Object-less Vision-language Model on Visual Question Classification for Blind People

Tung Le[1], Khoa Pho[1], Thong Bui[2,3], Huy Tien Nguyen[2,3] and Minh Le Nguyen[1]
[1]*School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, Japan*
[2]*Faculty of Information Technology, University of Science, Ho Chi Minh city, Vietnam*
[3]*Vietnam National University, Ho Chi Minh city, Vietnam*

Keywords: Visual Question Classification, Object-less Image, Vision-language Model, Vision Transformer, VizWiz-VQA.

Abstract: Despite the long-standing appearance of question types in the Visual Question Answering dataset, Visual Question Classification does not received enough public interest in research. Different from general text classification, a visual question requires an understanding of visual and textual features simultaneously. Together with the enthusiasm and novelty of Visual Question Classification, the most important and practical goal we concentrate on is to deal with the weakness of Object Detection on object-less images. We thus propose an Object-less Visual Question Classification model, OL–LXMERT, to generate virtual objects replacing the dependence of Object Detection in previous Vision-Language systems. Our architecture is effective and powerful enough to digest local and global features of images in understanding the relationship between multiple modalities. Through our experiments in our modified VizWiz-VQC 2020 dataset of blind people, our Object-less LXMERT achieves promising results in the brand-new multi-modal task. Furthermore, the detailed ablation studies show the strength and potential of our model in comparison to competitive approaches.

1 INTRODUCTION

Among multi-modal topics which requires advanced technologies in many cutting-edge areas, vision-language tasks are challenging and potential to dig much deeper due to the associated composition between images and texts. This reflects in many intriguing tasks such as Visual Question Answering (Le et al., 2020), Visual Commonsense Reasoning (Wang et al., 2020), and so on. Those works focus on analyzing, understanding, and retrieving the relationship among various modalities (Le et al., 2021a).

In the flow of vision-language studies, we propose an interesting and novel task named Visual Question Classification (VQC) to determine the category of visual questions. With an aid of a Visual Question Classification approach, it is useful to determine the answer space for the special kinds of questions such as Yes/No and Number ones. Although this is the first time this task is introduced, the category of visual question always exists in most Visual Question Answering datasets such as VQAv2.0 (Goyal et al., 2017), VizWiz-VQA (Gurari et al., 2018), etc. This fact reflects the enormous potential of question type which has not ever been used in the previous approaches.

Figure 1: The typical examples of low-qualified images.

Together with the potential of question type, our concerns also come from low-qualified images in practice. Throughout our work, object-less images are the representative of samples that have few objects recognized from Object Detection models. Typical examples of these kinds of images are presented in Figure 1. In these cases, it is not effective to utilize Object Detection models such as Faster R-CNN (Ren et al., 2015) to extract the object-based features. Therefore, an important and practical research question is how to take advantage of recent vision-language models to deal with the real-world images. Obviously, object-based approaches seem vulnerable due to the resolution of images. To over-

Le, T., Pho, K., Bui, T., Nguyen, H. and Nguyen, M.
Object-less Vision-language Model on Visual Question Classification for Blind People.
DOI: 10.5220/0010797400003116
In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022) - Volume 3*, pages 180-187
ISBN: 978-989-758-547-0; ISSN: 2184-433X

come the challenges of practical images, we propose an object-less generator to make use of Transformer-based image features in constructing the virtual objects, which is pragmatic and effective to replace the role of Object Detection component in recent vision-language models. Our proposed component is less influenced by side effects of poor quality images. It is completely suitable and effective to adapt into practical environment. However, there is no denying that object-based features are more complicated and efficient in the high-qualified images due to informative representation. Although our proposed architecture is not limited in poor-qualified vision, it is meaningful to emphasize the strength of our proposed architecture in the specific domain (i.e in VizWiz-VQA dataset of blind people).

