Transforming Image Captioning: Integrating SwinV2, CSwin, and DeiT Architectures into the Pure Transformer (PureT) Model

Austin Lamb
University of Southern California
Los Angeles, CA, USA
arlamb@usc.edu

Hassan Shah
University of Southern California
Los Angeles, CA, USA
hhshah@usc.edu

***Abstract*—This paper presents an in-depth exploration of transformer-based models for image captioning, a burgeoning field at the intersection of Natural Language Processing and Computer Vision. Our project replicates and enhances the Pure Transformer (PureT) model for image captioning, initially introduced by Wang et al. in 2022, using the 'Karpathy' offline test split of the Microsoft COCO dataset. We implemented a codebase that replicates the PureT model, with the goal of matching or surpassing its reported 138.2% CIDEr score. Our methodology involved experimenting with various transformer backbones, such as SwinV2, CSwin, and DeiT, while also refining the encoder and exploring novel approaches like SmallCap and datastore retrieval. Our findings reveal that not only were we able to replicate the original PureT results, but specific enhancements, particularly with the SwinV2 and CSwin backbones, led to improved efficiency and performance. The results of this study not only confirm the effectiveness of the original PureT model but also demonstrate the enhanced capabilities and efficiency of the SwinV2 and CSwin backbones, offering a contribution to the field of image captioning and providing useful insight for future advancements in transformer-based modeling.**

***Index Terms*—Image Captioning, Natural Language Processing, Computer Vision, Object Detection, Vision Transformers, Swin Transformers, Deep Learning, Multi-Modal Models**

I. Introduction

The advent of deep learning has revolutionized numerous fields in computing, notably Natural Language Processing (NLP) and Computer Vision (CV). One of the most intriguing applications at this intersection is image captioning, which aims to generate descriptive text for a given image. This task demands an understanding of the content within an image and the generation of coherent and contextually relevant language descriptions. The challenge lies in effectively bridging the gap between visual perception and linguistic expression.

Our research focuses on advancing the capabilities of end-to-end transformer-based models in image captioning. The motivation stems from the transformative impact of transformer models in both NLP and CV, which have shown remarkable success in understanding and generating human language and processing complex visual inputs. However, applying these models to image captioning, especially in a manner that synergistically combines visual and linguistic elements, remains an area ripe for exploration and enhancement.

Building on the foundational work by Wang et al. [1], our research centers on the Pure Transformer (PureT) model for image captioning. The PureT model represents a significant advancement in the field, employing an end-to-end transformer architecture that effectively integrates visual inputs with language generation capabilities. Our goal is to replicate the success of the PureT model, using the Microsoft COCO dataset, and then to extend its capabilities through various enhancements.

These enhancements include experimenting with different transformer backbones such as SwinV2 [2], CSwin [3], and DeiT [4] and refining the encoder. By exploring these avenues, we aim to push the boundaries of what transformer models can achieve in image captioning, striving for improvements in both accuracy and efficiency. Our work also involves exploring novel approaches like SmallCap [5] and datastore retrieval, further diversifying the methodologies applied in this field.

The significance of our research lies in enhancing an existing model and contributing to a deeper understanding of image captioning, particularly in the context of multi-modal applications. By advancing transformer models in this domain, our work helps to deepen the understanding of how visual and linguistic elements can be more effectively integrated. This research is another stepping stone towards more sophisticated and nuanced multi-modal AI systems, emphasizing the practical implications in CV and NLP.

Our main contributions are summarized as follows:

- We recreate the results of the PureT paper by Wang et al., validating the results of said paper.
- We integrate various transformer backbones, such as SwinV2, which notably enhances performance, CSwin, contributing to improved efficiency, and DeiT, into the PureT model, leading to improvements in efficiency and performance in image captioning tasks.
- Through extensive experimentation, we demonstrate the ability to match and improve the CIDEr metric for PureT. We achieve a CIDEr score equal to the original model's

[1]Code: https://github.com/hhsusc/Transformers-Image-Captioning

138.2% in fewer epochs, with reduced training time and fewer resources.
- We propose a potential modification in the encoder of the PureT model and explore novel approaches like SmallCap and datastore retrieval. These modifications are aimed at improving inter-modal interactions, efficiency, and costs.

