Probing the scene information in different layers of Vision and Language models

Anonymous ACL submission

1 Introduction

Scene recognition is a high level computer vision task which aims to determine the scene category(e.g., classroom, bus station, tennis court) through understanding the global property of an image(Zhang et al., 2022). Li et al. (2021) purpose a DNN which exploit the linguistic information in indoor place category recognition by fusing the semantic features in the caption and the region features in the image and outperforms previous methods on MIT-67 datase(Quattoni and Torralba, 2009). Combined the fact that the vision features and the language features are interacted with each other via the attention mechanism in the pre-training procedure of vision and language models and VL models are pre-trained with large scale pre-training data, it is natural to raise the questions: does the VL models have the ability to recognize the scene through the pre-training procedure?

Few researches in exploring the ability of the VL models in scene understanding. Cafagna et al. (2021) shows VisualBert(Li et al., 2019), LXMERT(Tan and Bansal, 2019) and CLIP(Radford et al., 2021) are capable of recognizing scenes and scene information are encoded in the representation of CILP, however the answer of whether the scene information is also encoded in the representations of transformer based VL models is still missing. Besides, it is hard to draw a conclusion of why some VL models outperform other VL models in the scene recognition task as different VL models are pre-trained in different settings including pre-training datatset and pre-training tasks, there are many possible compounds while comparing the performance of different VL models in the particular task. Thirdly, inspired by the observation that the degree of visual modality attends with the linguistic modality in higher layers is deeper than in the lower layers of the single stream model, and the degree of visual modality attends with the linguistic modality is more shallow in the higher layers than in the lower layers in the dual stream model(Cao et al., 2020), we are curious that whether the VL models can learn more scene information in the deeper layers than in the lower layers as there are more interventation between two modalities.

To tackle aforementioned problems, we apply VL models in the VOLTA framework(Bugliarello et al., 2021) in the scene understanding task, which unifies all setting of VL models exclude the way VL models embed linguistic features and visual features. We purpose a probing task to know whether there are scene information encoded in the embedding of the VL models and probe each layer of the VL models to see whether the deeper layers contains more scene information than the lower layers.

We make three primary contributions in this work:
(1) We observe that single stream models outperform dual stream models in the VOLTA framework in the task of scene recognition. Our results show that the difference between the performances of VL models in scene understanding mainly attribute to the initial embedding of each VL model.
(2) We observe that descriptive captions are preferred by VL models to describe the image than using scene captions.

2 Experiment design and Results

2.1 The simpler, the better: Single stream VL models outperform dual stream VL models in scene understanding

In order to know whether the VL models already have the ability to recognize a scene, we feed the scene captions in the HL1K test set and the re-

lated images into the VL model and use the image sentence alignment head of the VL model to classify the relationship between the input caption and the image. In the processed test set of the HL1K dataset, there are 50% the caption and the image are a match and there are 50% the caption is randomly chosen and is not a match with the image. The number of positive samples in the test dataset is 741 and the number of negative samples in the test dataset is 758. We conduct the same experiment but use the descriptive captions in the COCO dataset(Lin et al., 2015) and the corresponding images as the input to the VL models, aiming to compare the ability of the VL models to understand the scene of the given image and the ability of the VL models to understand the detailed semantics of the image. We apply the checkpoints of UNITER(Chen et al., 2020), VisualBERT, VL-BERT(Su et al., 2020) and LXMERT in the control set up of the VOLTA framework in aligning scene captions with the image and aligning descriptive captions with the image respectively. Table 1 shows the accuracy of LXMERT, VL-BERT, UNITER and VisualBERT in the VOLTA framework aligning the scene captions and descriptive captions with the image.

Model	**HL1K(%)**	**COCO(%)**
LXMERT	73.18	90.53
VL-BERT	79.19	95.13
UNITER	81.12	95.26
VisualBERT	81.12	96.93

Table 1: The accuracy of LXMERT, VL-BERT, UNITER and VisualBERT in the VOLTA framework aligning the scene captions and descriptive captions with the image.

