

Deep Learning-Based Color Transfer Biomedical Imaging Technique

MAHESH TUBAKI

Lecturer in E and C Engineering Department
maheshtubaki1@gmail.com
Government Polytechnic Belagavi

ABSTRACT

Text similarity targets are commonly used to train modern picture captioning algorithms. Models trained with text similarity objectives, on the other hand, tend to disregard unique and nuanced characteristics of a picture that distinguish it from others, because reference captions in public datasets generally identify the most conspicuous common things. We suggest leveraging CLIP, a multimodal encoder trained on large image-text pairings from the web, to calculate multimodal similarity and utilize it as a reward function in order to generate more detailed and distinctive captions. During the incentive calculation, this eliminates the requirement for reference captions entirely. We offer FineCapEval, a novel dataset for caption evaluation with fine-grained criteria: overall, background, object, and relations, to thoroughly assess descriptive captions. The suggested CLIP guided model generates more unique captions than the CIDEr-optimized model in our text-to-image retrieval and FineCapEval studies. We also show that our unsupervised grammar finetuning of the CLIP text encoder solves the basic CLIP reward degeneration problem. Finally, we show that, according to various parameters, annotators greatly prefer the CLIP reward over the CIDEr and MLE objectives.

Keywords: CIDEr, reward, optimization, Deep Learning, Medical Images

I. INTRODUCTION

Many applications, such as providing text keys for the image search engine and accessibility for the visually impaired, require describing a picture in depth and identifying features. The textual similarity between produced and reference captions is maximized in standard deep learning algorithms to train an image conditioned language model (Vinyals et al., 2015; Xu et al., 2015; Ronnie et al., 2017; Anderson et al., 2018). The reference captions of public datasets, on the other hand, frequently just describe the most conspicuous things in the photographs. As a result, models trained to maximize textual similarity with reference captions tend to create captions that are less distinctive and disregard the subtle details that differentiate one picture from others.

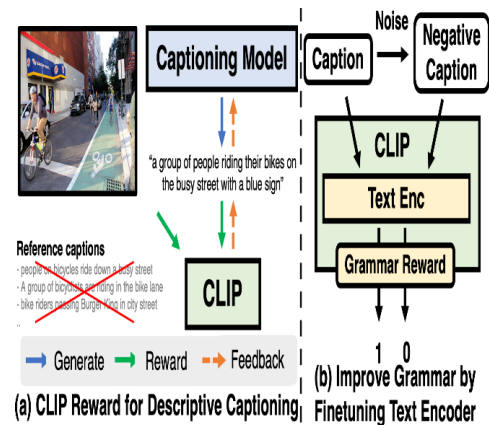


Fig No. 1: Over view of bio medical color transferring technique

To solve the problem, we propose employing CLIP (Radford et al., 2021), a multimodal encoder model trained on vast image-text data (mainly English) acquired from the web, and rewarding it based on its similarity scores (Sec. 3.1). Furthermore, we present a CLIP text encoder fine-tuning technique with synthetic negative caption augmentation to enhance the grammar of the captioning model without the need of any additional text annotations (Sec. 3.2). It's worth noting that our method fully eliminates the requirement for reference captions for calculating rewards. Figure 1

shows how we went about it. FineCapEval, a novel dataset that assesses captioning in several characteristics such as overall, backdrop, object, and relation between objects, is also introduced to thoroughly evaluate descriptive captions (Sec. 4).

We show that captions from models trained with CLIP reward are more distinctive and include more specific information than captions from CIDER (Vedantam et al., 2015)-optimized models in our tests on the MS COCO (Lin et al., 2014) dataset. CLIP-guided captions even outperform reference captions that were previously matched with photos in terms of text-to-image retrieval. We also show that fine-tuning our text encoder improves caption grammars greatly by minimizing degenerative artefacts like word repetition. We show that our CLIP-based incentives surpass text similarity targets by a substantial margin in all categories in fine-grained caption assessment with FineCapEval and human analysis.

