Evaluation of VLMs using SOTA evaluation metrics

Ujwal Pratap Krishna, Aryaman Bahukhandi

1 Introduction

With recent breakthroughs in large-scale generative AI models, the world is witnessing a profound surge in the potential of large language models, extensive vision language models, and other generative AI counterparts. These models exhibit a striking resemblance to human speech and cognitive abilities, propelling the industry forward due to their diverse applications.

However, this rapid advancement brings along a distinct set of challenges. Given their training on vast corpora of data, certain tasks exhibit a predominant bias towards the knowledge base learned during training, rather than task-specific information. Furthermore, these models are susceptible to hallucinations, producing outputs that may appear satisfactory to humans but are factually inconsistent. Consequently, there is an urgent need for a thorough evaluation of outputs from these large models to establish a more reliable correlation with human understanding.

Thus, our project seeks to assess a selection of cutting-edge vision language models widely employed across various vision-related tasks such as image captioning, image classification, and style transfer on the state-of-the-art metrics used to evaluate VLMs. Within this endeavor, we have evaluated Salesforce's InstructBLIP-FLAN-T5-XL, OpenAI's CLIP, Salesforce's InstructBLIP-Vicuna, HuggingFace's VIT-GPT2 and an encoder-decoder-based model custom-trained specifically for image captioning. All these vision language models were applied to perform zero-shot inference on a subset of Microsoft's MSCOCO dataset. Additionally, the custom encoder-decoder model was trained on the Flickr8k dataset.

The metrics utilized for the evaluation include Bleu_1, Bleu_2, Bleu_3, Bleu_4, METEOR, ROGUE_L, CIDEr, CIDEr-R, and SPICE.

2 Literature Review

2.1 InstructBlip

Recent advancements have shown encouraging outcomes by integrating vision and language models based on the foundation of pre-trained BLIP models. Our methodology involves multi-modal networks, specifically the Salesforce's Blip series, which merges a visual encoder with a language model[8]. This is akin to the approach adopted by Wenliang Dai et al.[5]. In their research, they carried out a systematic and comprehensive study on vision-language instruction tuning based on the pre-trained BLIP-2 models. They collected 26 publicly available datasets, encompassing a wide range of tasks and capabilities, and converted them into an instruction-tuning format. They also introduced an instruction-aware Query Transformer, which extracts informative features tailored to the given instruction. Their models achieved state-of-the-art zero-shot performance across all 13 held-out datasets. In our study, we limit our focus to the encoder-decoder-based InstructBlip-Flan-T5 and decoder-only-based InstructBlip-Vicuna VLMs.

2.2 CLIP

Clip models are neural network architectures focused on contrastive language-image pre-training. They were introduced in a 2021 paper "Learning Transferable Visual Models From Natural Language Supervision" by Radford et al[11].

The key innovation with clip models is that they are trained on a large dataset of text-image pairs to learn multimodal representations, without the need for explicit object labels. Specifically, the model tries to

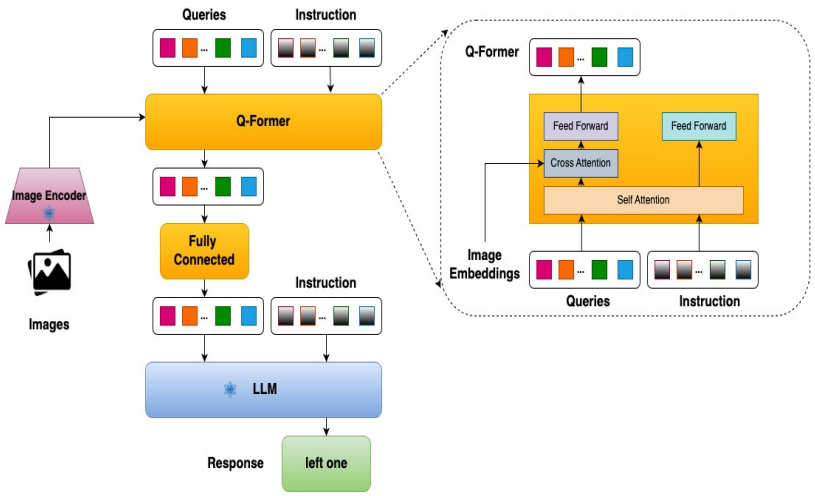

Figure 1: Model architecture of InstructBLIP as depicted in [*Wenliang Dai et al., 2023, "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning"*][5]. "The Q-Former extracts instruction-aware visual features from the output embeddings of the frozen image encoder, and feeds the visual features as soft prompt input to the frozen LLM.[5]