Our main contributions are as follows:

- We introduce a novel task, Visual Question Classification, in the cutting-edge area between vision and language. It is the first time this task is regarded as an independent problem despite its importance and concealed long-term viability.
- We propose the Object-less Visual Question Classification model which takes advantage of image features to generate the virtual objects and integrates the vision-language models to predict the category of visual questions.
- Experimental results and ablation studies on VizWiz-VQC 2020 dataset prove the effectiveness and robustness of our virtual objects against object-based models.

In the remain of this paper, some related works are shown in Section 2. Then, the detail of components and our model is represented in Section 3. Next, to prove the effectiveness of our model, the result of our model, as well as the comparison with the state-of-the-art models, are represented thoroughly in Section 4. In addition, we also present some samples and discussion in Section 5. Finally, the conclusion and future works are shown in Section 6.

2 RELATED WORKS

2.1 Text Classification

In recent years, with the development of Transformer architecture and transfer-learning, text understanding achieves significant improvements, especially in sentence classification. Well-known models in this trend such as BERT (Devlin et al., 2019), XLNet (Yang et al., 2019) is one of the competitive approaches in a lot of text classification datasets. Different from previous autoencoding and autoregressive approaches, XLNet proposes a new framework to learn the context of a word based on the contribution of all tokens via the permutation operation. Generally, the strength of these approaches depends on the self-supervised learning from huge datasets and the robustness of self-attention in Transformer architecture.

In contrast to inductive learning in the above approaches, transductive learning techniques are effective to model the relationship between texts via observing all data samples. In the traditional approaches, the category of text is considered by the local context the global relationship through textual graphs. In this kind of approach, BERT-GCN is the powerful approach that combines a large-scale pre-training language model and transductive learning. Through a heterogeneous graph of textual elements, TextGCN (Yao et al., 2019) is able to learn the text representation via the relationship matrix of nodes and weighted edges. To integrate external language model, BERT-GCN (Lin et al., 2021) proposes an ensemble classifier with the contribution of both BERT (Devlin et al., 2019) and TextGCN (Yao et al., 2019).

2.2 Vision-language Model

In the rapid growth of multi-modal information, vision-language models have been receiving interest from both research and industry. However, the difference of representation between images and texts is a huge hurdle in previous works. In recent years, there is a lot of effort to overcome this challenge in many vision-language tasks such as Visual Question Answering (Le et al., 2020), Visual Commonsense Reasoning (Wang et al., 2020), etc.

In previous approaches, visual and textual features are independently considered and combined by the multi-modal fusion function. These systems, however, accidentally ignore the composition between textual and visual objects. The typical example of this type is Vision-Text Transformer (Le et al., 2021b) which utilizes Vision Transformer (Dosovitskiy et al., 2021) and BERT (Devlin et al., 2019) to extract visual and textual features. The success of this model comes from the robustness of pre-trained vision and language model instead of the interaction between images and texts.

Recently, another branch where images and texts are much more considered simultaneously is pre-training vision-language models. One of the most successful vision-language models, LXMERT (Tan and Bansal, 2019), utilize bi-directional cross-

encoder attention to learn the visual and textual features together. In particularly, LXMERT utilizes the Faster R-CNN to extract the object-based features of images. After, the textual and visual features are intensified and combined by Transformer architecture. Besides, it also proposes some pre-training tasks to optimize the vision-language model. However, its bottle-neck comes from the external Object Detection which is profoundly affected by the images' quality.

3 METHODOLOGY

The performance of most current vision-language models depends on the quality of Object Detection system, which is the bottle-neck in practical domain, especially in poor-qualified images. With our observation and the trend of image processing, we propose an object-less generator which utilizes the visual features to eliminate the requirements of the external Object Detection models. Instead of starting from scratch, we take advantage of pre-trained models via transfer-learning to construct the powerful virtual objects. To prove the efficiency of our proposed component, we integrate it into one of the most successful vision-language models, LXMERT (Tan and Bansal, 2019).