II. RELATED WORK

Metrics for Image Captioning, Similarity measures based on n-grams or scene graphs, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), CIDER (Vedantam et al., 2015), and SPICE (Anderson et al., 2016), have traditionally been used to assess captions. Due to the restricted amount of reference captions or scene-graphs, such measures frequently fail to identify paraphrased sentences. Recent papers such as BERT Score (Zhang et al., 2019), ViLBERT Score (Lee et al., 2020a), UMIC (Lee et al., 2021), and CLIP Score (Hessel et al., 2021) suggest employing relevance scores computed by linguistic or multimodal models retrained on huge data to solve the problem.

Image captioning objectives, Models are trained using a maximum likelihood estimation (MLE) goal in standard deep learning-based picture captioning systems. MLE, according to Ranzato et al. (2016), has an exposure bias problem. Bengio et al. (2015) suggest a curricular learning technique called planned sampling to reduce exposure bias. With REINFORCE, Ranzato et al. (2016) propose training models by directly maximizing text

similarity between produced and reference captions (Williams, 1992). Self-critical sequence training (SCST) is a method proposed by Rennie et al. (2017) and Luo (2020) to stabilize the large variation of incentives by normalizing rewards.

Text similarity between produced and reference captions is the de facto standard reward function for captioning, as seen in Fig. 2. According to recent research, reference-trained captioning machines frequently overlook vital information from photos (Dai et al., 2017; Wang et al., 2017). Lee et al. (2019) utilize the accuracy of a visual question answering model as a motivator for models to create captions with enough information to answer a visual query. Dai and Lin (2017), Luo et al. (2018), and Liu et al. (2018) employ the self-retrieval score of image-text retrieval models as a reward, combining it with n-gram metrics to encourage captioning models to develop captions that are unique to each input picture.

For steady training, keep in mind that these works necessitate a careful balance between self-retrieval and text similarity aims. Our technique, on the other hand, eliminates the need for reference caption and text similarity metrics for reward computation with CLIP text encoder fine-tuning (Sec. 3.2).

III. METHODS

3.1: Image Conferencing Using CLIP

To lead an image captioning model, we suggest employing the CLIP (Radford et al., 2021) image-text similarity score. We utilize CLIP-S as our incentive, as recommended by Hessel et al. (2019):

$$CLIP - S(I, C) = w * \max(f^I I^T(c)) / |f^I| |f^T|$$

where I, c are the image and caption encoders, and $w = 2.5$ are the CLIP image and text encoders. Picture captioning models are encouraged to provide captions that contain more unique information about the input image by learning to optimize the contrastive model's image-text similarity. This training technique is depicted in Fig. 1 (a).


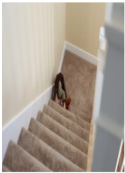
Image	Criteria	Annotations
(a) 	Background	white house, truck digging soil in front of the house, trees and bushes, house surrounded by a small garden, Mini excavator, houses, white and grey building, greenery, two houses, blue and white colored machine
	Object	a blue car, a blue car, black car, car, dozer, white and grey building, greenery, black car, green bushes
	Relation	parked in the front yard, in front, parked in front of, Parked, car standing on the road
	Overall	A blue car parked in the front yard of an off white house with a truck digging soil in front of the house. A blue car in front of a house surrounded by a small garden with trees and bushes in the background. A black car parked in front of a house with a mini excavator behind it with other houses in the background. A car and a dozer parked in front of two white and grey buildings and greenery on both sides. A black car standing on the road surrounded by green bushes on both sides and two houses and a blue and white colored machine in the background.
	Background	velvet carpet stairs, light-brown colored stairs, Off white wall, Cream painted walls, cream wall with straight line light
(b) 	Object	brown jumpsuit, kid, Toy, black jumpsuit, boy, brown clothes, toy, brown carpet, Little young boy, cotton carpeted stair, dark brown jumper dress, cream wall
	Relation	with its head on to, touching, Hiding, Holding, boy holding and playing with the toy, putting, wearing
	Overall	A child wearing a brown jumpsuit with its head on to the velvet carpet stairs. A kid is touching their head on a light brown colored stairs. A Kid wearing a black jumpsuit and holding a toy hiding below the stairs with off white wall in the background. A boy wearing brown clothes holding and playing with his toy and playing on a brown carpet on stairs with cream painted walls. Little young boy is putting his forehead on the cotton carpeted stair wearing dark brown jumper dress and background of cream wall with straight line light.
	Background	velvet carpet stairs, light-brown colored stairs, Off white wall, Cream painted walls, cream wall with straight line light
	Object	brown jumpsuit, kid, Toy, black jumpsuit, boy, brown clothes, toy, brown carpet, Little young boy, cotton carpeted stair, dark brown jumper dress, cream wall