predict if an image-text pair matches or doesn't match. The contrastive approach allows the model to learn powerful associations between visual concepts and language.

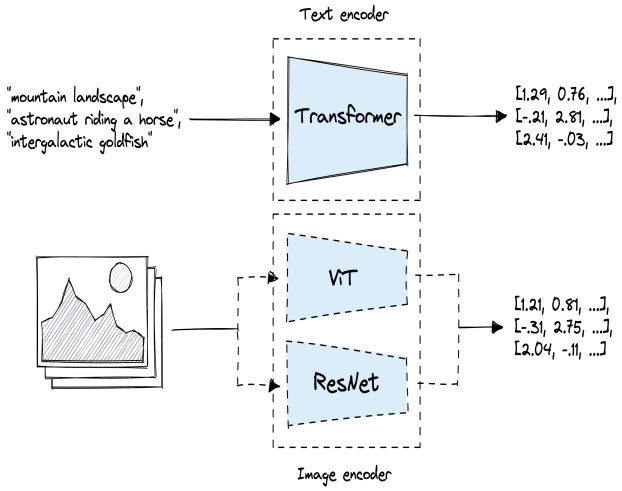

Figure 2: Model architecture of CLIP.[10]

2.3 Vit-GPT2 Image Captioning

Hugging Face's ViT-GPT2 represents a state-of-the-art approach in the domain of image captioning by amalgamating the strengths of Vision Transformer (ViT) and GPT-2. ViT, initially designed for vision tasks, extracts hierarchical visual features from images by breaking them down into patches, enabling efficient processing and understanding of visual information. These extracted features are then fed into GPT-2, a powerful language model well known for its text generation capabilities. By integrating ViT's ability to comprehend images and GPT-2's prowess in generating coherent and contextually relevant text, ViT-GPT2 offers a robust solution for image captioning.

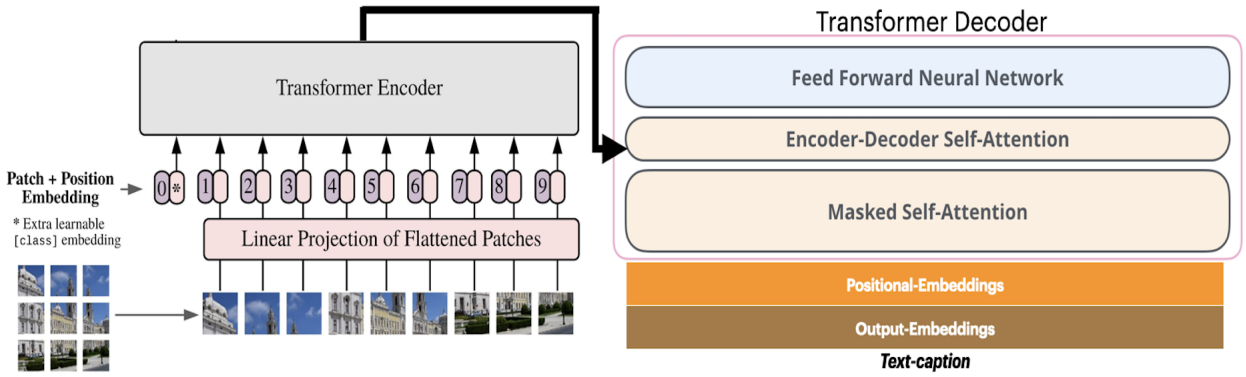

Figure 3: Model architecture of ViT-GPT2.[7]