LXMERT, VL-BERT, UNITER and VisualBERT can align images with descriptive captions accurately, the accuracy of all VL models aligning images with descriptive captions is above 90%. The accuracy of aligning images with scene captions in HL1K dataset is lower than the accuracy of aligning images with descriptive captions in COCO dataset for all VL models in our experiment. We believe there are two reasons to explain. The first reason is the VL model can align words in the input sentence to the related regions in the image implicitly through the pre-training procedure(Li et al., 2019),which might be helpful to align descriptive captions and images. Because descriptive captions in COCO dataset mainly describe an image by mentioning objects or people, attributes and relationships between objects or people in the image, VL models can align descriptive captions and image by aligning object tokens in the descriptive caption and the related region of interest in the image. We hypothesis that the tokens of attributes and relationships in the caption can be aligned with the corresponding regions of the image as this trend can be shown in the visualization of attention weights of some selected heads in VisualBERT(Li et al., 2019), which can also benefit VL models to recognize an image by detailed semantics. However, this pattern cannot be duplicated in aligning scene captions and images as scene captions do not contain any objects and their attributes and relationships in the image, scene captions are more abstract than descriptive captions, therefore VL models cannot align scene captions and images by aligning word tokens in the caption and the related regions of interested in the image. The other reason is all VL models in the VOLTA framework are pre-trained with the Conceptual Caption dataset. The Conceptual Captions dataset has 3.3M image text pair examples and the captions are created by harvesting and filtering Alt-text from HTML pages(Sharma et al., 2018). After text filtering, the captions also use objects and their relationships and attributes to describe the image, which is similar to the way COCO captions to describe the image, combining the enormous amount of training data therefore makes the performance of VL models aligning descriptive captions with the image better than the performance of aligning scene captions with the image. We also notice that a portion of the Conceptual Captions contain scene information, which is a stylistic difference with COCO captions and might contribute to the performance of the VL models aligning scene captions with the images.

The accuracy of VL-BERT, UNITER and VisualBERT aligning scene captions and descriptive captions with the image are higher than the accuracy of LXMERT aligning scene captions and descriptive captions with the image, this can demonstrate that single stream models can perform better than dual stream models in the scene understanding and understanding the detailed semantics of the image in the VOLTA framework. In the performances of all VL models in our experiment, VisualBERT have the highest accuracy in both aligning scene captions and descriptive captions with the image while LXMERT has the worst performance in both

aligning scene captions and descriptive captions with images. We think the differences between the performances of the single stream models in our experiment attribute to the differences in the initial embedding of the single stream models as the only variable of the single stream models in the controlled set up of the VOLTA framework is the way VL models embed the input visual features and linguistic features. We think the position embedding of visual features is not important in the scene recognition task as UNITER and VisualBERT which have the highest accuracy in understanding scenes do not use the position embedding when the visual features are embedded. Besides, we hypothesis that summarizing the visual features extracted on the whole input image and the token embedding if the input feature is a linguistic element and summarizing the [IMG] token and the visual features if the input feature is a visual element might not benefit scene recognition as VL-BERT uses this way to construct the token embedding and the visual feature embedding but performs the worst compared to the performances of scene understanding of VisualBERT and UNITER.

Table 2 2-5 5 show the precision, recall and F1-score of LXMERT, UNITER, VL-BERT and VisualBERT to classify the input caption and the image as a match or not a match. We notice for the input caption in both HL1K dataset and COCO dataset, the F1 score of classifying the caption is not a match with the image is higher than or equal to the F1 score of classifying the caption is a match with the image for all VL models. In the performances of all VL models to align scene captions in the HL1K dataset with the image, the recall of classifying the caption and the image is a match is lower than the precision of classifying the caption and the image is a match, while the precision of classifying the caption and the image is not a match is lower than the recall of classifying the caption and the image is not a match. This indicates VL models are more easily to misclassify the image and the caption is not a match when the ground truth is they are a match compared to misclassify the image and the caption is a match when the ground truth is they are not a match.

Caption dataset	Label class	P	R	F1
HL1K	match	89	52	66
	not match	67	93	78
COCO	match	94	86	90
	not match	88	95	91

Table 2: the precision, recall and F1 score of LXMERT on the task of aligning the scene captions and the descriptive captions with the image. Label class "match" represents the model classify the caption and the image is a match, label class "not match" means the model classify the caption and the image is not a match, the following tables use the same setting.

Caption dataset	Label	P	R	F1
HL1K	match	86	74	79
	not match	77	88	83
COCO	match	96	94	95
	not match	95	96	95

Table 3: the precision, recall and F1 score of UNITER on the task of aligning the scene captions and the descriptive captions with the image.

2.2 Detailed vs Abstract: Descriptive captions are preferred by VL models compared to scene captions

We are also interested in knowing whether the VL models prefer to use the scene caption or the descriptive caption to describe a given image. The most intuitive way to achieve this goal is to use the scene caption, the descriptive caption and the image as the input and compare the scores generated by the image sentence alignment head of the VL model, if the scene caption is assigned higher score to align with the image than the descriptive caption, then the scene caption is regarded "preferred" to describe the image than the descriptive caption by the VL model, if the descriptive caption is assigned higher score to aligned with the image, then the descriptive caption is regarded as "preferred" to describe the image than the scene caption. However, we can not directly adapt this design in our experiment as the image sentence alignment head in the VOLTA framework is trained to classify whether the input single caption is aligned with the image, therefore if we use the the scene caption and the descriptive caption and the image as the input to the VL model, the image sentence alignment head of the VL model can not generate the score of aligning the scene caption with the image and the score of aligning the descriptive caption respectively.

