CEUR-WS.org/Vol-3349/paper4.pdf

VinVL+L: Enriching Visual Representation with Location Context in VQA

Jiří Vyskočil[1,*], Lukáš Picek[1]

[1]Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Technická 8, Pilsen, Czech Republic

Abstract

In this paper, we describe a novel method – VinVL+L – that enriches the visual representations (i.e. object tags and region features) of the State-of-the-Art Vision and Language (VL) method – VinVL – with Location information. To verify the importance of such metadata for VL models, we (i) trained a Swin-B model on the Places365 dataset and obtained additional sets of visual and tag features; both were made public to allow reproducibility and further experiments, (ii) did an architectural update to the existing VinVL method to include the new feature sets, and (iii) provide a qualitative and quantitative evaluation. By including just binary location metadata, the VinVL+L method provides incremental improvement to the State-of-the-Art VinVL in Visual Question Answering (VQA). The VinVL+L achieved an accuracy of 64.85% and increased the performance by +0.32% in terms of accuracy on the GQA dataset; the statistical significance of the new representations is verified via Approximate Randomization. The code and newly generated sets of features are available at https://github.com/vyskocj/VinVL-L.

Keywords

Vision and Language, Visual Question Answering, Location Recognition, Oscar, VinVL

1. Introduction

Multi-modal understanding systems can answer general questions from visual and textual data. These questions are largely focused on objects and their relations, appearances, or behaviors. Rest of them are asked about the overall scene, such as location or weather. Most of multi-modal systems are split into visual and textual modules, followed by image-text alignment. Faster R-CNN [1] region features of detected objects are commonly used for visual representation and BERT [2] embeddings for the textual. However, such visual model only provides information about objects, from which the entire multi-modal system must decide simple questions like "*Are people inside or outside*?".

We intuitively feel that in general, the objects are related to indoor/outdoor scene division even if they cannot be directly assigned. They have a certain weight on the basis of which the correct answer can be decided. For example, cars, sky, and trees are more likely to belong to an outdoor scene, however, the scene may be indoors, and these categories can be detected through the garage door. In addition to [3, 4, 5], the mentioned paradigm of splitting image-text modules follows VinVL [6] based on Oscar [3] that additionally adds object tags, i.e., textual output from an Object Detection network, to region features. However, a clear cross-modal representation of the scene is still missing, which can harm the network, as shown in Figure 1.

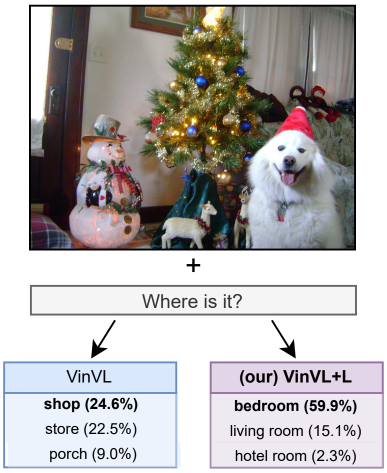

Figure 1: Example predictions of the proposed VinVL+L. We compare VinVL+L with the State-of-the-Art VinVL on the randomly selected input pair (i.e. image and question) from the GQA test set. The VinVL+L better aligns the answer to the question thanks to the enriched visual features.

Our method, based on VinVL, brings a new representation including information about the location into the system. This representation is obtained using a classification network trained on the Places365 dataset having a total of 365 location categories. All of these categories are directly split into one of the indoor and outdoor supercat-

---

26th Computer Vision Winter Workshop, Robert Sablatnig and Florian Kleber (eds.), Krems, Lower Austria, Austria, Feb. 15-17, 2023

*Corresponding author.
vyskocj@kky.zcu.cz (J. Vyskočil)
0000-0002-6443-2051 (J. Vyskočil); 0000-0002-6041-9722 (L. Picek)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

egories. All of these labels are then passed as scene tags to our VinVL+L method to predict the answers. Besides, we utilize scene features that are generated in the same way as the region features of Oscar/VinVL. Finally, we evaluate influences on answers while using these novelties. An example of the top 3 predictions of the VinVL and our VinVL+L is visualized in Figure 1. More examples are shown in Section 5.3 and Appendix A. Our contributions are:

- We enrich visual representations of the VinVL using the global information about the image - location.
- We present the effectiveness of each new cross-modal representation as we compare their related models including a reproduced version of the VinVL.
- We improve the VinVL in visual question answering (VQA) with an overall accuracy of **64.85%** on the GQA dataset.
- We provide data with the location context that we generated for the GQA dataset.

2. Related Work

Many Vision and Language (VL) methods, like [3, 6, 7, 8], focus on pre-training generic models by combining multiple datasets from different tasks. Then the models are fine-tuned to downstream tasks that include: image captioning, visual reasoning, or visual question answering. In this section, we briefly review recent approaches to VL tasks and their commonly used Vision Encoders, which are the most relevant for our work.