3 Methodology

3.1 Dataset

- **COCO Captions [3]:** The COCO Captions dataset is a widely-used collection in the field of computer vision and natural language processing. It consists of over 600,000 images annotated with 5 descriptive captions per image. The dataset covers a diverse range of scenes, objects, and activities, providing a rich resource for training and evaluating image captioning algorithms. Researchers often leverage this dataset to develop models that can generate accurate and contextually relevant captions for images. Its scale, diversity, and quality make it an ideal benchmark dataset in the domain for evaluation, fostering advancements in image understanding and caption generation tasks.
- **Flickr8k [6]:** The Flickr8k dataset is a well-known benchmark in the field of image captioning, containing 8,000 images from Flickr, each paired with five human-generated captions. This dataset is smaller in scale compared to some others but remains valuable due to its high-quality annotations and diversity in image content. It's often utilized for developing and evaluating image captioning models, allowing researchers to train algorithms to generate descriptive and contextually relevant captions for images, making it highly suitable for efficient image-caption task training of our custom baseline model despite being constrained by limited computing resources.

3.2 ResNET50 encoder-Attention LSTM decoder (Baseline Model)

The custom implemented Encoder-Decoder architecture for image captioning harnesses a combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to generate descriptive captions for images. The architecture comprises an EncoderCNN, leveraging a pre-trained ResNet-50 to extract features from images. The ResNet-50's last 2 layers corresponding to it's classifcation layers were removed to obtain the image features. This is input to an attention mechanism-equipped DecoderRNN responsible for generating captions. The attention mechanism is based on the Bahdanau attention algorithm [2] and allows the model to focus on different image regions while decoding, ensuring contextually relevant captions, LSTM is used as the choice of RNN to generate captions.

During training, the model uses the Cross-Entropy Loss function, ignoring padding tokens, and optimizes its parameters using the Adam optimizer with a learning rate of 3e-3. The training loop spanned 100 epochs on 6000 images of Flickr8k training data, where batches of size 256 are processed, incorporating image preprocessing techniques such as resizing, cropping, and normalization. The model's hyperparameters include an embedding size of 300, an attention dimension of 256, an encoder dimension of 2048, and a decoder dimension of 512. The model was trained on NVIDIA A100 GPU(115GB CPU RAM, 40GB GPU RAM, 1500 GB Disk space). The architecture's fusion of visual understanding via CNNs and sequential generation of text using attention-based RNNs enables it to produce contextually coherent captions for images, enhancing the understanding of image content through descriptive text making it a suitable baseline to

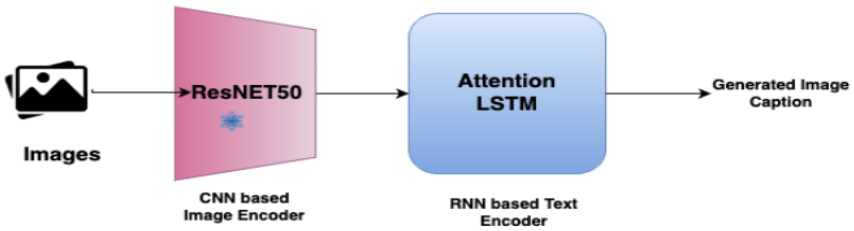

Figure 4: Model architecture of ResNET50 Image Encoder- Attention incorporated LSTM decoder.

evaluate our models under test.

