Speaking the Language of Sight

Stanford CS224N {Custom} Project

Jean Rodmond Jr Laguerre
Department of Computer Science
Stanford University
jeanlag1@stanford.edu

Vicky Wu
Department of Computer Science
Stanford University
vwu2010@stanford.edu

Abstract

Our aim is to address the challenge of producing overly generic image captions. Despite the expressive power of language and the adage "a picture is worth a thousand words," image captioning often falls short in providing rich and descriptive outputs. Our approach involves leveraging the BLIP-2 model, pretrained on the MSCOCO dataset, as the foundation. We then fine-tune this base model using the Maximum Likelihood Estimation (MLE) objective to enhance accuracy and conciseness. Subsequently, we incorporate the Semipermeable Maximum Likelihood Estimation (SMILE) objective to promote richness and descriptiveness in the generated captions. Finally, we adopt a hybrid approach, mixing MLE and SMILE into a single weighted average objective to strike a balance between accuracy and richness in the generated captions. By combining the strengths of MLE and SMILE, our model achieves significant improvements in caption quality, paving the way for more nuanced and contextually grounded image descriptions. Our work represents a step towards improving accessibility and inclusivity in the digital landscape, empowering individuals with visual impairments to access and enjoy a wide range of visual media.

1 Key Information to include

- Mentor (CS 224n): Kaylee Burns

2 Introduction

In the realm of accessibility, bridging the gap between visual content and individuals with visual impairments remains a pressing challenge. While the adage 'A picture is worth a thousand words' rings true, unlocking these visual narratives for those unable to see poses significant hurdles. Conventional image captioning methods, while commendable, often fall short in evoking the vivid imagery and emotional resonance inherent in visual experiences. They tend to produce descriptions that lack depth and fail to capture the nuances present in images, leaving much to be desired in terms of accessibility and storytelling (Karpathy and Fei-Fei, 2015; Vinyals et al., 2015; You et al., 2016). Our research endeavors to address this gap by introducing an innovative image captioning model that transcends these limitations, going beyond mere description, crafting immersive captions that resonate with readers on a deeper level. Specifically, we explore a novel objective semi-permeable maximum likelihood estimation (SMILE) that maximizes lexical diversity and caption length, metrics optimized for richness (Yue et al., 2023), and combine it with the accuracy and conciseness optimization of traditional image captioning MLE, to create a one-shot model that creates captions that strike a balance between expressiveness and informativeness. Through our exploration, we aim to not only enhance accessibility but also to push the boundaries of storytelling at the intersection of computer vision and natural language processing, promising a more inclusive and engaging experience for all.

Our performance evaluation of the BLIP-2 model, fine-tuned with different objectives such as MLE and SMILE, closely matched our initial expectations. Notably, the BLIP-2 + SMILE model demonstrated superior performance across various evaluation metrics, indicating an enhancement in richness and descriptiveness within the generated captions. The SMILE objective, designed to optimize richness through lexical diversity and caption length, indeed yielded captions that were richer and more contextually grounded. However, this improvement was accompanied by a slight decrease in CLIPScore, potentially suggesting a misalignment with the original optimization focus of BLIP-2. Conversely, the inclusion of the MLE objective led to captions that were more concise and accurate, although perplexity decreased slightly. Despite this, there was a significant improvement in CLIPScore, indicating better alignment with the visual context. Our hybrid SMILE + MLE model demonstrated notable enhancements across various reference-based metrics compared to raw SMILE, striking a balance between expressiveness and accuracy while improving alignment with reference captions.

Through a thorough analysis of model outputs, we observed discrepancies between our model-generated captions and ground truth captions. MLE-dominant captions prioritized accuracy but lacked specificity, while SMILE-dominant captions were richer but occasionally hallucinated attributes. These findings underscore the inherent trade-offs between accuracy and richness in caption generation, highlighting the complexities of optimizing image captioning models for both expressiveness and precision.

3 Related Work

Image captioning has been a subject of extensive research in the fields of computer vision and natural language processing. Early approaches typically relied on handcrafted features and shallow learning algorithms to generate captions for images. These methods often struggled to capture the semantic complexity and contextual understanding required for accurate and expressive captioning.

With the advent of deep learning techniques, particularly convolutional neural networks (CNNs) for image processing and recurrent neural networks (RNNs) for sequence generation, significant progress has been made in the field of image captioning. Models such as Show and Tell (Vinyals et al., 2015) and Show, Attend, and Tell (Xu et al., 2016) introduced end-to-end trainable architectures that combined CNNs for image feature extraction with RNNs for caption generation. These models demonstrated improved performance in generating coherent and contextually relevant captions compared to earlier approaches.

Subsequent research has focused on refining and extending these architectures to address various challenges in image captioning. Attention mechanisms have been integrated into models to enable them to selectively focus on different regions of the image while generating captions, leading to more accurate and detailed descriptions (Lu et al., 2017; Anderson et al., 2018). Additionally, techniques such as reinforcement learning have been employed to optimize captioning models for specific evaluation metrics, further improving their performance (Rennie et al., 2017).

Recent advancements in image captioning have also seen the exploration of multimodal architectures that leverage both visual and textual information for caption generation. Models such as VisualBERT (Li et al., 2019) and UNITER (Chen et al., 2020), Meshed-Memory Transformer (MMT) (Cornia et al., 2020), and VL-BERT (Su et al., 2020) integrate pre-trained language models with image features to enhance the understanding of visual context and improve caption quality.