Caption dataset	Label class	P	R	F1
HL1K	match	88	67	76
	not match	74	91	82
COCO	match	95	95	95
	not match	95	95	95

Table 4: the precision, recall and F1 score of VL-BERT on aligning the scene captions and the descriptive captions with the image.

Caption dataset	Label class	P	R	F1
HL1K	match	85	75	80
	not match	78	87	82
COCO	match	96	98	97
	not match	98	96	97

Table 5: the precision, recall and F1 score of VisualBERT on aligning the scene captions and descriptive captions with the image.

Therefore, we decide to make use of the logit score generated by the image sentence alignment head of the VL model in the VOLTA framework, as the logit score represents the degree of the certainty that VL models classify the input caption and the image as a match or not a match. If the logit score of classifying the caption and the image as a match is high and the ground truth is the same, this means the VL model have a high certainty to judge the image and the caption are aligned therefore the VL model prefer to use this kind of caption to describe the image. However, the VL model can also make wrong decisions which misclassifies the caption and the image as not a match. If the VL model misclassifies one kind of caption as not a match with the image but classifies another kind of caption as a match with the image accurately, we think the model prefers to use the kind of caption classified accurately to describe the image compared to use the kind of caption misclassified. Because the caption that is classified as not being aligned with the image means the VL model does not think the caption and the image are a match and will not use this caption to describe the image, but the VL model will use the kind of caption classified as a match with the image accurately to describe the image and this caption can be regarded it is preferred to describe the image by the VL model.

This is because the image sentence alignment head is a linear function and learned the optimal weights and biases to classify whether the linguistic features and the visual features which are encoded in the representations are a match in the pre-training procedure of the VL models, the representations generated from the scene caption has the same visual features and the way the linguistic features interact with the visual features as the representations generated from the descriptive caption, the differences between the representations generated from the scene caption and the the representations generated from the descriptive caption lie in the linguistic features and the attention score computed by these linguistic features with the same visual features, if the score of aligning the scene caption with the image is higher than the score of aligning the descriptive caption with the image, then this indicates the linguistic features in the scene caption or the way the linguistic features interacts with the visual features are more helpful than the linguistcic features in the descriptive features or the way the descriptive features interact with the visual features, thus the scene caption is preferred to describe the image than the descriptive caption in this case. if the representations generated from the linguistic features in the scene caption through the encoder of the VL model assigned the higher score by the image sentence alignment head than the descriptive caption, this indicates the linguistic features in the scene caption.

Based on the above ideas, we summarize four kinds of cases:

- If the scene caption and the descriptive caption are both classified as a match with the image, the kind of caption assigned the higher logit score to be classified as a match with the image by the image sentence alignment head is concluded being preferred by the VL model to describe the image compared to use another kind of caption.

- If the VL model misclassifies the scene caption as not a match with the image and classify the descriptive caption and the image are a match accurately, we conclude that the VL model prefers to use scene captions to describe the image than using descriptive captions to describe the image. If the VL model predicted descriptive captions are not aligned with the images wrongly and predicted scene captions are aligned with the images correctly, we regard that scene captions are preferred to be used to describe the image than using descriptive captions to describe the captions by the VL model.

Scene	Descriptive	Preferring scene captions(%)	Preferring descriptive captions(%)
match	match	19.04	80.96
match	not match	11.45	0
not match	match	0	88.55
not match	not match	52.80	47.20

Table 6: The experiment result of whether LXMERT prefers the scene caption or the descriptive caption. "match" in the first or second column means the model classify the caption is a match with the image, and "not match" means the modal classify the caption is not a match with the image. The number in the third column represents the percentage of the case the model prefers scene captions compared to descriptive captions, and the number in the fourth column represents the percentage of the case the model prefers descriptive captions compared to scene captions. This setting is used in the following tables.

- If the scene caption and the descriptive caption are both misclassified as not a match with the image, the kind of caption has the lower logit score to classify it as not a match with the image assigned by the image sentence alignment head is regarded being preferred by the VL model to describe the given image compared to using another kind of caption.