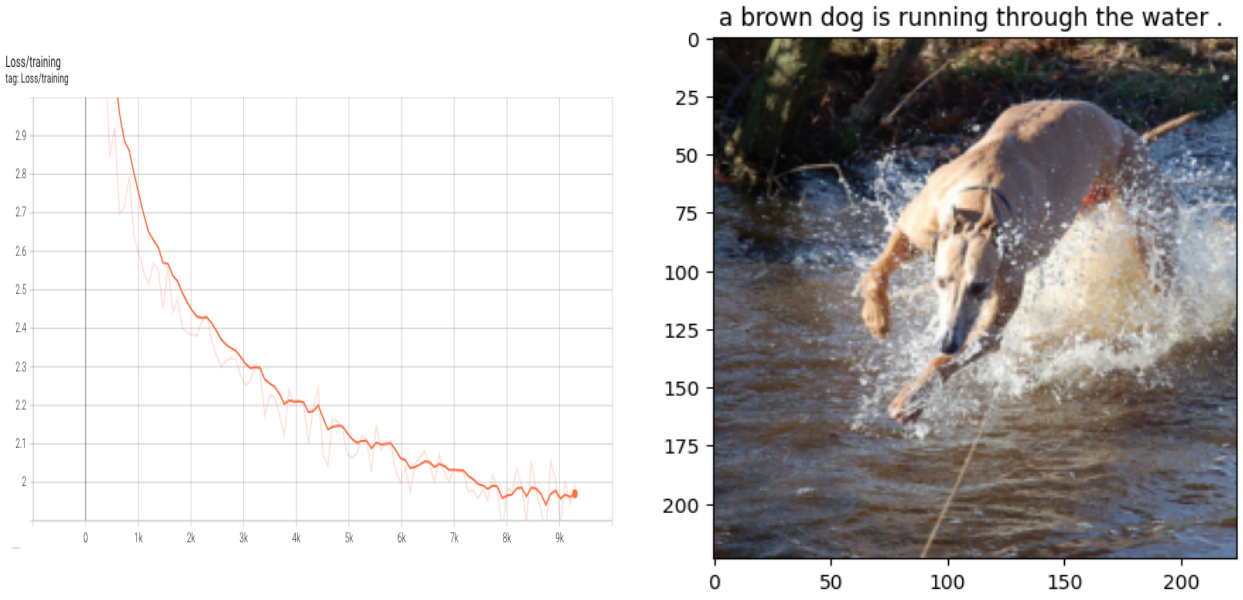

Figure 5: Left: Plot of Training Loss(y-axis) vs Total Steps(x-axis) for the custom CNN Encoder-RNN Decoder model; Right: Caption Generated during inference for a test image

Figure 5's right image displays the caption that was generated during inference on an unseen image. The picture displays a dog running across a river and the caption generated is fairly good in terms of capturing the overall information displayed. Figure 5's left image shows the decrease in loss as the number of training steps increases, the evaluations of this model along with the models under test will be presented using the evaluation metrics discussed in the Section 4.

3.3 Evaluated Vision Language Models (Inference Only)

In our research, we have performed a holistic evaluation of four Vision Language Models(VLMs) or equivalent models:

3.3.1 Salesforce's InstructBLIP-Vicuna:

This model [5] pairs a Vision Language Model (VLM) with a decoder-only Language Model (LLM) called Vicuna, developed by LMSYS and fine-tuned from Meta's LLAMA [13] family. Vicuna relies solely on attention to past tokens for predicting subsequent text tokens, excelling in generation-based tasks but underperforming in classification-based tasks due to its autoregressive nature.

3.3.2 Salesforce's InstructBLIP-Flan-T5-XL:

Another VLM from Salesforce [5], connecting to an encoder-decoder Language Model called Flan-T5 [4], developed by Google Research. Flan-T5 incorporates both bi-directional encoders and autoregressive decoders, showcasing strong performance in both generation-based and classification-based tasks. However, due to increased parameters, it tends to be slower during inference compared to other LLM variants.

3.3.3 OpenAI's CLIP:

This encoder-only VLM comprises separate text and image encoders, processing textual and visual inputs to create vector representations. CLIP [11] excels in understanding and relating visual and textual information, specifically designed for classification-based tasks but might lag in performance for generative tasks.

3.3.4 HuggingFace's ViT-GPT2:

Connecting a Vision Language Model with a decoder-only LLM, ViT-GPT2 [7] [9] combines OpenAI's GPT2 model with a Vision Transformer (ViT) developed by Google Research. Despite being smaller and lacking specialized components like Q-Formers or highly optimized Image Encoders, it serves as an integrated VLM suitable for various tasks, leveraging ViT for visual understanding and GPT2 for text generation.