In reviewing the literature, it becomes evident that while significant advancements have been made in image captioning, notable limitations persist, particularly concerning the richness and expressiveness of generated captions. The current state-of-the-art methods fall short in delivering captions that truly resonate with the depth and complexity of the visual world (Al-Malla et al., 2022; Krause et al., 2017; Yue et al., 2023; Shi et al., 2021). Al-Malla et al. (2022) propose an attention-based Encoder-Decoder model incorporating convolutional and object features, alongside an "importance factor" positional encoding scheme, to address quality shortcomings in caption generation. Krause et al. (2017) address the richness limitations of existing image captioning methods by introducing a model that generates entire paragraphs to describe images in finer detail, overcoming the constraint of compressing visual content into single sentences. Yue et al. (2023) propose a new training objective, Semipermeable Maximum Learning Estimation (SMILE), designed to optimize caption generation models for lexically rich and descriptive outputs. Shi et al. (2021) introduce a novel approach leveraging natural language inference and directed inference graphs to guide captioning models towards producing more detailed and informative descriptions.

Inspired by the "semipermeable maximum likelihood estimation" (SMILE) approach proposed by Yue et al. (2023), we aimed to develop a novel methodology that addresses these presiding challenges directly. Based on our deep examination of various open-source image captioning outputs, SMILE stood out for its ability to capture the subtle details of visual content. Refer to Figure 7 in the Appendix for specific examples of SMILE-output samples. Our work marks an advancement in image captioning, with the promise of enhancing accessibility, enriching storytelling, and pushing the boundaries of computational understanding of visual content.

4 Approach

We take the novel state-of-the-art BLIP-2 model and pretrain it on our dataset (described in a later section). We then add MLE to develop a basic fine-tuned model. Then we optimize it with SMILE. Finally, we run a hybrid learning objective, mixing MLE and SMILE, to refine our model even further.

4.1 Baselines:

We use the new, base version of BLIP-2, a state-of-the-art language and vision model pre-trained on 129M images and paired captions (Li et al., 2023). BLIP-2 is built on the standard BLIP model (Li et al., 2022), and it further optimizes performance on three objectives: image-text contrastive learning (ITC), image-grounded text generation (ITG) and image-text matching (ITM) (Li et al., 2023).

After carefully considering our proposal paper summary on Kreiss et al. (2023), we added one more baseline for comparison: CapEnrich, the latest descriptive image captioning system (Yao et al., 2023). To review, Kreiss et al. (2023) provided staunch support for the usage of reference-less metrics in generating image captioning for the visually impaired. We use this baseline as to have a metric to evaluate descriptiveness through the performance of CLIP self-retrieval, which uses the CLIP model to retrieve and recall an image along with its caption from a candidate pool, the hard retrieval pool constructed in CapEnrich (Yao et al., 2023).

4.2 BLIP-2 Architecture

In their paper, Li et al. (2023) introduces BLIP-2, a model comprising of an image encoder and a Large Language Model (LLM), operating in a two-stage process for vision-language representation learning. In the first stage, a module called Q-Former bridges the gap between the frozen image encoder and LLM. Q-Former includes an image

transformer for visual feature extraction and a text transformer for text encoding and decoding. Learnable query embeddings guide the interaction between the image transformer and frozen image features, with self-attention layers facilitating query-query and query-text interactions. Different self-attention masks control query-text interaction based on pre-training tasks. In the second stage, BLIP-2 optimizes three objectives—image-text matching, image-text contrastive learning, and image-grounded text generation—to encourage relevant visual representation extraction. This unified framework enables effective interaction between visual and textual modalities for diverse vision-language tasks.

We choose BLIP-2 as our baseline due to its state-of-the-art vision and language modeling capabilities. One of the key features of BLIP is its attention mechanism, allowing the model to focus on different parts of the image while generating captions (Li et al., 2022). Additionally, the novel BLIP-2 incorporates

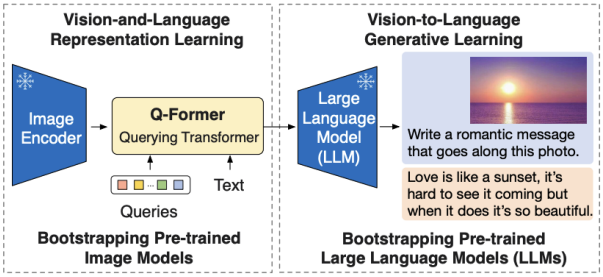

Figure 1: Overview of BLIP-2's framework (Li et al., 2023)

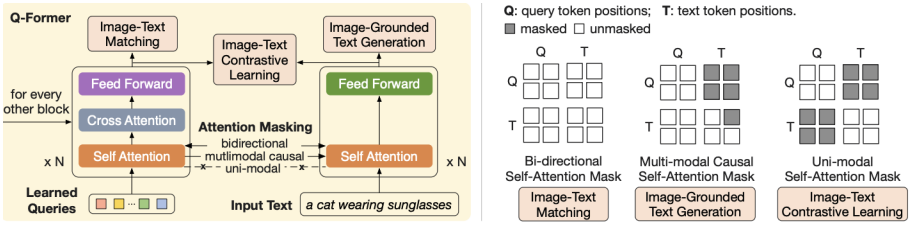

Figure 2: (**Left**) BLIP-2's Q-former and (**Right**) self-attention masking strategy (Li et al., 2023)

the Q-former architecture, enhancing its performance by leveraging structured attention mechanisms and efficient query processing (Li et al., 2023). This characteristic provides us with a baseline model already capable of generating accurate and descriptive captions by focusing on relevant image regions and utilizing efficient query processing. Building upon this foundation, we aim to further enhance captioning performance and richness of description.