Table 7,Table 8,Table 9, Table 10 shows the experiment result of whether the scene caption or the descriptive caption is preferred by LXMERT, VL-BERT, UNITER and VisualBERT to describe the images in our experiment respectively.
In the case of both the scene caption and the descriptive caption are classified as a match with the image, the quantitative difference between the percentage of preferring to use the descriptive caption to describe the image and the percentage of preferring to use the scene caption to describe the image is higher than 60%. We notice that LXMERT has the highest percentage of preferring to use the scene caption to describe an image although LXMERT has the lowest accuracy in the scene recognition, while VisualBERT has the lowest percentage of preferring to use the scene caption compared to the descriptive caption but VisualBERT has the highest accuracy in recognizing the scene. In the case the scene caption is predicted as a match with the image while the descriptive caption is misclassified as not a match with the image, and the case that the descriptive caption is classified as a match with the image accurately while the scene caption is misclassified as not a match with the image, the quantitative difference between the percentage of preferring to use the descriptive caption to describe the image and the percentage of preferring to use the scene caption to describe the image is higher than 77%. This indicates the probability of the case classifying the descriptive caption is aligned with the image accurately and misclassifying the scene caption is not aligned with the image is much higher than the probability of the case classifying the scene caption is aligned with the image accurately and misclassifying the descriptive caption is not aligned with the image for all VL models. LXMERT has the highest percentage of preferring to use scene captions to describe the image while VL-BERT has the lowest percentage of preferring to use scene captions to describe the image. In the case both the scene caption and the descriptive caption are misclassified as not a match with the image, the percentage of preferring to use scene captions to describe the image is higher than the percentage of preferring to use descriptive captions to describe the image for LXMERT and UNITER, while the percentage of preferring to use scene captions to describe the image is lower than the percentage of preferring to use the descriptive captions to describe the image for VL-BERT and VisualBERT. We believe the difference of the linguistic information in the input caption and the information encoded through the interaction of the visual modality and the linguistic modality via the attention mechanism cause the gap between the percentage of preferring to use scene captions to describe an image and the percentage of preferring to use descriptive captions to describe an image as they are the only difference between using the scene caption and the descriptive caption as the input of the single stream VL models in the VOLTA framework.

References

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs.

Scene	Descriptive	Preferring scene captions(%)	Preferring descriptive captions(%)
match	match	19.04	80.96
match	not match	11.45	0
not match	match	0	88.55
not match	not match	52.80	47.20

Table 7: The experiment result of whether LXMERT prefers the scene caption or the descriptive caption. "match" in the first or second column means the model classify the caption is a match with the image, and "not match" means the modal classify the caption is not a match with the image. The number in the third column represents the percentage of the case the model prefers scene captions compared to descriptive captions, and the number in the fourth column represents the percentage of the case the model prefers descriptive captions compared to scene captions. This setting is used in the following tables.

Scene	Descriptive	Preferring scene captions(%)	Preferring descriptive captions(%)
match	match	16.42	83.58
match	not match	7.45	0
not match	match	0	92.55
not match	not match	36.67	63.33

Table 8: The experiment result of whether VL-BERT prefers the scene caption or the descriptive caption.

Scene	Descriptive	Preferring scene captions(%)	Preferring descriptive captions(%)
match	match	17.61	82.39
match	not match	10.92	0
not match	match	0	89.08
not match	not match	54.84	45.16

Table 9: The experiment result of whether UNITER prefers the scene caption or the descriptive caption.

Scene	Descriptive	Preferring scene captions(%)	Preferring descriptive captions(%)
match	match	13.58	86.42
match	not match	7.91	0
not match	match	0	92.09
not match	not match	33.33	66.67

Table 10: The experiment result of whether VisualBERT prefers the scene caption or the descriptive caption.

Transactions of the Association for Computational Linguistics, 9:978–994.

Michele Cafagna, Kees van Deemter, and Albert Gatt. 2021. What vision-language models 'see' when they see scenes. *CoRR*, abs/2109.07301.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. *CoRR*, abs/2005.07310.

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.

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. *CoRR*, abs/1908.03557.

Pei Li, Xinde Li, Xianghui Li, Hong Pan, M. O. Khyam, Md. Noor-A-Rahim, and Shuzhi Sam Ge. 2021. Place perception from the fusion of different image representation. *Pattern Recognition*, 110:107680.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context.

Ariadna Quattoni and Antonio Torralba. 2009. Recognizing indoor scenes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 413–420.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. pages 2556–2565.

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.

Hao Tan and Mohit Bansal. 2019. LXMERT: learning cross-modality encoder representations from transformers. *CoRR*, abs/1908.07490.

Ji Zhang, Jean-Paul Ainam, Li-hui Zhao, Wenai Song, and Xin Wang. 2022. Scene recognition with objectness, attribute and category learning.