Fig No. 2 : Examples of FineCapEval. We combine the comments for each criterion from five separate human annotators for each image. For ‘overall’ criteria, we assess captions using CIDEr. We use word-level recall Rerword to evaluate captions for the remaining criteria.

Following Rennie et al. (2017), we use REINFORCE (Williams, 1992) to optimize our captioning model $P(c|I)$ using a self-critical baseline. The gradient of expected reward for the produced caption c is approximated by normalizing the reward of the beam search with the baseline reward b from greedy decoding.

3.2: Fine-tuning the CLIP Text Encoder to Improve the Image

The captioning model trained with the CLIP-S reward frequently creates grammatically erroneous captions since CLIP is not trained with a language modelling purpose (e.g., repeated words; see Table 3). Negative captions, created by randomly repeating/removing/inserting/swapping/shuffling tokens from the reference captions, are used to infuse grammatical information into the CLIP text encoder. In the appendix, we offer implementation details for such procedures. A grammar head, a two-layer perceptron, is introduced, which accepts the CLIP text feature $f^T(c)$ as input and outputs the probability that c is grammatically correct: $g(c) [0, 1]$.

For the grammar aim, we employ binary cross-entropy, with the label y being 1 for reference captions and 0 for negative captions: $y \log g(c)$. With the sum of the original CLIP objective and the grammars objective, we fine-tune the text encoder and grammar head together. During fine-tuning, the CLIP image encoder settings are fixed. Figure 1 depicts the fine-tuning procedure (b). We train captioning models with the reward enriched with the

grammar score after fine-tuning CLIP: $CLIP-S(I, c) + g(c) = R(I, c)$, where $g(c) = 2.0$

IV. FINECAPEVAL IS A DATA COLLECTION FOR EVALUATING FINE-GRAINED CAPTIONS.

FineCapEval is a new dataset for evaluating captions in four separate characteristics. We use 500 photos from the MS COCO (Lin et al., 2014) test2015 split and the Conceptual Caption (Sharma et al., 2018) val split to build FineCapEval. Then we ask five human annotators to write words about 1) the backdrop, 2) the items (and their qualities; i.e., color, form, etc.), the relation between objects (i.e., spatial relation), and 4) a descriptive caption that incorporates all three aspects for each image. The data gathering procedure is detailed in the appendix. For each of the four criteria, FineCapEval contains 1,000 photos with 5,000 annotations. Samples from the FineCapEval dataset are shown in Table 1.

V. EXPERIMENT

MLE, CIDEr, CLIP-S, CIDER+CLIP-S, and CLIPS+Grammar are the reward setups we compare. We undertake tests on the MS COCO (Lin et al., 2014) English captioning dataset with Karpathy split based on earlier work (Karpathy and Fei-Fei, 2015). We use n-gram-based metrics, embedding-based metrics, text-to-image retrieval scores, and FineCapEval to evaluate the model. We also conduct a five-criteria human evaluation to determine the human preference for the generated captions in many aspects.

Reward	N-Gram Based				Embed Based		
	Image Based				Image - Text Based		
	BLEU - 4	CIDEr	METEOR	ROUGE-L	BERT-S	CLIP-S	RefCLIP-S
MLE	33.9	112.3	27.21	28.9	56.32	0.921	1.13
CIDEr	41.52	125.9	28.9	28.25	59.21	0.9521	1.14
CLIP-S	7.2	13.25	18.34	19.32	31.6	0.8921	1.21
CIDEr + CLIP-S	35.2	125.32	29.91	29.9	59.21	0.9432	1.15
CLIP-S + Grammar	17.9	72.3	24.8	28.21	48.21	0.9321	1.16