4.3 Maximum Likelihood Estimation (MLE)

MLE has been a standard with many text generation NLP tasks, (Allahyari et al., 2017), (Stahlberg, 2020). Our objective to maximize the likelihood of a given label when predicting the current word w from some visul content v and a sequence of given previous words $w_<$. We then define the MLE token-level loss function as:

$$\mathcal{L}_{\text{MLE}} = -\sum_{j}^{|\mathcal{V}|} y_j \log \hat{P}^{\mathcal{V}}\left(w \mid w_<, v; \theta\right) \tag{1}$$

The summation goes through all the words of the vocabulary V where $\hat{P}^{V}$ is the predictive probability distribution over V and y_j is the j-th element of the one-hot label vector. This section details your approach to the problem.

4.4 Semipermeable Maximum Likelihood Estimation (SMILE)

This was proposed by Yue et al. (2023) as being a solution to the conciseness optimizing component of MLE, which is counter-intuitive to the training objective of producing descriptive image captions. Given a target sequence caption string $D = [w_1, ..., w_N]$, a subset of the vocabulary V_D can be formed where $V_D = \{w_i | w_i \in D\}$. The SMILE token-level loss function is then defined as:

$$\mathcal{L}_{\text{SMILE}} = -\sum_{j}^{|\mathcal{V}_D|} y_j \log \hat{P}^{\mathcal{V}_D}\left(w \mid w_<, v; \theta\right), \qquad \hat{p}_j = softmax\left(\mathbf{z}_j\right) = \frac{\exp\left(\mathbf{z}_j\right)}{\sum_{k \in \mathcal{V}_D} \exp\left(\mathbf{z}_k\right)} \tag{2}$$

where $\hat{P}^{V_D}$ is the predictive probability distribution over V_D and the probability for the j-th word in V_D, and $\hat{p}_j$ assigns probabilities over the subset V_D instead of the entire vocabulary. Thus in SMILE, we do not get penalized for the addition of more word terms beyond the ground truth caption as they do not contribution to the denominator in 2. This allows for longer and richer optimizations as the prediction loss assigned to the label is no longer penalized in the SMILE objective by the model.

4.5 Hybrid objective: MLE + SMILE

Though SMILE improves descriptiveness of captions, it does this at the cost of accuracy. MLE is accuracy optimized, maximizing the probability of generating the ground truth caption. Thus, striking a balance, we propose the following

mixed loss objective:

$$\mathcal{L}_{\text{hybrid}} = \lambda \cdot \mathcal{L}_{\text{MLE}} + (1 - \lambda) \cdot \mathcal{L}_{\text{SMILE}} \tag{3}$$

where $\lambda \in [0, 1]$. We will do testing as to find the best λ parameter that strikes the best balance between descriptiveness and accuracy on this weighted sum combined loss objective.

4.6 Originality and References

Utilizing BLIP-2 as a baseline is original, although the model itself is not (Li et al., 2023). MLE is not original (Allahyari et al., 2017), (Stahlberg, 2020). Additionally, SMILE is not original (Yue et al., 2023). However, utilizing a combined hybrid objective is original.

The code for BLIP-2 is not original and can be found here(Li et al., 2023). Yue et al. (2023) did provide starter code for a SMILE image captioning Hugging Face model, but there was no support for BLIP-2 nor for further tuning, so we utilized their code base for guidelines as to how to format and brainstorm our approach but had to implement it ourselves. We had to code up MLE and the hybrid approach ourselves.

5 Experiments

5.1 Data

We use MSCOCO as our dataset (Lin et al., 2014). MSCOCO contains about 120K images, and each image has five human-annotated captions (Figure 8). We pick this dataset due to its notable feature: detailed pixel-level segmentation (Figure 9). This attribute, accessible through COCO Explorer, allows us to delve into precise object boundaries, which are crucial for developing the nuanced understanding required to generate rich and descriptive image captions. "To ensure a robust evaluation, we adopt the Karparthy split, a widely-used seminal split for image captioning models, allocating 5,000 images to both the validation and test sets (Karpathy and Fei-Fei, 2015).

Reference image captions

A pizza on a pan sitting on a table.
A close up of a pizza in a pan on a table.
A pizza sits on a plate on a dark surface.
A person sitting at a table where a pizza is sitting.
A pizza topped with different toppings is brought to a table.

Figure 3: Sample image with captions

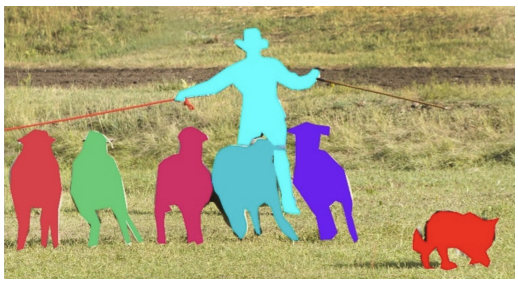

Figure 4: Sample of pixel-level individual object instance segmentation

5.2 Evaluation method

In evaluating richness, we employ metrics utilized by Yue et al. (2023), including caption length and lexical diversity (i.e., the number of unique words). Additionally, we utilize self-retrieval at R@1 and R@5, along with CLIPScore, as employed by Yao et al. (2023) to assess descriptiveness. These metrics gauge the model's proficiency in retrieving and aligning information, ensuring the generation of more contextually relevant and visually grounded captions; the pairing of the two fosters a more accurate and context-aware description of the visual content, as recommended by Kreiss et al. (2023). Perplexity is used to measure the language modeling performance, evaluating the model's ability to predict word sequences with a focus on capturing and quantifying uncertainty. Furthermore, in later stages of development, we incorporate BLEU, METEOR, CIDEr, SPICE, and ROUGE metrics to assess the model's robustness and generalizability through traditional reference-based metrics.