All the above models were used for inference on the ms coco subset for caption generation using NVIDIA A100 GPU(115GB CPU RAM, 40GB GPU RAM, 1500 GB Disk space). and were set at the hyperparameters (where applicable): $temperature = 0, max\_length = 128, top\_p = 0.9, num\_beams = 5, length\_penalty = 1.0, repitition\_penalty = 1.5$. The above models cover a broad range of categories to test ranging from small to large language models and from encoder-only to encoder-decoder to decoder-only backed LLMs and provide a good test model suite to be evaluated with our custom baseline model.

4 Evaluation Metrics

4.1 BLEU

BLEU, a widely used metric for assessing machine-translated output quality, calculates precision by comparing n-grams in the machine translation to those in reference translations. The score, ranging from 0 to 1, is derived as the geometric mean of modified n-gram precisions, adjusted for brevity penalties in shorter outputs compared to reference translations.

4.2 METEOR

It evaluates machine translation by aligning unigrams in the machine output with human-made references. Unique features like stemming and synonym matching address issues in BLEU, ensuring a strong correlation with human judgment at the sentence or segment level.

4.3 ROGUE_L

ROUGE_L evaluates text summaries by comparing them to human-made reference summaries, emphasizing longer common subsequences as indicators of higher quality. The resulting score, ranging from 0 to 1, represents the average F1-score across all reference summaries for a given summary, with higher scores denoting superior summaries.

4.4 CIDEr and CIDEr_R

The CIDEr metric [14] evaluates the quality of image captions by contrasting them with reference captions authored by humans. Its premise lies in the notion that good captions should not merely replicate the words and structure of reference captions but also convey equivalent meaning and content. It makes use of an N-gram similarity, TF-IDF weighting to calculate the frequency of word's occurrences. The similarity with

gold referrences is found using cosine similarity. The final score is a geometric mean of the cosine similarities for different n-gram lengths. CIDEr_R is a variant of the CIDEr metric that emphasizes the importance of rare words in the evaluation of image captions.

4.5 SPICE

This metric [1] is an improvement over the other metrics in certain aspects. Most of the evaluation metrics use some kind of string matching or frequency counting algorithm that focuses majorly on the syntactic aspect of caption generation but lack in representing the semantic meaning of the reference captions. Spice retains semantic connections by using a semantic parser to create a scene graph of the generated captions. The scene graph is a combination of the objects, relations and attributes involved in the caption. The scene graph of the generated caption is checked against the scene graph of the reference captions to check the fraction of objects, relations and attributes retained in the output. The SPICE metric computes an F-score, which is a harmonic mean of precision and recall, over the logical tuples that represent the semantic propositions in the scene graphs.

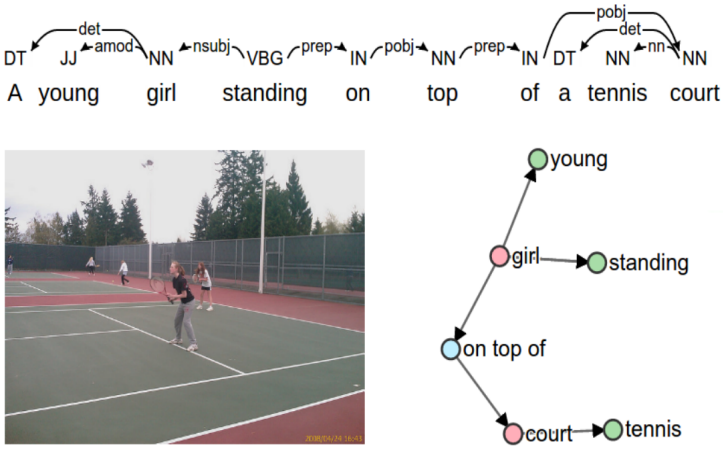

Figure 6: Example of a scene graph with the input image.