3.1 Object-less Generation

3.1.1 Image Feature Extraction

Together with the success of Transformer architecture in Natural Language Processing, more and more powerful approaches have appeared in many Computer Vision tasks such as Image Classification (Dosovitskiy et al., 2021), Image Super-Resolution (Parmar et al., 2018), etc. Transformer architecture is more promising than traditional Convolution Neural Network (CNN) to capture self-attention on pixels. With the fewest modifications from Transformer architecture, Vision Transformer model treats an image as a series of patches instead of pixels. This mechanism is efficient to maintain the spatial relationship in images and drop the computational cost.

Specifically, an image is split into a list of N fixed-size patches which are flattened and mapped into the featured space via linear projection. Then, each patch is injected by its position information via position embedding to keep the spatial relation of patches in the original image. The detail of the Vision Transformer in our image feature extraction is presented in Figure 2. We also notice that Vision Transformer is used as the external feature extractor instead of an internal

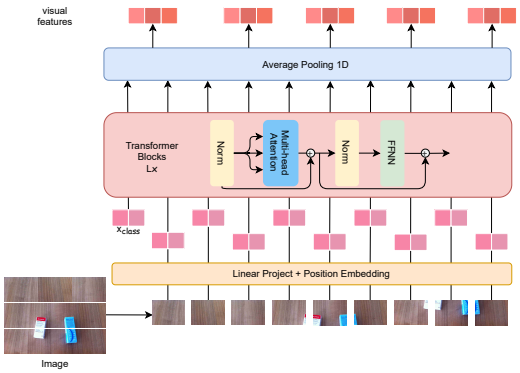


Figure 2: The detailed architecture of our Image Feature Extraction.

component in our model, which is similar to most previous image processing approaches. This mechanism is ideal enough to increase the speed of our system. Besides, to reduce the computation cost, we also integrate the stack of Average Pooling 1D in the visual features to obtain the generalized representation.

3.1.2 Object-less Generation

In our model, we assume that image features are sufficient to generalize visual contents such as objects, colors, backgrounds, and so on. Therefore, the global representation of images is a great alternative to output from an external Object Detection model. Furthermore, annotated data of image classification is easier and more reasonable than object-based resources. Even that, our proposed component is completely accomplished to integrate the portable image system into understanding visual content.

In the previous components, an image is represented into the feature vector $f_I \in \mathbb{R}^{d \times k}$ where d is the dimension of the output layer in Vision Transformer and k is the number of patches in images. In most vision-language approaches, an object consists of two main characteristics extracted by a specific region of interest (RoI). Firstly, with each selected RoI, the RoI features are derived from the mean-pooled convolution layer of the Object Detection model. Secondly, the position of bounding boxes in images is also utilized to represent objects. However, we also consider some thresholds to eliminate the uncertain objects. In traditional approaches, object-based features are often extracted by the process in the Up-Down model (Anderson et al., 2018). Besides, to maintain consistency in an input layer, the number of objects is often equal to 36.

Based on the configuration of previous approaches in image processing, our virtual objects also includes two kinds of features. With each patch $f_i \in f_I$, RoI

features of corresponding virtual object o_i are generated by a Feed-forward Neural Network Equation 1.

$$r_i = \tanh\left(W_r^T f_i + b_r\right) \quad (1)$$

Where $W_r \in \mathbb{R}^{d\times 2048}, b_r \in \mathbb{R}^{2048}$.

The dimension of RoI features is similar to the Up-Down model (Anderson et al., 2018) for the original objects. The number of virtual objects is, however, based on the number of patches in Vision Transformer instead of the Up-Down model (Anderson et al., 2018).

After generating the RoI features of virtual objects, we also learn the corresponding position via Feed-forward Neural Network models without an activation function in Equation 2.

$$p_i = W_p^T f_i + b_p \quad (2)$$

Where $W_p \in \mathbb{R}^{d\times 4}, b_p \in \mathbb{R}^4$. In most Object Detection models, we need to determine the specific format of bounding boxes such as width-height, point-length, etc. However, our Position Generator is efficient enough to learn spatial information in all configurations. Similar to RoI Generation, the number of position features are based on the number of patches from Vision Transformer (Dosovitskiy et al., 2021).