5.3 Experimental Details

Our experiments can be split into the four phases: (**1**) pretraining BLIP-2 on MSCOCO, (**2**) fine-tuning the **1** with the MLE objective, (**3**) fine-tuning **2** with the SMILE objective, and (**4**) fine-tuning **1** on a hybrid objective.

We ran with `alpha` $= 0.4$, `weight decay` $= 0.05$, `initial lr` $= 3e-4$, `lr decay` $= 0.9$, `image size` $= 224$, and `max epochs` $= 20$ for pretraining BLIP-2. This phase took the longest, at 18+ hours given that we had to set `batch size` $= 16$ due to memory constraints. We stopped training at around 18 hours. We noticed convergence at around 3 epochs, so we subsequently set the remaining experiments to `max epochs` $= 5$ as to conserve compute. The

rest of the trials ran around 11-13 hours, so we would run our models overnight and discuss results in the morning. Training plots for our SMILE model can be found in the Appendix.

5.4 Results

Our results for our fine-tuned BLIP-2, BLIP-2 + MLE, and BLIP-2 + SMILE models largely aligned with our expectations. were mostly as expected:

Table 1: BLIP-2, BLIP-2 + MLE, and BLIP-2 + MLE + SMILE Performance

Model	Cap. Len.	Lex. Div.	R@1	R@5	CLIPScore	PPL
`CapEnrich`	13.301	1.498	9.487	22.592	79.153	62.345
`BLIP-2`	10.032	1.404	6.702	16.604	77.312	95.804
`BLIP-2 + MLE`	10.025	1.400	6.497	16.591	77.552	67.564
`BLIP-2 + SMILE`	24.431	4.557	10.135	24.146	72.035	94.231
`Yue et al. (2023)'s SMILE`	22.3	4.5	10.0	24.5	75.0	95.6

Specifically, we anticipated that the SMILE model's richness optimization component would yield captions that were richer, more descriptive, and more contextually grounded in the given input picture. This expectation was largely confirmed, as the BLIP-2 + SMILE model exhibited higher scores in all evaluation metrics except for CLIPScore and PPL; pretrained BLIP-2 outperformed in PPL by 1.78%, while CapEnrich, pretrained BLIP-2, and BLIP-2 + MLE models outperformed in CLIPScore by 9.88%, 7.33%, and 7.65%, respectively.

For the lower CLIPScore within the BLIP-2 models, it is possible that there was a misalignment between what the SMILE training objective aimed to optimize and the effectiveness of BLIP-2's advanced attention and Q-former architecture. The pretrained BLIP-2, without further fine-tuning, is already optimized to generate accurate, descriptive, relevant, and highly contextually grounded captions. The addition of the SMILE objective may have unintentionally introduced some divergence from BLIP-2's original optimization focus, potentially leading the model to prioritize richness metrics like 'lexical diversity' over preserving relevant information crucial for contextual richness. This discrepancy is particularly evident when comparing the pretrained BLIP-2 to the BLIP-2 + MLE model, where their performance is nearly identical, with minor variations likely attributable to random factors during dataset splitting. Therefore, we can attribute the lower CLIPScore of the BLIP-2 + SMILE model to the unintended shift in optimization focus introduced by the SMILE objective.

We do expect to observe a lower CLIPScore in CapEnrich though, as this metric serves as our upper bound or goal. CapEnrich is specifically designed to maximize CLIPScore and generate contextually relevant descriptions of visual content (Yao et al., 2023). However, the SMILE objective, which focuses on optimizing richness through maximizing lexical diversity and caption length, doesn't prioritize maximizing context relevance. Therefore, lower CLIPScore in the SMILE model with this expectation.

The slight decrease in PPL, although marginal at just 1.78% lower than the pretrained BLIP-2, can likely be attributed to the inclusion of the MLE objective. Our approach involved pretraining BLIP-2, followed by optimization with the MLE objective, and then further refinement with the SMILE objective. Notably, there was a significant 41.79% decrease in PPL from pretrained BLIP-2 to BLIP-2 + MLE. Since MLE aims to optimize for conciseness and accuracy, it may have led to the generation of shorter and more precise captions, potentially at the expense of capturing nuanced details and uncertainties within the language, leading to a more mechanical sounding caption. This ultimately is probably not the case, as the caption length and lexical diversity between pretrained BLIP-2 and BLIP-2 + MLE are nearly identical. Other factors, such as subtle changes in word choice or syntactic structure introduced by the MLE optimization, may contribute to the observed decrease in perplexity.

Additionally, our BLIP-2 + SMILE ended up outperforming Yue et al. (2023)'s SMILE model, which inspired our method. While our model exhibits slightly lower performance in CLIPScore and PPL, we consider these differences negligible given the constraints of our project timeline and the computational resources available.

