Improving cross-modal attention via object detection

Yongil Kim
Seoul National University
miles94@snu.ac.kr

Yerin Hwang
Seoul National University
dpfls589@snu.ac.kr

Seunghyun Yoon
Adobe Research
syoon@adobe.com

Hyeongu Yun
Seoul National University
youaredead@snu.ac.kr

Kyomin Jung
Seoul National University
kjung@snu.ac.kr

Abstract

Cross-modal attention is widely used in multimodal learning to fuse information from two modalities. However, most existing models only assimilate cross-modal attention indirectly by relying on end-to-end learning and do not directly improve the attention mechanisms. In this paper, we propose a methodology for directly enhancing cross-modal attention by utilizing object-detection models for vision-and-language tasks that deal with image and text information. We used the mask of the detected objects obtained by the detection model as a pseudo label, and we added a loss between the attention map of the multimodal learning model and the pseudo label. The proposed methodology drastically improves the performance of the baseline model across all performance metrics in various popular datasets for the image-captioning task. Moreover, our highly scalable methodology can be applied to any multimodal task in terms of vision-and-language.

1 Introduction

Several studies have been conducted on multimodal learning models that deal with various modalities having different characteristics, such as image, text, and speech information. Vision-and-Language, which includes VisualQA[1] and Image Captioning[2], is one of the most extensively studied fields that achieves the desired output by receiving image and text information as input. Similar to other multimodal learning methods, a model that mixes the two modalities well should be designed for producing the desired representation.

Most multimodal models that display the best performance use attention mechanisms to mix two or more modalities. Attention techniques help the model learn the most important parts on its own and share relevant details between the two modalities. Unlike attention applied in one modality, such as self-attention, cross-modal attention is applied between two modalities. In particular, cross-modal attention has become essential in multimodal models because transformer models based on attention techniques exhibit strength in multimodal learning tasks. Additionally, many studies have displayed excellent performances using cross-modal attention in vision-and-language research.

However, most existing vision-and-language models only leverage the cross-modal attention module, and there have been no attempts to improve it. In other words, no study has directly trained a cross-modal attention module using an objective function. In this paper, we propose a methodology that directly improves cross-modal attention in vision-and-language research, thus increasing the model's overall performance.

Specifically, we propose a method to improve attention mechanisms using object-detection models. Recently, powerful object-detection models, such as detectron2[3], have emerged that show

36th Conference on Neural Information Processing Systems (NeurIPS 2022), New Orleans, Louisiana.

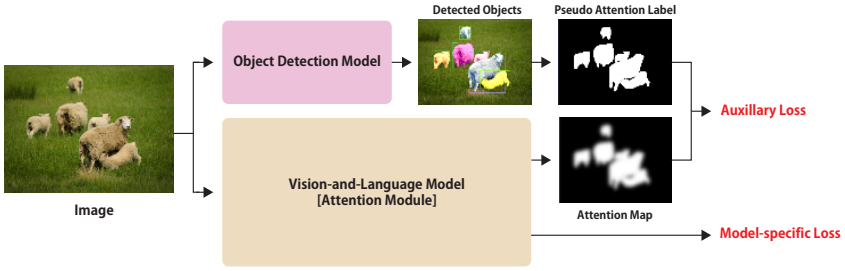

Figure 1: Overall model architecture. We propose a task-agnostic methodology to make direct improvements to the attention module using the object detection model in the vision-and-language.

performances similar to or even exceeding those of humans. We leveraged the ability of these object-detection models to improve cross-modal attention through pseudo-self-supervised learning. As shown in Figure 1, the object mask obtained by passing the input image through the object-detection model was used as a pseudo label, and an auxiliary loss was configured such that the attention map output by the cross-modal attention module of the vision-and-language model was similar to the pseudo label.

Upon performing experiments on multiple datasets by applying our methodology to an image-captioning task, we observed a significant improvement in performance compared to the existing baseline model across all metrics with all datasets. The proposed methodology is a new training scheme that seeks to improve the performance of cross-modal attention. Furthermore, we highlight that our highly scalable methodology can be applied task-agnostically to all vision-and-language tasks.