II. Related Work

The development of image captioning systems, a critical intersection of Computer Vision (CV) and Natural Language Processing (NLP), has seen significant advancements with the emergence of transformer models. Early approaches in image captioning predominantly relied on template-based methods or convolutional neural networks (CNNs) coupled with recurrent neural networks (RNNs). The seminal work by Vaswani et al. in "Attention Is All You Need" [6] introduced the transformer architecture in NLP, revolutionizing approaches to sequence-to-sequence tasks. This breakthrough was followed by notable adaptations in CV, as evidenced by Dosovitskiy et al. in "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" [7], which further expanded the application of transformers to visual data through the advent of the Vision Transformer (ViT). The Pure Transformer (PureT) model by Wang et al. [1] represents a significant leap in progress in the field, showcasing a purely transformer-based approach for image captioning capable of effectively capturing complex interdependencies between visual and textual elements without the need for pre-training stages typical of object detectors.

Our research is grounded in these foundational studies, and we seek to enhance the PureT model by exploring various transformer backbones and architectural refinements. The Microsoft COCO dataset, a widely acknowledged benchmark in image captioning research, is the basis for our experiments. Its extensive collection of labeled images and unique captions per image offers a comprehensive platform for assessing our model's performance. By situating our work within the context of these pivotal studies, we aim to contribute to the ongoing discourse on the efficacy and performance of transformer models in image captioning.

Furthermore, we draw inspiration from advancements in transformer architectures, such as the Swin Transformer V2 from Liu et al. [2] and the CSwin Transformer from Dong et al. [3], which introduce novel approaches to handling visual inputs. Our exploration includes not only these advanced backbones but also innovative methodologies like SmallCap and datastore retrieval, as proposed by Ramos et al. [5], to enhance the model's captioning capabilities. Through this process, our study builds upon the existing literature.

III. Approach

A. Replicating PureT Model

Our approach to enhancing the PureT model for image captioning began with a thorough replication of the original model as presented by Wang et al. [1]. This replication was crucial to establish a baseline for our enhancements and to ensure an understanding of the model's workings. We utilized the 'Karpathy' [8] offline test split of the Microsoft COCO dataset, widely recognized for its diversity and complexity in image captioning tasks [1].

B. Training Strategy and Phases

The first phase of our methodology involved a two-stage training process for the model. Initially, we employed cross-entropy (XE) loss for 20 epochs. The XE loss is defined as:

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log(p(y_t^*|y_{1:t-1}^*)) \quad (1)$$

where $y_{1:T}^*$ represents the target ground truth sequence, and θ denotes the parameters of our model. This objective function is a standard approach in training deep learning models. It is particularly effective in early stages, where the model learns to approximate the probability distribution of the training data.

Subsequently, we transitioned to Self-Critical Sequence Training (SCST) for an additional 30 epochs. SCST, as introduced by Rennie et al. [9], uses a different approach, focusing on optimizing a non-differentiable metric (in our case, the CIDEr score). The SCST method is expressed as follows:

$$LR(\theta) = -E_{y_{1:T} \sim p_\theta}[r(y_{1:T})] \quad (2)$$

where r(·) is the CIDEr score. This objective function involves training the model to improve its performance based on the scores from its best predictions. This training regime shifts the focus from accurately predicting the next word to generating higher-quality captions and fine-tuning the model's performance.

This dual-phase training strategy was instrumental in balancing the initial learning phase, grounded in syntactic accuracy (XE), with the subsequent fine-tuning phase aimed at semantic optimization (SCST). By employing these complementary training methods, we ensured a comprehensive development of the model's capabilities, maintaining the key aspects of the PureT model's transformer-based architecture, which effectively integrates visual and linguistic processing.

C. Transformer Architecture Enhancements and Modifications

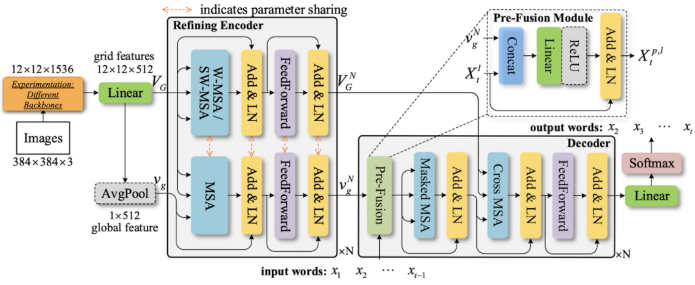

Fig. 1. Enhanced PureT Model Architecture: Experimentation with Various Transformer Backbones. This figure illustrates the incorporation of different transformer backbones, like SwinV2, CSwin, and DeiT, into the PureT framework, highlighting the key areas of model experimentation and improvement.