Table 2: Hybrid Model Performance with Varying λ

λ	Caption Length	Lexical Diversity	R@1	R@5	CLIPScore	PPL
1.00	9.7902	1.366	6.526	16.257	75.646	93.686
0.75	11.774	1.089	6.303	17.328	74.153	65.585
0.50	10.578	1.364	6.526	17.444	75.392	65.291
0.25	12.422	1.660	7.248	18.001	74.029	67.120
0.10	12.331	1.851	7.389	17.723	74.424	67.160
0.05	14.379	2.233	8.355	20.282	74.375	72.369
0.00	23.340	4.392	9.791	24.012	73.751	93.599

The table above depicts the performance trend of our hybrid SMILE + MLE model as λ, ranging from 0 to 1, interpolates between SMILE and MLE objectives. A decrease in lambda leads to richer, more descriptive captions, evident from increased caption length and lexical diversity. Conversely, an increase in lambda results in more concise and accurate captions. Metrics such as R@1, R@5, and CLIPScore improve with lower lambda values, reflecting better retrieval and

alignment, while perplexity (PPL) increases, indicating a decline in language modeling performance. This trade-off highlights the balance between richness and accuracy controlled by lambda in caption generation.

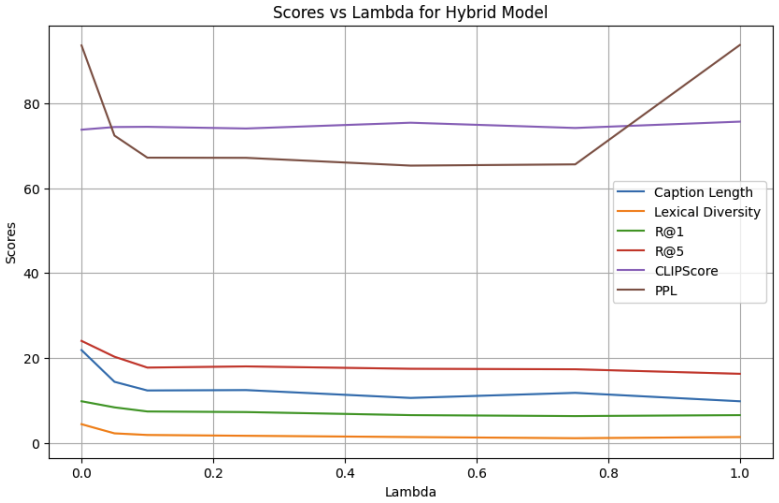

Figure 5: Trend lines for varying λ

However, there is a notable discrepancy in the trend, regarding perplexity (PPL). PPL initially decreases as λ decreases, indicating an improvement in language modeling performance associated with richer captions. This decrease aligns with the expectation that more descriptive captions capture a wider range of language patterns, reducing uncertainty in word prediction. However, beyond a certain point, PPL starts to increase as lambda approaches 1. This reversal could be attributed to the increasing influence of the MLE objective, which, while optimizing for accuracy, may tend to produce more predictable, repetitive language patterns, resulting in higher perplexity.

6 Analysis

For analysis, we incorporate traditional reference-based metric performance as to assess the model's robustness and generalizability across broad image captioning tasks. We use our BLIP-2 + MLE, BLIP-2 + SMILE, and BLIP-2 + SMILE + MLE @$\lambda = 0.5$ to produce the following results:

Table 3: Reference-based Metric Performance

Model	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**	**METEOR**	**ROUGE-L**	**CIDEr**	**SPICE**
MLE	0.7890	0.6375	0.5048	0.3968	0.3095	0.5997	1.3330	0.2378
SMILE	0.3326	0.1692	0.0919	0.0490	0.1828	0.3021	0.0978	0.1260
SMILE + MLE	0.4222	0.2408	0.1442	0.0862	0.2164	0.3711	0.2406	0.1589

The combination of the SMILE + MLE demonstrates notable improvements across various evaluation metrics when compared to raw SMILE. Specifically, the combined approach yields higher Bleu scores across all n-gram orders, indicating enhanced alignment with reference captions in terms of word overlap. Moreover, the METEOR score shows improvement, suggesting that the combined captions are more linguistically similar to reference captions. Similarly, the ROUGE-L score exhibits enhancement, reflecting improved overlap between generated and reference captions. Additionally, the slight increase in the CIDEr score indicates improved consensus-based image description evaluation. Furthermore, the SPICE score demonstrates improvement, highlighting enhanced semantic content in the generated captions.

It is worth noting that while the SMILE + MLE approach may underperform compared to traditional raw MLE in terms of some reference-based metrics, this can be attributed to MLE optimizing specifically for accuracy, the objective that these metrics evaluate. Nonetheless, the observed improvements align with our hypothesis of the addition of MLE to our objective enhancing these reference-based metric scores, as MLE optimizes for accuracy.

Furthermore, we conducted a comprehensive review of the captions generated by our models. This evaluation encompassed examining samples from the MSCOCO dataset to compare the generated output with ground truth captions. Additionally, we analyzed captions for ambiguous real-world inputs, including both our own images and AI-generated images.

We reviewed 50 MSCOCO images in total, and we found that the image in Figure 6 particularly emphasized the primary themes of discrepancies between our model's outputs and the ground truth captions. In scenarios where MLE predominates, our model tends to prioritize accuracy by generalizing rather than providing specific details. For instance, instead of mentioning a baseball field, the model simply describes the men as being "in dirt." Similarly, it offers a generic

TRUTH: The man at bat readies to swing at the pitch while the umpire looks on

λ = 0.25: a group of men that are standing up in the dirt with baseball bats.