Table 1 : Over-all Evaluation matrix

Training and Model Architecture, our captioning model architecture is CLIP-Res50 Transformer (Shen et al., 2019). CLIP-Res50 is used to extract visual features, and a transformer encoder-decoder (Vaswani et al., 2017) is used to simulate conditional language. To extract 2048-dimensional visual information, we scale photos to 224x224. A 6-layer encoder and 6-layer decoder make up the transformer. With 8 V100 GPUs, we train our model with the MLE goal for 15 epochs and then with variable incentives for another 25 epochs (total 40 epochs). For beam search decoding, we utilize beam size 5. We use PyTorch (Paszke et al., 2017), PyTorch Lightning3, and Hugging Face Transformers to build a training pipeline (Wolf et al., 2019).

Metrics based on N-grams, BLEU-4 (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), METEOR (Banerjee and Lavie, 2005), and ROUGE-L are among the genes we describe (Lin, 2004).

Metrics based on embedding, we provide BERT-S (Zhang et al., 2019) and CLIP-S/RefCLIP-S (Hessel et al., 2021) as well as CLIP-S/RefCLIP-S (Hessel et al., 2021). 4 BERT-S compares reference captions to generated captions in terms of textual similarity, CLIP-S compares input pictures to generated captions in terms of image-text similarity, and RefCLIP-S averages textual (with reference captions) and image-text similarity.

Text-to-Image Retrieval is a technique for retrieving images from text, to assess the uniqueness of the produced captions, we report the

recall of the reference picture using a text-to-image retrieval model. CLIP ViT-B/32 is used as the retrieval model (Radford et al., 2019).

FineCapEval, we use word-level recall, reword [0, 1], to evaluate caption performance for background, object, and relation criteria. Details of the Reword computation may be found in the appendix. CIDEr is used to assess overall caption performance.

Evaluation by humans, we display a pair of captions from CLIP-S+grammar reward (ours) with CIDEr reward and with MLE baseline to human annotators from Amazon Mechanical Turk5 to assess the captions in terms of human preference. Then, based on five factors, we ask them to choose a better caption (overall, background, object, attribute, relation). We ask 10 annotators 50 paired selection questions for each of the five criteria. For caption generation, we use 50 pictures from FineCapEval.

VI. RESULT

6.1 : Distinctive Captions are aided by CLIP.

Table 2 shows that models with CLIP-S and CLIPS+Grammar rewards outperform baselines in terms of image-text metrics (CLIP-S / RefCLIP-S) and text-to-image retrieval. Their retrieval scores are, interestingly, even higher than the retrieval score with reference captions. This demonstrates the uniqueness of their captions created by them. For picture (a) in Table 3, our CLIP-S+Grammar reward model classifies the rainy weather as 'wet,' but the CIDEr reward model does not.

Reward	Text - to - Image Retrieval		
Reference Caption	30.1*	55.21*	65.21*
MLE	22.08	46.21	59.21
CIDEr	21.92	46.21	59.01
CLIP-S	42.51	72.21	82.36
CIDEr + CLIP-S	24.51	50.21	63.25
CLIP-S + Grammar	39.21	64.52	75.92

Table No. 2 : Text to Image Retrieval

Text similarity metrics (n-gram metrics and BERT-S) are lower in our CLIP-S and CLIP-S+Grammar models than in the CIDEr model. However, models with CLIP-S and CLIP-S+Grammar rewards frequently output captions that include fine-grained information that is not available in the reference captions, which can help to solve the poor scores on these reference-based metrics. The CLIP-S+Grammar model describes the restaurant's "blue sign" in picture (b) in Table 3, although none of the reference captions do.

6.2 : CLIP Text Encoder is being fine-tuned to improve grammar.

Table 3 illustrates that adding the grammar incentive (CLIPS+Grammar) successfully mitigates the CLIP-S award's degeneration (e.g., word repetition). Table 2 indicates that including grammar reward improves all text similarity measures (e.g., CIDEr by +60).