2 Related works

2.1 Cross-modal attention

Multimodal learning models that handle two modalities with different characteristics in one model require a process of fusing two modality embeddings. Ever since the attention technique, which allows models to focus on the important parts and derive information on their own, has been devised, studies have focused on mixing information and deriving performance using cross-modal attention for different modalities.

Like other uni-modal models, many multimodal studies barely utilize the attention technique, with only few existing studies on improving the attention effect. Recently, some studies have been conducted to enhance the effectiveness of cross-modal attention beyond simple utilization. For example, recent studies [4, 5] have attempted to improve the efficacy of attention between two modalities by reducing the difference in characteristics between the two modalities.

However, to date, few studies exist on the direct improvement of attention because in most existing models, the attention technique is assimilated through end-to-end learning, making it challenging to design a direct objective function. In this paper, we propose a direct improvement method for cross-modal attention via designing objective functions using an object-detection model. .

2.2 Surrogate IOU

The Invention-of-Union (IOU) metric is used as an indicator of detection performance based on the degree of overlap between the output of the object-detection model and the actual label. It is widely used for model evaluation in semantic segmentation, instance segmentation, and object-detection tasks. Additionally, attempts have been made to use it as a direct objective function to improve the results of the IOU metric. However, because the IOU metric is indifferentiable, it is not available in most gradient-based frameworks. Therefore, in several studies [6, 7], the IOU score was directly used for learning through a differentiable surrogate function.

In this study, we used the surrogate IOU loss function to learn the cross-modal attention of the vision-and-language model. While previous studies have focused on improving the performance of

the object-detection model itself, this study differs by using a surrogate IOU to design an objective function between the results of the attention module and the object-detection model.

3 Methods

3.1 Proposed Methodology for improving cross-modal attention

Algorithm 1 Improving cross-modal attention

Input : image, sequence, **ODM**(Object Detection Model), **VLM**(Vision-and-Language Model)
Output : *training_Loss*
objects, *masks* ← **ODM**(image)
attentions, model-specific Loss ← **VLM**(image, sequence)
  for object in objects **do**
    if object is in sequence **then**
      merge $masks[object]$
      $ICA\_Loss[object] =$ **Surrogate_IOU**$(masks[object], attentions[object])$
    end if
  end for
average *ICA_Loss*
training_Loss = *model-specific_Loss* + $\lambda_{Aux} * ICA\_Loss$
return *training_Loss*

We propose a training scheme that can be applied to vision-and-language tasks that deal with image and text information, regardless of the task. As shown in Figure 1, the proposed methodology utilizes an object-detection model. Many approaches use the ability of an object-detection model in the vision-and-language tasks [8, 9], but, to the best of our knowledge, our study is the first to introduce and use it as a factor for directly improving the attention module.

Algorithm 1 provides our novel training scheme for improving cross-modal attention. First, the object-detection model extracts object instances from the input image and corresponding masks. Then, among the labels of the objects detected, we select the one that matches the word token in the input sequence. Additionally, the masks of the instances corresponding to these critical objects are merged. For example, in Figure 1, sheep are found as multiple instances by the object-detection model, and as "sheep" is included in the caption, the masks of the instances corresponding to the sheep are merged. The merged mask is then used as a pseudo-ground-truth label for the attention map corresponding to the word token. Because of the bias in the detection model, it is called a pseudo-ground-truth label. However, as the performance of recent object-detection models exceeds human capabilities, they can be considered as approaching human-annotated ground-truth labels.

Finally, the surrogate IOU is measured with the attention map output by the attention module in the vision-and-language model and the pseudo-attention label and is used as an auxiliary loss. For the surrogate IOU loss, the Lovász hinge extension of Jaccard loss [6] is used. The Lovász hinge extension of Jaccard loss converts the Jaccard index (also called the IOU score) into a differentiable loss function for use in a continuous optimization framework. For a segmentation output $\tilde{y}$ and ground truth y^*, mispredicted pixels for class c can be expressed as the following formula.

$$M_c(y^*, \tilde{y}) = \{y^* = c, \tilde{y} \neq c\} \cup \{y^* \neq c, \tilde{y} = c\}$$

Subsequently, Jaccard loss is rewritten as shown below, and this Jaccard loss is employed with empirical risk minimization.