Vision Encoders Convolutional Neural Networks (CNNs) gained popularity in image classification when AlexNet [9] won the ImageNet 2012 competition. In the subsequent period, models with skip-connections [10, 11, 12, 13] with blocks having small feed-forward networks in parallel connections [12, 14, 15], or with a focus on optimization [16, 17, 18, 19, 20] were created. In recent years, Transformer-based methods, such as Vision Transformer [21], or its modification with shifted windows [22], gained favor thanks to computational efficiency and accuracy. These image classification models are often used as backbone architectures in object detection to predict bounding boxes with a classification of each object in the image. The most popular detectors are the one-shot Yolo-based architectures [23, 24] and two-shot Faster R-CNN-based architectures [1], which are generally slower but more accurate than the one-shot ones. The image classification or object detection models are further used as Visual Encoders in the VL tasks.

BERT-based VL Methods End-to-end methods such as MDETR [7] use a pre-trained image classification backbone to extract features and concatenate them with word embeddings taken from a BERT-based model [2, 25]. However, some existing VL methods [3, 4] reuse the extracted features from another approach, e.g., a bottom-up mechanism [5] that extracts object regions via Faster R-CNN, to fine-tune a novel method with an unchanged visual model. These methods include Oscar [3], which introduces object tags as cross-modal representation to improve the alignment of the image-text pairs. Based on Oscar, VinVL [6] improves visual representation by pre-training larger model on multiple object detection datasets. Since this method holds State-of-the-Art results on the GQA dataset [26] and represents the image as a set of regional features while suppressing global scene information, we decided to improve the alignment of the cross-modal representation by location recognition.

3. Datasets

The early datasets, such as VQA [27] and COCO-QA [28], contain only the core annotation needed for the vision question answering: an image, a question, and a desired one-word answer. However, we are interested in dataset containing richer annotations to recognize types of locations in the image input. It does contain the GQA dataset, but only for part of the images. Therefore, there are two existing datasets Places365 and GQA suitable for our task. Both datasets are thoroughly described below.

Places365 [29] This dataset consists of 365 location categories that we can directly map to indoors/outdoors category. The balanced training set varies from 3,068 to 5,000 images per location category, while the validation set consists of 50 images per category.

GQA dataset [26] This dataset consists of 22,669,678 questions (from which the test2019 split contains 4,237,524 questions) over 113,018 images with 1,878 possible answers to open and binary yes/no questions. In addition to questions and answers, each image contains annotations of objects, the relations between them, and their attributes. Besides, each image contains global information in the form of location and weather, the distribution of which is shown in Table 1. Regarding the evaluation of the results, the following metrics are used:

- *Accuracy* – overall accuracy, primary metric,
- *Binary* – accuracy of yes/no questions,
- *Open* – accuracy of open questions,
- *Consistency* – overall accuracy including equivalent answers,

- *Plausibility* – relative number of answers making sense with respect to the dataset,
- *Validity* – relative number of answers that are in the question scope,
- *Distribution* – overall match between the distributions of true answers and model predictions.

Table 1
GQA dataset. Distribution of annotated global information about the scenes on the training and validation split.

Metadata	Training	Validation
# of images	74,942	10,696
with weather	6,600 *(8.8%)*	952 *(8.9%)*
with location	23,370 *(31.2%)*	3,265 *(30.5%)*
indoors	4,520 *(19.3%)*	638 *(19.5%)*
outdoors	18,850 *(80.7%)*	2,627 *(80.5%)*

4. Methodology

The Vision and Language (VL) approaches are commonly divided into two phases: pre-training and fine-tuning. In pre-training, multiple datasets of different tasks are combined to create generic models. In fine-tuning, these models are then trained on each of these datasets, called downstream tasks. In this study, we focus on improving the current State-of-the-Art VinVL [6] in GQA dataset [26]. This improved version learns the image-text representation with respect to the global information of an entire image, such as indoors/outdoors, which is given by novel scene tags and features.

4.1. Adding locations to VinVL

Based on VinVL [6], we present an extended architecture with scene tags and features. In our work, these representations are simply generated using a fine-tuned classification network on the Places365 dataset [29] with an accuracy of up to 96.1% in case of binary indoor/outdoor classification (see Section 5.1 for more details). Scene tags are the predicted location categories. Scene features are made in the same style as their object counterparts, i.e., as a 2,048 feature vector (obtained via Global Average Pooling) concatenated with top-left & bottom-right corners, and height & width. Besides, the novel scene representations are prepended before the object ones so that the scenes in the embeddings always have the same position for each image-text pair input, as outlined in Figure 2.

Even though we do not perform pre-training on various tasks with the new representation, in general, the yet-established pre-training objective of Oscar/VinVL [6] can be followed. The change is only in the definition of the (w, q, v) triple input, where w is the word embedding sequence of the text, q is the word embedding sequence of the **scene** and object tags detected from the image, and v is the visual embedding sequence of the **entire image** and all detected regions. This input can be viewed from two different perspectives as [3, 6]:

$$x \triangleq [\underbrace{\underbrace{w,q}_{Q\&A}\ ,\ \underbrace{v}_{img}}_{Dictionary\ View}]\ or\ [\underbrace{\underbrace{w}_{caption}\ ,\ \underbrace{q,v}_{tags\&img}}_{Modality\ View}] \tag{1}$$

where *Dictionary View* defines Masked Token Loss $\mathcal{L}_{MTL}$, applied on the discrete token sequence $h \triangleq [w, q]$, to predict the masked tokens h_i based on their surrounding tokens $h_{\backslash i}$:

$$\mathcal{L}_{MTL} = -\mathbb{E}_{(h,v)\sim\mathcal{D}} \log p(h_i | h_{\backslash i}, v) \tag{2}$$