λ = 0.75: A group of three men who are standing up against each other in front of home plate.

Figure 6: MSCOCO image captioned sample

description of the individuals in the image as "a group of men," rather than providing specific attributes. Despite the emphasis on accuracy, we observed instances where the model's output could still be inaccurate. For instance, the model incorrectly identifies the presence of multiple "baseball bats" in the image, whereas in reality, there is only one.

When SMILE dominates, our model indeed furnishes richer and more descriptive captions. Although the description of the people remains generic, it specifies the exact count: three. Initially, we questioned the accuracy of our model's output, only to discover that there were indeed three individuals in the image upon closer inspection. Notably, our model even identified the presence of "home plate," a detail overlooked in the ground-truth caption. However, we encountered a common issue with SMILE-dominant models, wherein they tend to hallucinate attributes to enrich the captions, often resulting in inaccuracies. In this instance, despite not significantly extending the caption length or lexical complexity, the model erroneously described the three men as "standing up against each other." Yet, upon further reflection, we realized that the stance of the players resembled a confrontational posture, suggesting potential over-fitting of the model to conflict scenarios rather than baseball-specific ones. This observation underscores a limitation of our approach: while we meticulously analyze image context, there remains a gap in integrating disparate elements seamlessly. Consequently, despite correctly identifying the baseball context, the model struggled to reconcile the stance of the players with the presence of home plate and other visual cues.

7 Conclusion

Our study aimed to enhance image captioning by leveraging novel objectives like SMILE alongside traditional methods such as MLE. Through experimentation, we found that integrating SMILE led to richer, more contextually grounded captions, albeit with a slight decline in CLIPScore, while MLE improved caption conciseness and accuracy, with notable gains in CLIPScore. Our hybrid approach combining SMILE and MLE struck a balance between expressiveness and accuracy, yielding captions that outperformed raw SMILE across various metrics.

Key achievements of our work include the development of a model capable of generating immersive captions that resonate on a deeper level, contributing to both accessibility and storytelling in image captioning. Additionally, our findings underscore the importance of balancing competing objectives in caption generation, offering insights into the complexities of optimizing image captioning models.

However, our study also has limitations. One primary limitation is the potential divergence from the original optimization focus of the base model when incorporating additional objectives, as evidenced by the slight decrease in CLIPScore with SMILE. Additionally, while our hybrid approach showed promise, further refinement is needed to fully exploit the synergies between SMILE and MLE objectives. Moreover, our analysis of model outputs revealed trade-offs between accuracy and richness, highlighting the ongoing challenges in optimizing image captioning systems.

In summary, our study advances the understanding of image captioning methodologies and offers a promising framework for future research. By addressing the limitations identified and refining our approach, we can continue to push the boundaries of image captioning technology, ultimately improving accessibility and storytelling in visual content interpretation.

8 Team Contributions

Jean built the MLE and SMILE model, and Vicky built the hybrid model. Jean did BLIP-2 pretraining, MLE fine-tuning, and SMILE fine-tuning training and testing. Vicky did the hybrid model training and testing. Both members wrote up the report.

References

Mohammad A Al-Malla, Ahmed Jafar, and Nizar Ghneim. 2022. Image captioning model using attention and object features to mimic human image understanding. *Journal of Big Data*, 9(1):20.

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. 2017. Text summarization techniques: A brief survey.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137. IEEE.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs.

Elisa Kreiss, Eric Zelikman, Christopher Potts, and Nick Haber. 2023. Contextref: Evaluating referenceless metrics for image description generation.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollar. 2014. Microsoft COCO: Common Objects in Context. `https://cocodataset.org/#home`.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning.

Zhan Shi, Hui Liu, and Xiaodan Zhu. 2021. Enhancing descriptive image captioning with natural language inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 269–277, Online. Association for Computational Linguistics.

Felix Stahlberg. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2016. Show, attend and tell: Neural image caption generation with visual attention.

Linli Yao, Weijing Chen, and Qin Jin. 2023. Capenrich: Enriching caption semantics for web images via cross-modal pre-trained knowledge.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention.

Zihao Yue, Anwen Hu, Liang Zhang, and Qin Jin. 2023. Learning descriptive image captioning via semipermeable maximum likelihood estimation.

A Appendix (optional)

Model	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE-L	CIDEr	SPICE
MLE	0.7911	0.6364	0.5032	0.3982	0.3123	0.6003	1.3340	0.2386
SMILE	0.3298	0.1715	0.0927	0.0502	0.1807	0.2996	0.0973	0.1266
SMILE + MLE	0.4244	0.2383	0.1454	0.0868	0.2175	0.3724	0.2398	0.1603