$$\Delta J_C : M_c \in \{0,1\}^p \mapsto \frac{|M_c|}{|\{y^* = c\} \cup M_c|}$$

However, because the Jaccard loss is indiiferentiable, it is unable to directly apply it to the gradient descent framework. Previous studies [5, 10] proved that this Jaccard loss has submodular character-

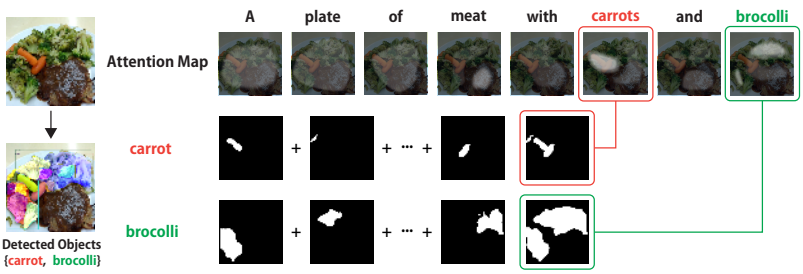

Figure 2: The proposed methodology applied to the image captioning task. When the word token of the text sequence matches the detected object label, the masks of the object instances are merged. After that, an auxiliary loss is constructed between the merged mask(middle and bottom) and the attention map(top) of the word token using the surrogate IOU.

istics and that the convex closure of the submodular set is tight and computable in polynomial time. The convex closure of the submodular set function is a Lovász extension, which is defined as follows.

Definition 1 [[6], Def.2]. *The* Lovász extension *of a set function* $\Delta : 0, 1^p \mapsto \mathbb{R}$ *such that* $\Delta(0) = 0$ *is dfeined by*

$$\overline{\Delta} : m \in \mathbb{R}^p \mapsto \sum_{i=1}^{p} m_i g_i(m)$$

with $g_i(m) = \Delta(\{\pi_1, ..., \pi_i\}) - \Delta(\{\pi_1, ..., \pi_{i-1}\})$, π *being a permutation ordering the components of the* m *in decreasing order, i.e.* $x_{\pi 1} \geq x_{\pi 2} \cdots \geq x_{\pi p}$.

Finally, this surrogate IOU is composed of the following Lovász hinge applied to the Jaccard loss by applying the Lovász extension to the hinge loss $m(F)$ for the output score F.

$$ICA\_loss = \overline{\Delta_{J_c}}((m(F))$$

This ICA(Improving Cross-modal Attention) loss was reflected in the objective function of the original model additionally. For our setting, output score F is attention map feature rather than segmentation output. The "*model-specific Loss*," the original objective function, is different for each task of vision-and-language. Because the proposed method is a task-agnostic methodology that can be applied regardless of the task, it can be used for any loss through additional auxiliary loss.

3.2 Application to Image Captioning task

The proposed methodology was applied to image captioning, a vision-and-language task. The model used for image captioning is shown in Figure 2. The Mask R-CNN [11], used as the object-detection model, receives an image as the input and outputs the class label and the mask of the detected object. We also checked whether the class label of the detected object given in natural language exists in the text sequence of the vision-and-language model. In the case of image captioning, if the object detected by Mask R-CNN was included in the word tokens of the caption given as input, the object was considered to be critical and the mask was extracted. This was used as a pseudo-attention label.

Moreover, as shown in Figure 2, the image was passed through the attention module of the model, and an attention map corresponding to each word token was generated. Subsequently, the attention map corresponding to the word token of the caption and the critical object word token corresponding to the class label predicted by Mask R-CNN were extracted. Finally, the auxiliary loss was configured as a surrogate IOU using this attention map and the pseudo-attention label previously output by Mask R-CNN.

4 Experiment

We conducted an experiment using image captioning in a vision-and-language task. The show-attend-tell [12] model was used as the baseline. This baseline model utilizes the attention module and attention with an image corresponding to each word token of the caption. For the dataset,

Table 1: Experiment Result on image captioning.