Our primary enhancements come from experimenting with different transformer backbones. We explored the integration

of advanced architectures like SwinV2 [2], CSwin [3], and DeiT [4] into the PureT model. Each of these backbones brought unique strengths to the model, particularly in how they processed visual inputs and interacted with the textual components of the model. For instance, SwinV2 offered improvements in handling spatial hierarchies, while CSwin introduced cross-shaped window attention mechanisms for efficient visual processing.

Additionally, we experimented with modifications in the encoder size and architecture. This process included adjusting the number of transformer heads and fine-tuning the balance between grid and global feature dimensions. These modifications were aimed at optimizing the model's capacity to process and integrate visual and textual information more effectively.

Another aspect of our approach was exploring alternative methodologies like SmallCap and datastore retrieval [5]. SmallCap, in particular, provided a promising avenue for generating captions by conditioning on both the input image and related captions from a datastore, thereby introducing a novel retrieval-augmented captioning method. This approach was expected to enhance the contextual relevance of the captions generated by the model.

D. Computational Hardware Setup

The experiments leveraged a robust computational setup featuring the Nvidia RTX 4090 GPU, equipped with 24 GB of GDDR6X VRAM, advanced tensor cores, and high processing speeds, ideal for large-scale data handling and efficient deep learning model training. Complementing the GPU, an Intel i9-12900k CPU with 32 GB of DDR5 RAM ensured balanced computational power, enhancing the overall efficiency and feasibility of our advanced transformer model experiments.

IV. Experiments

A. Dataset and Evaluation Metrics

Our research employs the MSCOCO dataset, an extensive collection of over 200,000 labeled images, each annotated with 5 unique reference captions. This dataset is a benchmark for object detection, segmentation, and captioning studies. The 2014 version of the dataset contains 82783 for training, 40504 for validation, and 40775 images for online testing. For our experiments, we used the well-established 'Karpathy' [8] split to redivide the training images: 113,287 for training, 5,000 for validation, and 5,000 for offline evaluation. The dataset size exceeds 25 GB, capturing a wide array of 80 object categories and more than 1.5 million object instances, thus providing a diverse and detailed basis for our image captioning model.

For evaluation, while multiple metrics exist for evaluating the quality of generated captions, our study focused on the CIDEr (Consensus-based Image Description Evaluation) metric [10], which offers a comprehensive measure of the similarity of machine-generated captions to human-written references. CIDEr uses Term Frequency-Inverse Document Frequency (TF-IDF) weighting to evaluate the significance of each word, representing captions as vectors and calculating cosine similarity for consensus scoring across all n-grams. The CIDEr metric is expressed as follows:

$$\text{CIDEr}_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{g^n(c_i) \cdot g^n(s_{ij})}{\|g^n(c_i)\| \|g^n(s_{ij})\|} \tag{3}$$

considering the sum of consensus scores for all n-grams in the candidate caption, thereby providing a holistic view of caption quality.

B. Experimental Settings

Our experiments are carefully structured to replicate and build upon the work done by Wang et al. [1]. We set our model's embedding size D to 512 and used 8 transformer heads, with the refining encoder and decoder consisting of 3 blocks each. The initial training phase under Cross-Entropy (XE) loss spans 20 epochs, setting a batch size of 10 and warming up with 10,000 steps. This phase is followed by 30 epochs of training under Self-Critical Sequence Training (SCST) with a fixed learning rate of 5×10^{-6}. We employ the Adam optimizer for both phases and a beam size of 5 during validation and testing.