At a quick glance, our generation is too easy to consider as the critical component in the novel vision-language models. However, we need to notice that the features of objects depend on the bounding boxes. Therefore, each object is represented locally in each specific region of an image instead of considering the global features. In object-less images, there are a few objects with high confidence. It is the reason that object-based features dwell on redundant regions in images for low confident objects. Obviously, localization is indeed challenging in CNN-based approaches. In our generation, object features are created by the global representation of images, so our virtual objects are ideal to capture global information and learn local regions in images. Although our generation is simple and transparent, its generalization is quite high and efficient. The performance of our proposed generator is proved in the detailed results and ablation studies.

3.2 Object-less Visual Question Classification

The key component in previous vision-language approaches is to expand the input and embedding layer for processing both image and text simultaneously. However, most of them often depend on the external Object Detection models. With the fewest modification of the vision-language model, we propose a novel Object-less Visual Question Classification model. Our model takes advantage of the pre-trained LXMERT model with our virtual objects. with our proposed architecture, it is ideal to evolve gradually through the development of the Computer Vision and Vision-Language model.

Firtly, after generating the virtual object $o = \{o_i\}_{i=1}^m$ from our object-less generator, virtual object features r_i and position p_i are normalized by LayerNorm function (LN). Then, the position-aware embedding is calculated by adding the information of normalized r_i and p_i in Equation 3. From this aggregation, image representation becomes the combination of virtual object features and position information.

$$v_i = \frac{(LN(W_F r_i + b_F) + LN(W_P p_i + b_P))}{2} \quad (3)$$

Next, the visual features v and textual features q are intensified by Multi-head attention in Transformer architecture. To combine the multi-modal information, LXMERT proposes a cross-modality encoder between images and texts. In particularly, this component is the bi-direction multi-head attention from images to texts and vice versa. It is so similar to guided-attention (Yu et al., 2019) in previous works. However, the success of the LXMERT model comes from the pre-training strategies. This process allows vision-language models to learn the multi-modal data instead of single modality in traditional approaches.

In our architecture, we take advantage of the pre-trained LXMERT model to prove the strength of our virtual objects. Therefore, LXMERT architecture is almost maintained. Instead of utilizing the object-based feature, our OL-LXMERT model obtains the virtual objects via the internal object-less generator. It is pragmatic and effective to deploy in both practice and research. Even that, our proposed generator may completely replace the role of the Object Detection module in vision-language models in this task.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

In the consistent motivation of our work, we focus on the VizWiz-VQA dataset for blind people. All samples of VizWiz-VQA are collected by blind people. Therefore, images in the VizWiz-VQA dataset are often object-less and poor-qualified. Studying in this domain is ideal to raise the public interest for the disabled especially for blind people. It is humane and necessary to help them overcome their difficulties in

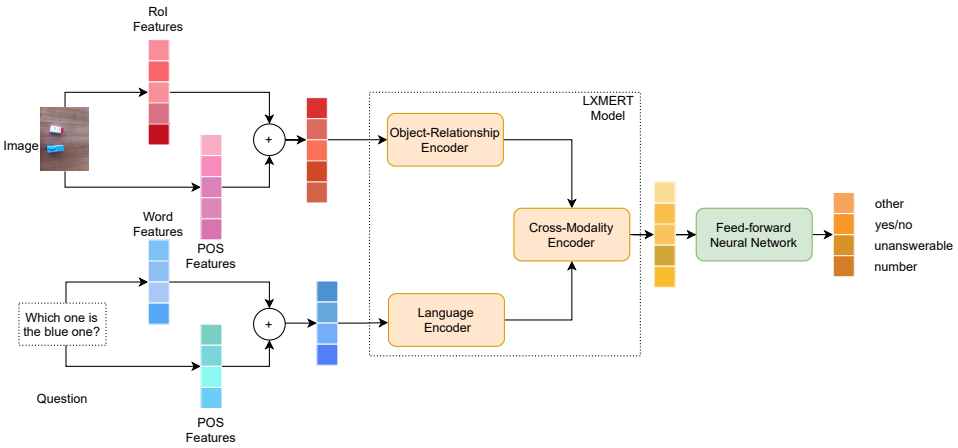


Figure 3: OL-LXMERT: The integration of object-less generator and vision-language model for Visual Question Classification.

the real world via deploying advanced technologies, especially in vision-language's understanding.