Table 4: Reference-based Metrics for Training

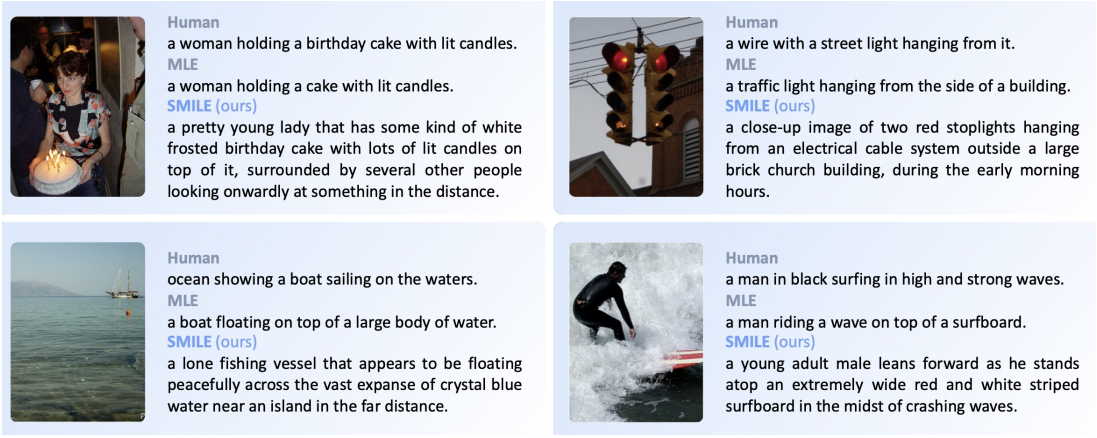

Figure 7: Descriptive caption samples generated by Yue et al. (2023) using their SMILE-optimized model, paired with the ground-truth human annotation and default MLE captions

Reference image captions
A pizza on a pan sitting on a table.
A close up of a pizza in a pan on a table.
A pizza sits on a plate on a dark surface.
A person sitting at a table where a pizza is sitting.
A pizza topped with different toppings is brought to a table.