5 Preliminary Results

Model	Caption
GroundTruth	*a adult sheep stands by a tree as some baby sheep look on*
Baseline Model	group of sheep standing in front of a fence
InstructBlip-FLAN-T5	a sheep and two lambs resting in the shade of a tree
InstructBlip-Vicuna	a group of sheep in a grassy field near a tree
CLIP	A group of sheep laying on top of a tree.
ViT-GPT2	a sheep and a lamb are standing in the grass sheep on top

Figure 7: Comparison of generated captions from all the models with ground truth and baseline model

After all the VLMs were used to infer for the 250 images, the captions were tested against the gold references [12] and the following scores were calculated:

From this table, the following notable observations were made:

Model	Bleu_1	Bleu_2	Bleu_3	Bleu_4	ROGUE_L	METEOR	CIDEr	CIDEr_R	SPICE
RESNET50 - attention LSTM (BASELINE)	0.579	0.404	0.279	0.191	0.396	0.195	0.505	0.523	0.133
InstructBLIP-FLAN-T5-xl	0.821	0.658	0.512	0.387	0.588	0.297	0.977	0.993	0.233
InstructBLIP-VICUNA	0.747	0.605	0.478	0.365	0.57	0.305	0.952	0.969	0.241
CLIP	0.771	0.612	0.462	0.341	0.571	0.278	0.868	0.873	0.199
VIT-GPT2	0.686	0.506	0.367	0.267	0.507	0.244	0.732	0.765	0.165

Figure 8: Metric table for all models evaluated

- InstructBLIP-FLAN-T5-xl outperforms all of the models in most of the metrics except SPICE and METEOR, where InstructBLIP-Vicuna outperforms. This can be attributed to the fact that these models are either encoder-decoder or decoder-only architecture which usually achieve better human correlation for generative tasks like image captioning as compared to encoder-only models.
- InstructBLIP based models also outperform other models because it uses a Query Transformer. The Q-Former retrieves instruction-aware visual features from the output embeddings of the static image encoder. These visual features are then utilized as soft prompt input for the unaltered LLM. In our context, these instruction-aware visual features are applied to the task of caption generation.
- Our baseline model despite being smaller and lacking Q-former or optimized image encoders performs very similar to VIT-GPT2 model despite being trained on a much smaller dataset compared to GPT2, which is makes it a noteworthy baseline.
- It's also observed that the CIDEr and CIDEr_R scores 4.4 for all VLM's seem somewhat unrealistic with a very high human correlation, This is attributed to the fact that given the time and resources inferencing on a very large and representative test suite was not feasible so the inference is done for randomly sampled 250 images from the MSCOCO dataset.
- It's also observed that the SPICE score 4.5 of all the models is low as compared to the other metrics, This can be attributed to the fact that SPICE doesn't consider syntactic structure of the captions, instead it only looks for the objects, relations and attributes in the generated caption's scene graph and compares it to the gold reference's scene graph.

