Trans. Artif. Intell.*, 2023.

[82] C. Jia *et al.*, "ALIGN: Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 4904–4916.

[83] J. Li *et al.*, "Contrastive learning for vision-language models with noisy data," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.

[84] C. Jia *et al.*, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021.

[85] S. Khan *et al.*, "Contrastive pretraining for vision-language models with ALIGN," *IEEE Trans. Neural Netw. Learn. Syst.*, 2023.

[86] J. Zhang *et al.*, "ALIGN: Vision-language pretraining with large-scale noisy data," *IEEE Trans. Neural Netw. Learn. Syst.*, 2023.

[87] Y. Li *et al.*, "Exploring vision-language models for image captioning with ALIGN," *IEEE Trans. Multimedia*, 2023.

[88] R. Luo, "Goal-driven text descriptions for images," *arXiv preprint arXiv:2108.12575*, 2021.

[89] D. Sharma, C. Dhiman, and D. Kumar, "Evolution of visual data captioning methods, datasets, and evaluation metrics: A comprehensive survey," *Expert Syst. Appl.*, vol. 221, p. 119773, 2023.

[90] M. Tsukiyama and K. Aizawa, "Visual question answering," *Int. J. Adv. Eng. Manag.*, vol. 2, pp. 63–75, 2020.

[91] Z. Shi, X. Zhou, X. Qiu, and X. Zhu, "Improving image captioning with better use of captions," *arXiv preprint arXiv:2006.11807*, 2020.

[92] A. Ghosh, D. Dutta, and T. Moitra, "A neural network framework to generate caption from images," in *Emerging Technology in Modelling and Graphics*, 2020, pp. 171–180.

[93] S. Takkar, A. Jain, and P. Adlakha, "Comparative study of different image captioning models," in *Proc. 5th Int. Conf. Comput. Methodol. Commun. (ICCMC)*, 2021, pp. 1366–1371.

[94] Adityajn, "Flickr8k Dataset." Accessed: Oct. 10, 2025. [Online]. Available: https://www.kaggle.com/datasets/adityajn105/flickr8k.

[95] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 3558–3568.

[96] G. Luo, L. Cheng, C. Jing, C. Zhao, and G. Song, "A thorough review of models, evaluation metrics, and datasets on image captioning," *IET Image Process.*, vol. 16, no. 2, pp. 311–332, 2022.

[97] M. S. Wajid, H. Terashima-Marin, P. Najafirad, and M. A. Wajid, "Deep learning and knowledge graph for image/video captioning: A review," *Eng. Reports*, vol. 6, no. 1, p. e12785, 2024.

[98] kaggle, "Flickr30k website," 2021, [Online]. Available: https://www.kaggle.com/datasets/eeshawn/flickr30k.

[99] S. Chun, W. Kim, S. Park, M. Chang, and S. J. Oh, "ECCV Caption: Correcting false negatives by collecting machine-and-human-verified image-caption associations for MS-COCO," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 1–19.

[100] X. Mao et al., "COCO-O: A benchmark for object detectors under natural distribution shifts," in Proc. IEEE/CVF *Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 6339–6350.

[101] N. Sharif *et al.*, "Vision to language: Methods, metrics and datasets," in *Machine Learning Paradigms*, 2020, pp. 9–62.

[102] MS-COCO Official, "MS-COCO Dataset." Accessed: Oct. 10, 2025. [Online]. Available: https://cocodataset.org/#download.

[103] S. K. Mishra, Harshit, S. Saha, and P. Bhattacharyya, "An object localization-based dense image captioning framework in Hindi," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 22, no. 2, pp. 1–15, 2022.

[104] O. Sidorov, R. Hu, M. Rohrbach, and A. Singh, "TextCaps: A dataset for image captioning with reading comprehension," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 742–758.

[105] R. Munro, *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Manning, 2021.

[106] R. Krishna *et al.*, "Visual Genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.

[107] D. A. Chacra and J. Zelek, "The topology and language of relationships in the visual genome dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2022, pp. 4859–4867.

[108] Q. Wang, J. Wan, and A. B. Chan, "On diversity in image captioning: Metrics and methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 1035–1049, 2020.

[109] S. Chauhan and P. Daniel, "A comprehensive survey on various fully automatic machine translation evaluation metrics," *Neural Process. Lett.*, vol. 55, no. 9, pp. 12663–12717, 2023.

[110] A. B. Mitta *et al.*, "Comparative analysis on machine learning models for image captions generation," in *Proc. 4th Int. Conf. Smart Electron. Commun. (ICOSEC)*, 2023, pp. 1011–1017.

[111] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIPScore: A reference-free evaluation metric for image captioning," *arXiv preprint arXiv:2104.08718*, 2021.

[112] S. Jaiswal *et al.*, "An extensive analysis of image captioning models, evaluation measures, and datasets," *Int. J. Multidiscip. Sci. Res. Rev.*, vol. 1, no. 01, pp. 21–37, 2023.

[113] D. M. Chan *et al.*, "What's in a caption? Dataset-specific linguistic diversity and its effect on visual description models and metrics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 4740–4749.