TABLE I
Performance Metrics by Backbone and Training Phase

Backbone	XE Training						SCST Training					
	B-1	**B-4**	**M**	**R**	**C**	**S**	**B-1**	**B-4**	**M**	**R**	**C**	**S**
SwinV1	77.1	36.6	28.9	57.2	**121.5**	21.9	82.1	40.9	30.2	60.1	**138.2**	24.2
SwinV2	78.0	37.1	28.9	57.7	**122.5**	22.0	82.4	41.4	30.1	60.2	**138.2**	24.0
CSwin	75.4	34.9	27.0	55.6	**122.4**	20.2	N/A	N/A	N/A	N/A	N/A	N/A
DeiT	75.9	35.2	28.3	56.4	**117.0**	21.3	N/A	N/A	N/A	N/A	N/A	N/A

C. Experiment 1: PureT Replication with SwinTransformer Backbone

The first experiment aimed to replicate the baseline results achieved with the Pure Transformer (PureT) model for image captioning using the SwinTransformer backbone, as specified by Wang et al. [1]. We used the same hyperparameters as the original study, including the word embedding dimension and attention features. The model was trained using XE loss, achieving a CIDEr score of 121.5% in 20 epochs with 12 GB of VRAM, and then further trained with SCST, which pushed the CIDEr score to 138.2% in 27 epochs utilizing 23 GB of VRAM. Each epoch in the XE phase took approximately 1 hour, while the SCST phase saw a training time of 4-5 hours per epoch.

Upon completion of the training phase, our results closely mirrored the findings reported in the original study, with the CIDEr metric score of 138.2% aligning with the published benchmark. This successful replication validated our approach and reinforced the PureT model's robustness as a solid foundation for further experimentation.

D. Experiment 2: SwinV2 Transformer Backbone

The second experiment introduced the SwinV2 Transformer [2], adapted from the official Microsoft repository, as the backbone of our PureT model. SwinV2 introduces improvements like moving layer normalization from front to back and using scaled cosine attention to deal with unstable training and decaying activations. SwinV2 also introduced log-spaced coordinates, enhancing transfer learning and fine-tuning across diverse datasets. Our training incorporated various datasets: each backbone was initially trained on ImageNet 22k (224x224 resolution), fine-tuned on ImageNet 1k (384x384 resolution), and then applied to the MSCOCO dataset (384x384 resolution) to optimize performance across different image contexts.

The PureT model with the SwinV2 backbone improved the model's visual processing capabilities and efficiency. The XE training phase slightly increased the CIDEr metric to 122.5% in 20 epochs, each taking 45-50 minutes with 12 GB of VRAM. Remarkably, SCST training matched the target CIDEr score in just 11 epochs, suggesting the potential for higher efficiency, with training times reduced to 3 hours per epoch and 23 GB of VRAM, and potential for even higher performance with extended training.

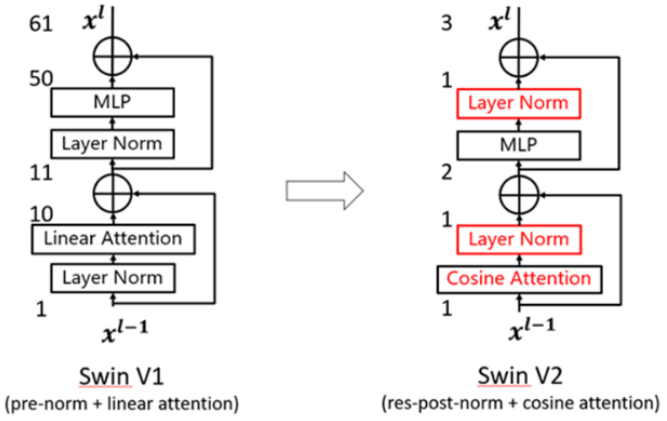

Fig. 2. Comparison of Swin V1 and Swin V2 Architectures. This figure showcases the distinct features of Swin V1 and Swin V2 backbones, with emphasis on layer normalization and scaled cosine attention in Swin V2, marked in red.

E. Experiment 3: CSwin Transformer Backbone

For the third experiment, we implemented the CSwin Transformer [3] backbone, characterized by its unique cross-shaped attention mechanism and locally-enhanced position encoding. The backbone was integrated to improve efficiency, as evidenced by reduced grid feature dimensions (1152). In the XE training phase, the model achieved a CIDEr score of 122.3% in 20 epochs, each taking 40-45 minutes and using 9 GB of VRAM. Due to time constraints and long training times, the SCST training phase was not attempted, but the initial results indicated a potential for enhanced efficiency.