Figure 8: Sample image with captions

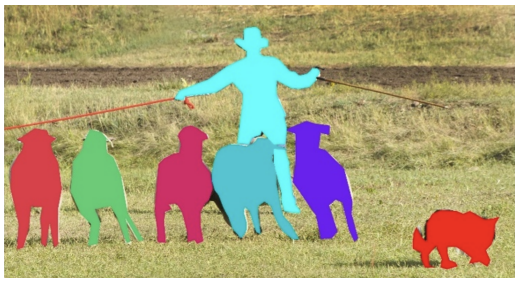

Figure 9: Sample of pixel-level individual object instance segmentation

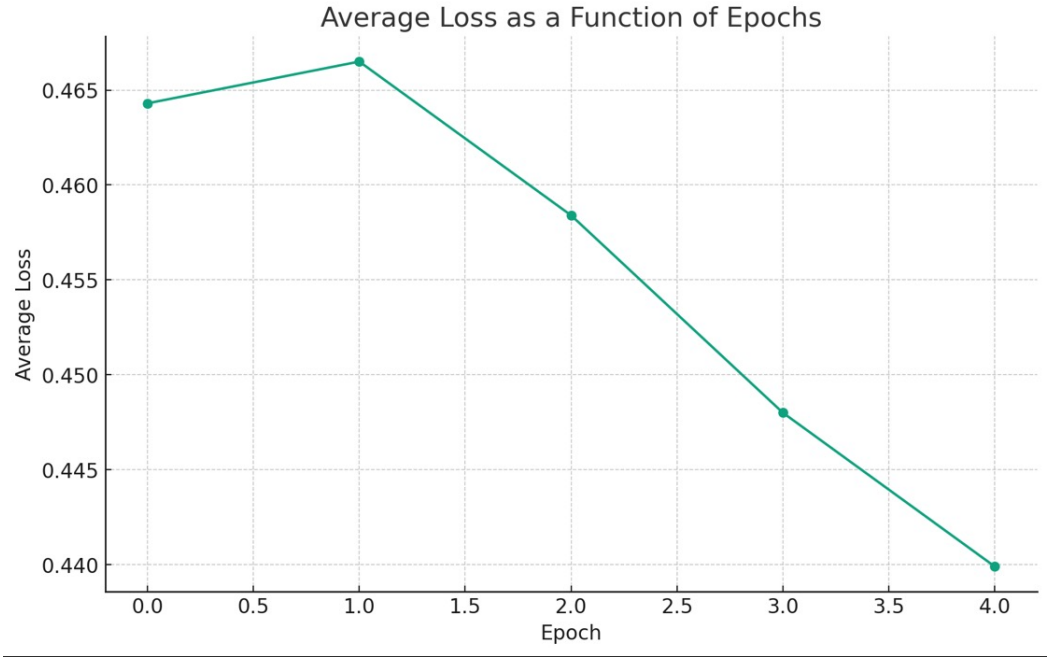

Figure 10: Progression of loss through training

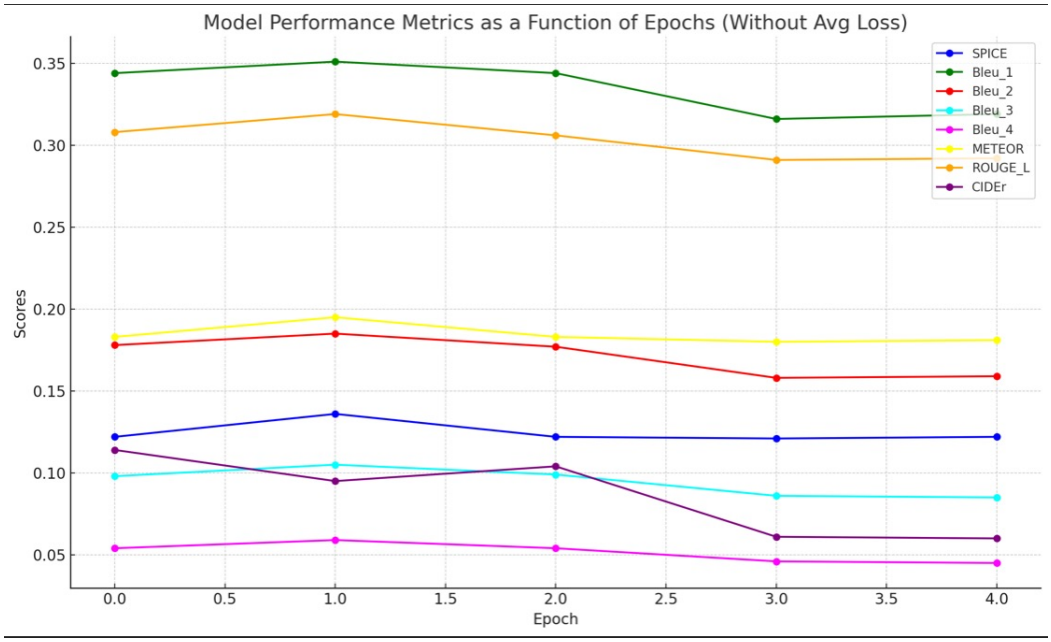

Figure 11: Reference-based metric scores through training

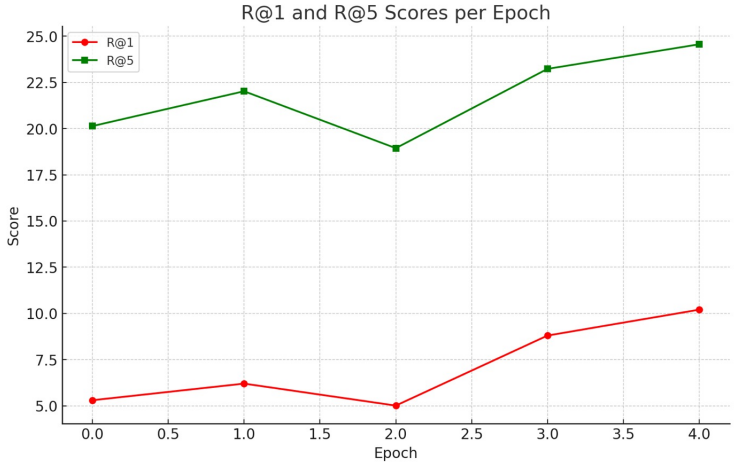

Figure 12: Progression of recall scores through training

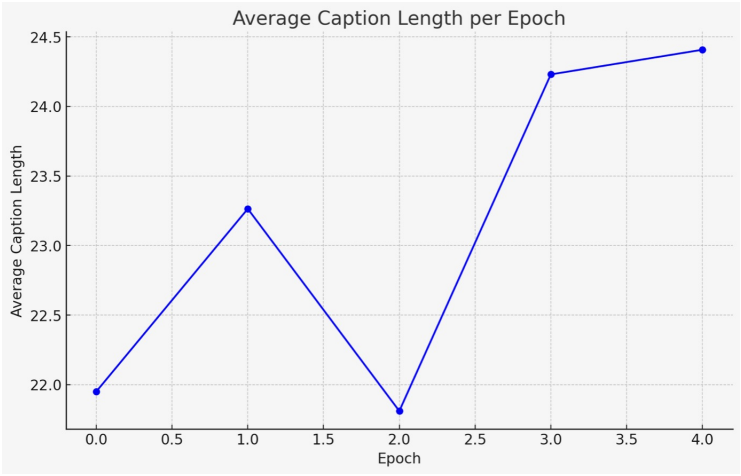

Figure 13: Progression of caption length through training

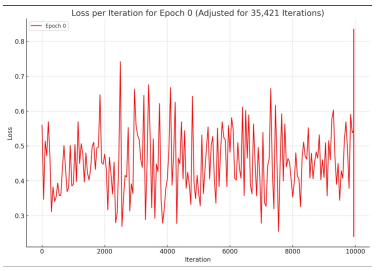

Figure 14: Caption for Figure 1

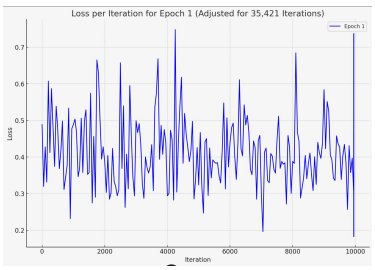

Figure 15: Caption for Figure 2

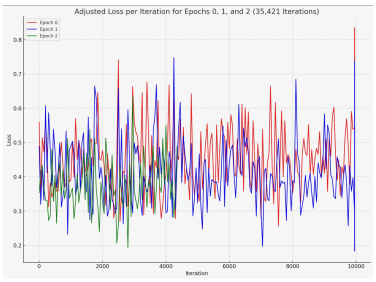

Figure 16: Caption for Figure 3

λ = 0.75: A close - up picture shows what appears to be an extremely long exposure shooting device next to a bright yellow fire hydrant, surrounded by some ornamental elements such as palm frondies and roots

Figure 17: Caption generated for AI generated Image

λ = 0.25: A close-up shot of two young adults shaving their furry gray and white husky dog's foreheads as they sit on the hard wood floor next to each other side.
λ = 0.75: a pretty young lady kneeling next to an adorable gray and white collie dog, which appears to be biting someone's fingers with both hands-on top part of them

Figure 18: Caption generated for real-world image