Dataset	Model	BLEU-1	BLEU-4	CIDEr	METEOR	ROUGE-L	SPICE
Flickr8k	**Show-Attend-Tell*	0.6460	0.2339	0.5874	0.2032	0.5306	0.1495
	+ICA Loss	**0.6576**	**0.2453**	**0.5939**	**0.2256**	**0.5479**	**0.1713**
Flickr30k	**Show-Attend-Tell*	0.6557	0.2334	0.4789	0.1930	0.5085	0.1330
	+ICA Loss	**0.6597**	**0.2351**	**0.4798**	**0.1963**	**0.5132**	**0.1351**
MSCOCO	**Show-Attend-Tell*	0.7050	0.3036	0.9378	0.2472	0.5684	0.1776
	+ICA Loss	**0.7110**	**0.3149**	**0.9392**	**0.2488**	**0.5792**	**0.1793**

*our Implementation

we used the most commonly used datasets for image captioning: Flickr8k[13], Flickr30k[14], and MSCOCO[15]. Additionally, the following natural language generation performance indicators were used: BLEU-1, BLEU-4[16], CIDEr[17], METEOR[18], ROUGE-L[19], and SPICE[20]. Both the existing baseline model and the proposed model use the directly implemented model, and the official performance index implemented in PyCoCo was used as a performance indicator. Among the class labels discovered by Mask R-CNN during the experiment, the average number of critical objects belonging to the caption was 2.3 per instance, and the loss was calculated by taking the mean average loss of the corresponding essential objects; the λ_{Aux} value was set to 0.1.

The experimental results are listed in Table 1. The ICA Loss improves the cross-modal attention loss, which is the auxiliary loss added in this study. The model to which the proposed auxiliary loss is added shows much better performance across all datasets and performance indicators. In particular, when ICA Loss is added to the Flickr8K and MSCOCO datasets, performance indicators such as BLEU-4 and ROUGE-L are markedly improved. Thus, the proposed method for directly improving attention has an apparent positive effect on model performance improvement.

5 Broader Impact

As mentioned earlier, the proposed methodology can be applied to all vision-and-language studies in a model-agnostic manner. The mask output by the object-detection model and the class label generated in the corresponding natural language can be used in vision-and-language tasks. For example, in addition to the image captioning used in this study, it can be applied to visual question answering tasks to help find a critical object in a question and directly improve the corresponding cross-modal attention. Further, in the natural language for a visual reasoning task, if there is an object with overlapping prediction classes in the object-detection model for two images in the natural language, whether these objects play an important role can be determined.

6 Conclusion

In this paper, we proposed a method to improve performance by borrowing an object-detection model from the existing vision-and-language research. We presented a methodology for enhancing attention through direct learning by applying a differentiable surrogate IOU objective function to the output mask of the object-detection model and the attention map of the model. The proposed method can be extended to any task in the future, and its feasibility was verified experimentally by the significant improvement in performance for the image-captioning task.

7 Acknowledgements

We thank anonymous reviewers for their constructive and insightful comments. K. Jung is with ASRI, Seoul National University, Korea. This work was supported by Sumsung Electronics and the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2022. This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)]. This work was partially funded by gifts from Adobe Research.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[3] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[4] Yongil Kim, Hyungu Yun, Seungpil Won, and Kyomin Jung. Improving video-qa system through modality alignment. In *Proceedings of the Korea Computer Congress*, pages 570–572, 2021.

[5] Hyeongu Yun, Yongil Kim, and Kyomin Jung. Modality alignment between deep representations for effective video-and-language learning. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2759–2770, 2022.

[6] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4413–4421, 2018.

[7] Gattigorla Nagendar, Digvijay Singh, Vineeth N Balasubramanian, and CV Jawahar. Neuro-iou: Learning a surrogate loss for semantic segmentation. In *BMVC*, page 278, 2018.

[8] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *Advances in Neural Information Processing Systems*, 32, 2019.

[9] Zhongliang Yang, Yu-Jin Zhang, Yongfeng Huang, et al. Image captioning with object detection and localization. In *International conference on image and graphics*, pages 109–118. Springer, 2017.

[10] László Lovász. Submodular functions and convexity. In *Mathematical programming the state of the art*, pages 235–257. Springer, 1983.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[12] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

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

[14] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[17] 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*, pages 4566–4575, 2015.

[18] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[19] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, 2004.

[20] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016.