To create the VizWiz-VQC dataset, we derive the visual question information from VizWiz-VQA 2020 dataset. However, VizWiz-VQA is the concealment of test set. Therefore, we recommend a modification as follows:

- Consider the validation set in VizWiz-VQA as the test set of VizWiz-VQC for Visual Question Classification task.
- Split the original training set in VizWiz-VQA into the training and validation set of VizWiz-VQC dataset with the ratio 80% : 20%

The detail of our extracted VQC dataset from VizWiz-VQA 2020 is presented in Table 1. We also mention the question type distribution of our data in Table 2.

Table 1: The analysis of our dataset – VizWiz-VQC.

	Train	**Val**	**Test**
No. Samples	16,418	4,105	4,319
No. Question Types	4	4	4
Avg. Words/Question	6.76	6.74	7.26
Avg. Objects/Image (thresh = 0.4)	2.98	3.07	2.88

Table 2: The distribution of question type in VizWiz-VQC.

	Train	**Val**	**Test**
% Other	66.91	66.92	62.31
% Yes/No	4.67	4.65	4.51
% Unanswerable	26.95	26.97	32.07
% Number	1.47	1.46	1.11

Obviously, our choice and recommendation is suitable for object-less as well as real-world images. Particularly, in VQAv2.0 (Goyal et al., 2017), annotators are required to give a question of objects in images. Obviously, object-less images are often eliminated in most previous VQA datasets. Therefore, we consider VizWiz-VQA 2020 as the best choice for the challenges of object-less images. The most special characteristic of VizWiz-VQA is based on its data collection process where all samples are taken by blind people via their personal phones. It is the reason that most images in this dataset are poor-qualified and in low resolution. Specifically, the number of objects is approximately 2.9 per image. This ratio is much lower than the other datasets in the same configuration of Faster R-CNN. Together with the difficulty in object-less images, questions recorded by blind people are challenging for text understanding.

In evaluation, similar to previous approaches in Question Classification, we also consider F1-score as the main metric in Visual Question Classification. In addition, we also mention the precision and recall of our classifier in all comparisons.

4.2 Results

Coming from the brand-new appearance of VQC in the multi-modal tasks, it is too hard to choose the competitive baselines in this problem. As a pioneer in the task, we suggest comparing VQA task to two kinds of models including (i) general text classification and (ii) multi-modal VQA. In the first comparison, we mention two SOTA text classification models which are based on the latest technologies of XLNET (Yang et al., 2019) for pre-trained language understanding and BERT-GCN (Lin et al., 2021) for Graph Neural Network. These approaches only receive textual questions (Q) as an input instead of both images (I) and texts (Q). In the aspect of multi-modal systems, we compare our model to VT-

Table 3: The detailed comparison of our Object-less approach against the competitive baselines.

	Model	Precision	Recall	F1
Q	**BERTGCN (Lin et al., 2021)**	0.59	0.61	0.59
	XLNet (Yang et al., 2019)	0.57	0.58	0.57
QI	**VT-Transformer (Le et al., 2021b)**	0.59	0.65	0.61
QI	**LXMERT (Tan and Bansal, 2019)**	0.63	**0.70**	0.66
QI	**OL–LXMERT (Our model)**	**0.67**	0.69	**0.68**

Transformer (Le et al., 2021b) which is one of the most successful models in VizWiz-VQA task. The reason for this choice comes from the similar image feature extractor between two models. With a few modifications in the last prediction layer, we consider VT-Transformer as the most related approach in comparison to our model. Besides, to prove the strength of our model in object-less domain, we also present the performance of the original LXMERT together with object=based features from Faster R-CNN (Ren et al., 2015) models.