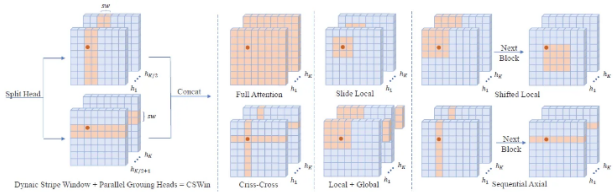

Fig. 3. CSwin Transformer: Cross-Shaped Window Attention Mechanism. This illustration explains the cross-shaped window attention mechanism in the CSwin Transformer, demonstrating its role in increasing the model's efficiency.

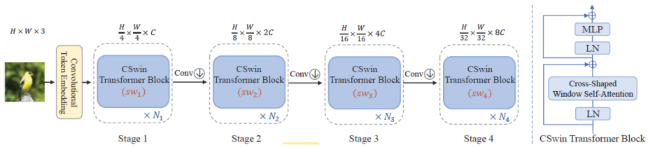

Fig. 4. Architecture Overview of the CSwin Transformer. The figure provides a detailed view of the CSwin Transformer architecture.

F. Experiment 4: DeiT Backbone

The fourth experiment explored the DeiT backbone [4], sourced from Hugging Face's transformers library. Adjustments were made to the grid features dimension (768) and encoder window size (17x17) to accommodate the DeiT model's specifics. The XE training phase yielded a CIDEr score of 117.0% in 20 epochs, with an efficient training time of 30 minutes per epoch and 7.5 GB of VRAM usage. The SCST training was not attempted as the preliminary results were slightly lower than with other backbones, but the DeiT model demonstrated promising efficiency.

TABLE II
TRAINING TIME, VRAM USAGE, AND NUMBER OF EPOCHS BY BACKBONE

Backbone	XE Training			SCST Training		
	Time/Epoch	VRAM	Epochs	Time/Epoch	VRAM	Epochs
SwinV1	1 hour	12 GB	20	4-5 hours	23 GB	27
SwinV2	45-50 mins	12 GB	20	3 hours	23 GB	11
CSwin	40-45 mins	9 GB	20	3 hours*	18 GB*	N/A
DeiT	30 mins	7.5 GB	20	2.5 hours*	15 GB*	N/A

*SCST training times and VRAM usage are approximations based on single epoch training or other experiments.

G. Results of Experiments

In our comparative analysis, each transformer backbone demonstrated unique strengths in enhancing the PureT model. SwinV2, adapted from the official Microsoft repository, showed remarkable improvements in CIDEr scores and training efficiency, surpassing SwinV1 in both aspects. The CSwin backbone, characterized by its cross-shaped attention mechanism, displayed increased efficiency and comparable CIDEr scores in the XE training phase despite not undergoing SCST training. DeiT, though slightly behind in CIDEr scores, stood out for its significant training efficiency. These results underscore the varying potentials of these backbones in improving

the performance and efficiency of image captioning models, highlighting their distinct contributions.

V. Discussion

In this section, we reflect on our paper's comprehensive outcomes and future directions:

A. Key Achievements and Contributions

Our study achieved notable success in replicating and enhancing the PureT model. We successfully implemented the PureT model with the original SwinTransformer, achieving benchmark CIDEr scores. The introduction of SwinV2 not only matched but surpassed the original model's performance. While the CSwin and DeiT backbones didn't undergo SCST training, they demonstrated potential in improving efficiency and performance in XE training.

B. Opportunities for Future Work

Many opportunities for future research have emerged from our findings. Hyperparameter optimization remains crucial, offering the potential to elevate model performance further. Additionally, exploring novel architectural enhancements could pave the way for more sophisticated and efficient image captioning models. For example, introducing Block Static Expansion [11] in the encoder could further enhance model capabilities.