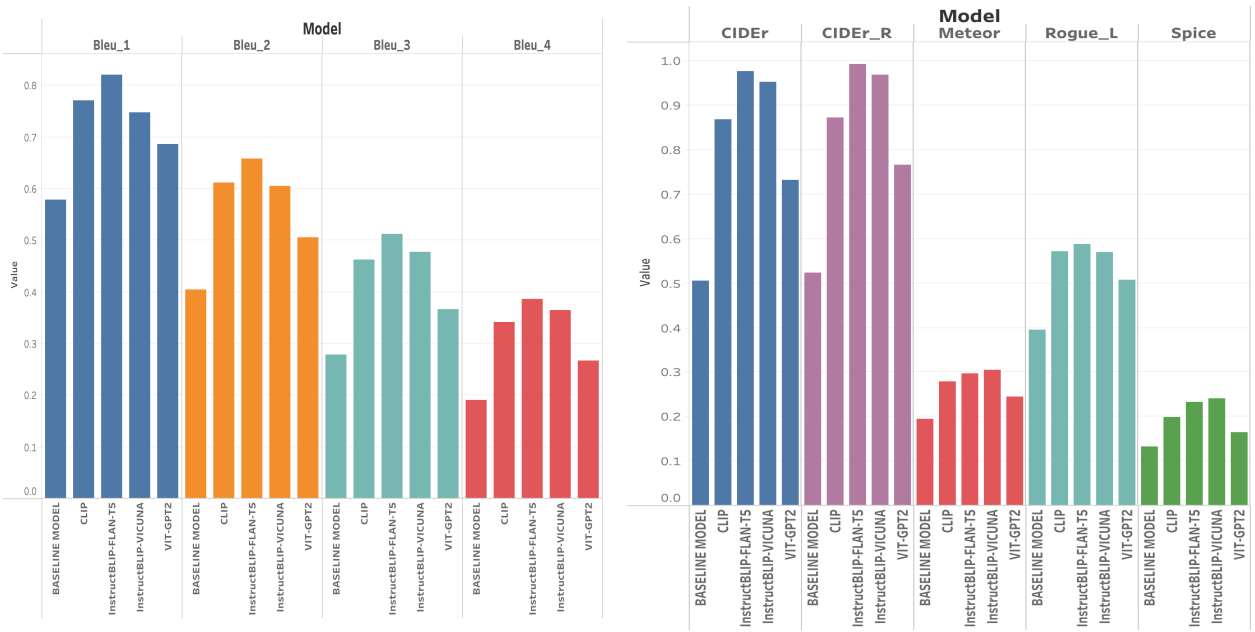

Figure 9: Visualizing different models scores- *Left:* BLEU_N Gram scores; *Right:* CIDEr, CIDEr_R, Meteor, Rougue_L and Spice scores

6 Conclusion and future work

We were able to successfully evaluate some of the most popular and widely used VLMs against our custom trained CNN-encoder RNN-decoder model for a set of SOTA evaluation metrics widely used in the domain of Generative AI. The valuable insights drawn from this evalaution can be used to make better decisions when choosing VLMs in the future by the user.
For the future work, we can extend this evaluation to a more hollistic test suite and make it more generizable. We can use more sampling strategies to create a more representative and larger test suite.
We can also extend the same evaluation methods to other vision tasks like image classification, style-transfer, object detection, etc.

7 Acknowledgements

The authors would like to express their sincere gratitude to Professor Hamed Pirsiavash and TA Melissa Liu for their in- valuable guidance and support towards the completion of this research project. As the professor and TA for graduate-level course 271 Machine Learning and Discovery, their extensive expertise in Machine Learning and Computer Vision has been instrumental in shaping this work.

References

[1] Peter Anderson et al. "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*. Springer. 2016, pp. 382–398.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2016. arXiv: 1409.0473 [cs.CL].

[3] Xinlei Chen et al. *Microsoft COCO Captions: Data Collection and Evaluation Server*. 2015. arXiv: 1504.00325 [cs.CV].

[4] Hyung Won Chung et al. *Scaling Instruction-Finetuned Language Models*. 2022. arXiv: 2210.11416 [cs.LG].

[5] Wenliang Dai et al. *InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning*. 2023. arXiv: 2305.06500 [cs.CV].

[6] Micah Hodosh, Peter Young, and Julia Hockenmaier. "Framing image description as a ranking task: Data, models and evaluation metrics". In: *Journal of Artificial Intelligence Research* 47 (2013), pp. 853–899.

[7] Ankur Kumar. *The Illustrated Image Captioning using transformers*. URL: https://ankur3107.github.io/blogs/the-illustrated-image-captioning-using-transformers/.

[8] Junnan Li, Hoi Steven, and Rose Donald. *Blip: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation*. 5 Nov. 2022. URL: blog.salesforceairesearch.com/blip-bootstrapping-language-image-pretraining/.

[9] nlpconnect/vit-gpt2-image-captioning. *nlpconnect/vit-gpt2-image-captioning*. URL: https://huggingface.co/nlpconnect/vit-gpt2-image-captioning.

[10] pinecone. *Multi-modal ML with OpenAI's CLIP*. URL: https://www.pinecone.io/learn/series/image-search/clip/.

[11] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV].

[12] ruotianluo. *COCO-Caption*. URL: https://github.com/ruotianluo/coco-caption/tree/master.

[13] Hugo Touvron et al. *LLaMA: Open and Efficient Foundation Language Models*. 2023. arXiv: 2302.13971 [cs.CL].

[14] 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*. 2015, pp. 4566–4575.