In Table 3, we show the performance of our Object-Less models (OL–LXMERT) in comparison to the existing state-of-the-art in text classification and multi-modal approaches. Our model obtains promising results against the competitive baselines in both single and multiple modalities. With the same architecture of LXMERT, our Object-less model is more efficient and encouraging to overcome the challenges in images from blind people. Instead of depending on external Object Detection models, our OL–LXMERT is ideal enough to take advantage of visual features to generate the local and global object representation. In addition, these results also reveal that Visual Question Classification is challenging and distinctive. Without the contribution of image (I), VQC is indeed arduous in text classification. This task is fully worthy enough to be considered in the independent problem in the multi-modal area.

4.3 Ablation Studies

In this part, we also conduct some ablation studies to emphasize the contribution of our proposed components. Firstly, we also present the reason that we consider Visual Question Classification as an independent task. With a cursory glance, VQC is quite similar to the text classification task in NLP. The results in Table 4, however, reflect this task's differences and challenges. In text classification, the category is determined by the context of natural language. However, the type of visual question is based on the relationship between images and texts. Especially, in VizWiz-VQC, questions are spoken by the blind in their daily lives, so it contains a lot of redundant information. In Table 4, the single modality model can not overcome the challenges in the VQC task. With the combination of images and questions, our approach and previous multi-modal system have enough features to give a correct choice.

Table 4: The contribution of images and texts in multi-modal VQC task.

	Model	Precision	Recall	F1
Q	BERT	0.58	0.63	0.59
I	Vision Transformer	0.56	0.57	0.55
QI	OL–LXMERT	**0.67**	**0.69**	**0.68**

Secondly, we also emphasize the strength of visual features from Transformer architecture against traditional Convolution Neural Work models. In this comparison, we utilize the image features from the latest improvement of the CNN-based model as EfficientNet (Tan and Le, 2019). Our detailed comparison between Transformer-based and CNN-based image feature extraction is presented in Table 5. Obviously, Vision Transformer obtains the best performance against the traditional approaches in CNN.

The visible question in this comparison is the weakness of virtual objects in CNN-based approaches against the Object-based model of Faster R-CNN. However, when we consider the architecture of Faster R-CNN carefully, it is easy to realize that both EfficientNet and Faster-RCNN are also deployed by the CNN models. Moreover, the annotated information of Object Detection is more greatly enriched than image classification. Besides, LXMERT (Tan and Bansal, 2019) architecture is designed and trained to satisfy the object-based models. Therefore, with the same fundamental architecture from CNN, the performance of Faster RCNN is better than EfficientNet in the LXMERT model.

However, as we mentioned above, the drawback of Convolution Neural Network models is based on the mechanism of kernels which only observe limited regions in images. In Object-less images, there are a few objects in high confidence, which means the features of local objects are less meaningful to cover all content of images. On the contrary, our model takes advantage of the Transformer-based approach to extract the global features in the object-less images.

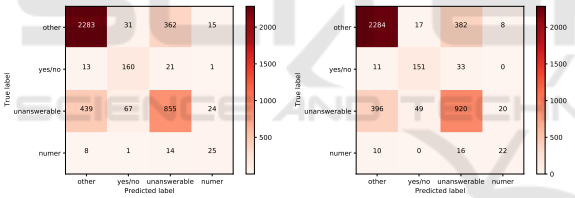
Table 5: The comparison of Transformer-based and CNN-based image feature extraction.

	Model	Precision	Recall	F1
Object-based	**Faster-RCNN (Ren et al., 2015)**	0.63	**0.70**	0.66
Objectless-based	**EfficientNet-b7 (Tan and Le, 2019)**	0.58	0.63	0.59
	EfficientNet-b6 (Tan and Le, 2019)	0.56	0.57	0.55
	ViT (B_16)	**0.67**	0.69	**0.68**

With the same process of virtual objects, the global features from Transformer architecture obtain significant results against CNN-based representation in both Image Classification and Object Detection models. It also reflects that our virtual images are auspicious enough to integrate the global and local features in images.

5 DISCUSSION

As a result, our virtual objects are created to reduce the dependence on object-based features in vision-language models. However, there are some special cases where the content of objects is so important. Through the detailed confusion matrix in Figure 4, the strength and weaknesses of our virtual objects are demonstrated clearly.