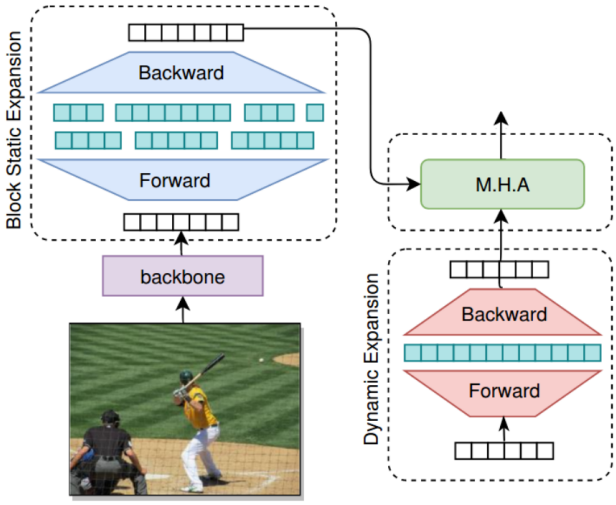

Fig. 5. Illustration of Block Static Expansion in Image Captioning. The expansion mechanism transforms the input data during the forward phase and performs the reverse operation in the backward pass, enabling the network to process inputs unconstrained by the number of elements. Block Static Expansion provides a way of performing these operations over a collection of arbitrary and diverse lengths at the same time [11].

C. Exploration of SmallCap

While waiting for other models to train, we explored SmallCap [5] as a potential solution to alleviate high costs and training times associated with large-scale image captioning models. Although we did not directly experiment with SmallCap, its lightweight nature and retrieval augmentation approach make it an intriguing option for future research. SmallCap's architecture, involving a pre-trained CLIP encoder and GPT-2 as a decoder. The model offers an innovative approach for generating multiple (N=4) contextually relevant captions and then selecting the best among them with reduced training demands. Its ability to adapt to new domains without retraining and utilize retrieval from target-domain data positions it as a promising solution to the challenges of high costs and lengthy training times in large-scale captioning models.

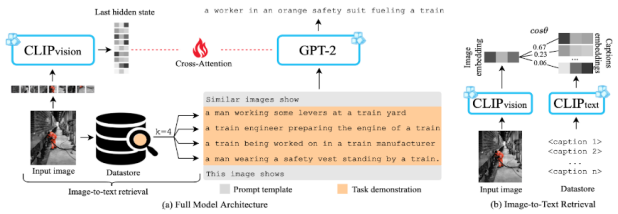

Fig. 6. SmallCap Image Captioning Model. This figure illustrates the SmallCap model, emphasizing its use of a CLIP encoder and GPT-2 decoder for effective caption generation

VI. Conclusion

Our study contributes to the ongoing development of image captioning techniques, showcasing the effectiveness of incorporating various transformer backbones into the Pure Transformer (PureT) model. We have demonstrated that enhancements like SwinV2 and CSwin can significantly improve the model's performance and efficiency. Our findings underscore the transformative potential of these advanced architectures in Natural Language Processing and Computer Vision. Looking ahead, novel approaches like SmallCap and architectural innovations like Block Static Expansion suggest the potential for further advancements. This paper provides a foundation for further advancements in transformer-based image captioning, highlighting the potential of these models to bridge visual perception and linguistic expression effectively.

References

[1] Y. Wang, J. Xu, and Y. Sun, "End-to-End Transformer Based Model for Image Captioning," 2022. [Online]. Available: https://arxiv.org/abs/2203.15350

[2] Z. Liu et al., "Swin Transformer V2: Scaling Up Capacity and Resolution," 2021. [Online]. Available: https://arxiv.org/abs/2111.09883

[3] X. Dong et al., "CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows," 2021. [Online]. Available: https://arxiv.org/abs/2107.00652

[4] H. Touvron et al., "Training data-efficient image transformers & distillation through attention," 2020. [Online]. Available: https://arxiv.org/abs/2012.12877

[5] R. Ramos, B. Martins, D. Elliott, and Y. Kementchedjhieva, "SmallCap: Lightweight Image Captioning Prompted with Retrieval Augmentation," 2022. [Online]. Available: https://arxiv.org/abs/2209.15323

[6] A. Vaswani et al., "Attention Is All You Need," 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[7] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[8] A. Karpathy, and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," 2014. [Online]. Available: https://arxiv.org/abs/1412.2306

[9] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical Sequence Training for Image Captioning," 2016. [Online]. Available: https://arxiv.org/abs/1612.00563

[10] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based Image Description Evaluation," 2014. [Online]. Available: https://arxiv.org/abs/1411.5726

[11] J. C. Hu, R. Cavicchioli, and A. Capotondi, "ExpansionNet v2: Block Static Expansion in fast end to end training for Image Captioning," 2022. [Online]. Available: https://arxiv.org/abs/2208.06551