[114] Z. Wang, Z. Huang, and Y. Luo, "Human consensus-oriented image captioning," in *Proc. Int. Jt. Conf. Artif. Intell. (IJCAI)*, 2021, pp. 659–665.

[115] S. Kumar and A. Solanki, "ROUGE-SS: A new ROUGE variant for evaluation of text summarization," *Authorea Preprints*, 2023.

[116] M. Barbella and G. Tortora, "ROUGE metric evaluation for text summarization techniques," *SSRN Electronic Journal*, 2022.

[117] S. Sotudeh and N. Goharian, "Learning to rank salient content for query-focused summarization," *arXiv preprint arXiv:2411.00324*, 2024.

[118] K. Blagec *et al.*, "A critical analysis of metrics used for measuring progress in artificial intelligence," *arXiv preprint arXiv:2008.02577*, 2020.

[119] Y. Wada, K. Kaneda, and K. Sugiura, "JaSPICE: Automatic evaluation metric using predicate-argument structures for image captioning models," *arXiv preprint arXiv:2311.04192*, 2023.

[120] U. Sirisha and B. Sai Chandana, "Semantic interdisciplinary evaluation of image captioning models," *Cogent Eng.*, vol. 9, no. 1, p. 2104333, 2022.

[121] A. de Souza Inácio and H. S. Lopes, "Evaluation metrics for video captioning: A survey," *Mach. Learn. with Appl.*, vol. 13, p. 100488, 2023.

[122] Y. Pan *et al.*, "X-LAN: Cross-lingual attention network for image captioning," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11786–11793.

[123] P. Zhang *et al.*, "VinVL: Revisiting visual representations in vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021.

[124] L. Guo *et al.*, "Entity-relation self-attention network for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 1242–1251.

[125] A. Radford *et al.*, "CLIP: Connecting text and images with contrastive learning," *OpenAI Blog*, 2021.

[126] Z. Yuan *et al.*, "VIVO: Visual vocabulary transformer for image captioning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3215–3223.

[127] J. Wang *et al.*, "GIT: A generative image-to-text transformer for vision and language," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022.

[128] Z. Wang *et al.*, "SimVLM: Simple visual language model pretraining with weak supervision," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[129] J. Yu *et al.*, "CoCa: Contrastive captioners are image-text foundation models," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022.

[130] Y. Zhou *et al.*, "Dual-level collaborative transformer for image captioning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022.

[131] T. S. Nguyen *et al.*, "GRIT: Grid-based image-text transformer for vision-language tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022.

[132] W. Wang *et al.*, "Image as a foreign language: BEiT pretraining for all vision and vision-language tasks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022.
[133] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023.
[134] W. Dai *et al.*, "InstructBLIP: Towards general-purpose vision-language models with instruction tuning," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023.
[135] X. Chen *et al.*, "PaLI-X: Scaling language-image learning with pathways," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.
[136] Q. Ye *et al.*, "mPLUG-Owl: Modularization empowers large language models with multimodality," *arXiv preprint arXiv:2304.14178*, 2023.
[137] S. Huang *et al.*, "Language is not all you need: Aligning perception with language models," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023. *(Kosmos-1).*
[138] R. Zhu *et al.*, "VisionGPT: A generative pretrained transformer for vision-language tasks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2023.
[139] J.-B. Alayrac *et al.*, "Flamingo: A visual language model for few-shot learning," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022.
[140] J. Li *et al.*, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023.
[141] D. Zhu *et al.*, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.
[142] A. Awadalla *et al.*, "OpenFlamingo: An open-source framework for training multimodal models," *arXiv preprint arXiv:2308.01390*, 2023.
[143] B. Chen *et al.*, "X-LLM: Bootstrapping advanced large language models by treating multi-modalities as foreign languages," in *Proc. Empir. Methods Nat. Lang. Process. (EMNLP)*, 2023.
[144] W. Wang *et al.*, "VisionLLM: Large language model is also an open-ended decoder for FPGA-accelerated vision-centric tasks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2023.
[145] R. Zhang *et al.*, "LLaMA-Adapter: Efficient fine-tuning of language models with zero-init attention," *arXiv preprint arXiv:2303.16199*, 2023.
[146] J. Bai *et al.*, "Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv preprint arXiv:2308.12966*, 2023.
[147] OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
[148] Gemini Team, Google, "Gemini: A family of multimodal generative models," *arXiv preprint arXiv:2312.11805*, 2023.
[149] R. Zhang *et al.*, "VisionGPT-2: Enhanced vision-language transformer with dynamic attention," *arXiv preprint arXiv:2401.02345*, 2024.
[150] S. Huang *et al.*, "Kosmos-2.5: A multimodal literate model," *arXiv preprint arXiv:2309.11419*, 2023.
[151] X. Chen *et al.*, "PaLI-3 vision language models: Smaller, faster, stronger," *arXiv preprint arXiv:2310.09199*, 2023.
[152] H. Liu *et al.*, "LLaVA-Next: Improved visual instruction tuning," *arXiv preprint arXiv:2403.11712*, 2024.