Image	Reward	Captions
	CIDEr	a window of an airport with planes on the runway
	CLIP-S	several rows of planes parked outside a terminal window area with fog outside a terminal window motion position area motionn
	CLIP-S+ Grammar	a lot of airplanes parked on a wet airport terminal
	Reference Captions	An airport filled with planes sitting on tarmacs.
		The view of runway from behind the windows of airport.
		a truck driving towards some planes parked on the runway
	CIDEr	a group of people riding bikes down a city street
	CLIP-S	several cyclists moving and bicycles near a restaurant and a blue advertisement outside a red brick building motion stance p
	CLIP-S+ Grammar	a group of people riding their bikes on the busy street with a blue sign
	Reference Captions	people on bicycles ride down a busy street
		A group of people are riding bikes down the street in a bike lane
		bike riders passing Burger King in city street
A group of bicyclists are riding in the bike lane.		
		Bicyclists on a city street, most not using the bike lane

Table 3 : On MS COCO Karpathy test split pictures, captions were generated by models with varying incentives.

Criteria	CLIP-S + Grammar	Win	Lose	Tie
Overall	v.s. MLE	50	41.91	8.9
	v.s. CIDEr	51.21	31.25	19.65
Background	v.s. MLE	53.8	36.25	13.54
	v.s. CIDEr	53.25	26.31	21.56
Object	v.s. MLE	53.11	36.65	12.48
	v.s. CIDEr	56.21	33.21	12.39
Attribute	v.s. MLE	58.21	38.71	7.21
	v.s. CIDEr	56.21	39.21	7.58
Relation	v.s. MLE	45.21	44.56	12.34
	v.s. CIDEr	49.56	38.91	12.36

Table 4 : Values of CLIP Text Encoder is being fine-tuned to improve grammar.

6.3: Caption Evaluation on a Finer Scale

FineCapEvaluation, Table 2's four right columns demonstrate that CLIP-S and CLIP-S+Grammar outperform CIDEr on all four FineCapEval criteria: overall, background, object,

and relation. The object criteria has the lowest gap, implying that MS COCO reference captions explain more object information than background or object relationships.

Human evaluation findings are shown in Table 4 for five criteria: overall, backdrop, object, attribute, and relation. Using 50 pictures from the Conceptual caption (Sharma et al., 2018) val split, we select 50 captions from a model trained with CLIP-S+grammar reward (ours), CIDEr reward, and MLE baseline. We invite ten human annotators to choose a better caption between ours and another approach for each of the five criteria. Across the board.

VII. CONCLUSION

By optimizing CLIP's multimodal similarity score and fine-tuning its text encoder to enhance grammar, we present a unique training technique for picture captioning models. CLIP reward removes the need for reference captions and associated bias in reward calculation. FineCapEval, a dataset for fine-grained caption assessment, is also introduced. We show the usefulness of our proposed technique using qualitative examples and improvements in text-to-image retrieval, FineCapEval, and human assessment on fine-grained criteria. Future research will focus on fine-tuning CLIP reward models with desired writing styles for various applications, as well as enhancing the synthetic augmentation process by including external data suited for grammars with advanced linguistics competence.

Millions of online image-text pairings were utilised to train the CLIP models we deployed. According to Birhane et al. (2021), large-scale datasets frequently contain explicit and problematic image-text combinations. The use of CLIP reward to train image captioning models is meant as a research output, as indicated by the CLIP model card⁶, and any actual use case of the models is out of scope.

Because our captioning and CLIP models were trained on English datasets, they should only be used in English-language scenarios. Future study will investigate the extensions in many languages, as our suggested approach is not confined to English and may simply be extended to other languages.

REFERENCE

- [1]. Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In ECCV.
- [2]. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and LeiZhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In CVPR.
- [3]. Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In ACL Workshop
- [4]. Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. In NIPS, pages 1–9.
- [5]. Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes.
- [6]. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, Jong Wook, Kim Chris, Hallacy Aditya, Ramesh Gabriel, Goh Sandhini, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2018. Learning Transferable Visual Models From Natural Language Supervision. In ICML.
- [7]. Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chana, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In NIPS Workshop.
- [8]. Kishore Papineni, Salim Roukos, Todd Ward, and Wj Wei-jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In ACL
- [9]. Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence Level Training with Recurrent Neural Networks. In ICLR, pages 1–15.
- [10]. Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2016. How Much Can CLIP Benefit Vision-and-Language Tasks? In ICLR.
- [11]. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In NIPS.