(a) LXMERT. (b) OL–LXMERT.

Figure 4: The detailed confusion matrix between LXMERT and OL–LXMERT.

Firstly, our Object-less LXMERT model outperforms the object-based LXMERT in total. Specifically, our model has significant success in the unanswerable samples which are the characteristic features of the VizWiz dataset against the previous works such as VQAv2.0 (Goyal et al., 2017), CLEVR (Johnson et al., 2017), etc. The main reason for the appearance of unanswerable samples in VizWiz comes from collection process from blind people. Obviously, these samples are too poor-qualified to detect any objects in images. Therefore, in these cases, our virtual objects are indeed efficient to predict the category of visual questions.

With the *other* question, both LXMERT and OL-LXMERT are equivalent in precision, recall and F1-score. Nevertheless, the recall of OL-LXMERT is worse than LXMERT model in *number* and *yes/no* question. It means that the object-based model focuses on the relationship between objects and keywords in question to predict the type of question. It also comes from the characteristic of the *number* and *yes/no* visual samples whose content is about the existence and quantity of objects. However, the precision of the LXMERT model in these kinds of questions is worse than OL-LXMERT. Although these samples tend to relate to the appearance of objects, most images in the VizWiz dataset are object-less. Therefore, our object-less model is more powerful to cover the global and local features through virtual objects.

Obviously, our virtual objects prove their strength and robustness in four kinds of questions. Especially, in unanswerable questions, our virtual objects clearly prove their importance and potential. Even, in the mainstream of object-based models in number and yes/no visual question, our object-less also obtains significant precision against LXMERT. However, we can not deny that the low recall of OL-LXMERT in the object-based question is also the weakness of our virtual objects. It encourages us to deploy the delicacy combination between real and virtual objects in the future works.

6 CONCLUSION

In this paper, we emphasize the importance and necessity of the Visual Question Classification task as an independent problem in the multi-modal area, especially in object-less images for blind people. We also propose the Object-Less LXMERT model to take advantage of the pre-trained vision-language model with the fewest modification via transfer learning. Our OL-LXMERT model is efficient to generate the virtual objects for replacing the role of the external Object Detection models in previous vision-language approaches. Through our detailed comparison and ablation studies, our Object-less LXMERT model achieves significant results against the competitive baselines in both single and multiple modalities in our extracted VizWiz-VQC 2020.

ACKNOWLEDGEMENTS

This work was supported by JSPS Kakenhi Grant Number 20H04295, 20K20406, and 20K20625. This research also was funded by the University of Science, VNU-HCM, Vietnam under grant number CNTT 2021-11.

REFERENCES

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. (2017). Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., and Bigham, J. P. (2018). Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Le, T., Nguyen, H. T., and Nguyen, M. L. (2020). Integrating transformer into global and residual image feature extractor in visual question answering for blind people. In *2020 12th International Conference on Knowledge and Systems Engineering (KSE)*, pages 31–36.

Le, T., Nguyen, H. T., and Nguyen, M. L. (2021a). Multi visual and textual embedding on visual question answering for blind people. *Neurocomputing*, 465:451–464.

Le, T., Nguyen, H. T., and Nguyen, M. L. (2021b). Vision and text transformer for predicting answerability on visual question answering. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 934–938.

Lin, Y., Meng, Y., Sun, X., Han, Q., Kuang, K., Li, J., and Wu, F. (2021). Bertgcn: Transductive text classification by combining gnn and bert. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., and Tran, D. (2018). Image transformer. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4055–4064.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28, page 91–99. Curran Associates, Inc.

Tan, H. and Bansal, M. (2019). LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111.

Tan, M. and Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Wang, T., Huang, J., Zhang, H., and Sun, Q. (2020). Visual commonsense r-cnn. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Yao, L., Mao, C., and Luo, Y. (2019). Graph convolutional networks for text classification. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pages 7370–7377.

Yu, Z., Yu, J., Cui, Y., Tao, D., and Tian, Q. (2019). Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6281–6290.