Modality View defines Contrastive Loss $\mathcal{L}_{CL}$ for the image representation $h' \triangleq [q, v]$, which is "polluted" by randomly replacing q with another sequence of tags from the dataset $\mathcal{D}$. To distinguish the original pair $(y = 1)$ from the polluted one $(y = 0)$, a binary classifier $f(.)$ as a fully-connected layer is applied on the top of the [CLS] token. This loss function is defined as [3]:

$$\mathcal{L}_{CL} = -\mathbb{E}_{(w,h';y)\sim\mathcal{D}} \log p(y | f(w, h)) \tag{3}$$

Alternatively, VinVL [6] applies the 3-way Constrastive Loss $\mathcal{L}_{CL3}$ on $h^* \triangleq [w, q, v]$, instead of the binary $\mathcal{L}_{CL}$ used in Oscar [3], to predict whether the (w, q, v) triplet is the original one $(c = 0)$, contains a polluted w $(c = 1)$, or contains a polluted q $(c = 2)$:

$$\mathcal{L}_{CL3} = -\mathbb{E}_{(h^*;c)\sim\mathcal{D}} \log p(c | f(w, q, v)) \tag{4}$$

By fusing Equation 2 and 4, or 2 and 3, the full pre-training objective is:

$$\mathcal{L}_{Pre-training} = \mathcal{L}_{MLT} + \mathcal{L}_{CL3\ (or\ CL)} \tag{5}$$

4.2. Implementation Details

We use the same feature-vector size (i.e. 2,048) in order to match the size of VinVL. These features are then concatenated with positions and sizes, as described in Section 4.1. The models used from the Timm library [30] are: resnext50d_32x4d [13], gluon_inception_v3 [15], mobilenetv3_small_100 [18], gc_efficientnetv2_rw_t [20], vit_large_patch16_224_in21k [21], and swin_base_patch4_window7_224_in22k [22]. All models are fine-tuned for 20 epochs with SGD and Focal Loss. We use an initial learning rate of 0.01 and we reduce it with a plateau scheduler. The batch size is 64 with 2 accumulation steps. We use horizontal flip (probability of 50%), random resized crop (scale from 0.8 to 1.0), and random brightness contrast (probability of 20%).

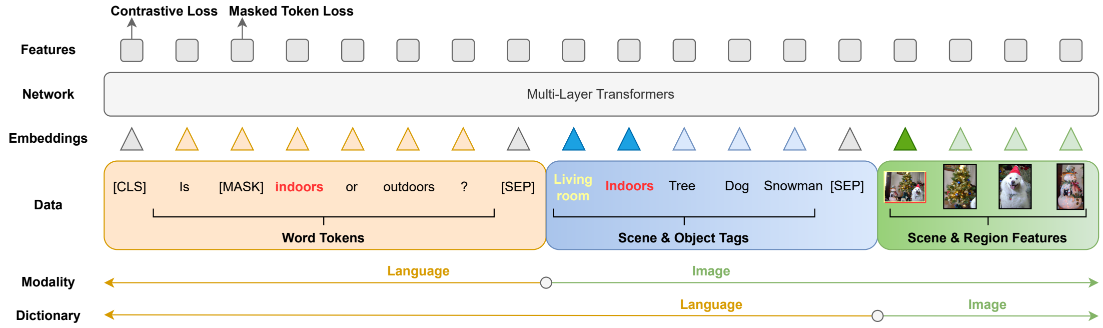

Figure 2: Illustration of VinVL+L. We represent the image-text pair as a quintuple [word tokens , scene tags , object tags , scene features , region features], where word tokens, object tags and region features are taken from VinVL [6]. Scene tags and features are proposed to improve the alignment of cross-domain semantics. The example shows a case where only detected objects could be classified as outdoors rather than indoors.

In the case of the VL model, we use the pre-trained Oscar+$_{\text{BASE}}$ with VinVL features and follow their presented procedure which is the same as the original Oscar, i.e., pre-training on the unbalanced "all-split" of the GQA dataset for 5 epochs, and fine-tune the best model with respect to overall accuracy on the "balanced-split" for 2 epochs. All the results are shown in Section 5.2 together with a reproduction of VinVL that we improve.

5. Experiments

Our approach is divided into two separate steps. First, we adapt several image classification models to the Places365 dataset and select the most accurate model to generate a visual representation for the VL model. Then, we fine-tune the VL model using its original and our new visual features.

5.1. Location Recognition

We selected several pre-trained image classification networks in order to cover a certain range of different approaches to location recognition. These methods include those focused on high inference speed, methods containing skip-connections, parallel paths, or transformers. Results of fine-tuned models on the Places365 dataset are shown in Table 2. ResNeXt-50, EfficientNetV2, and ViT-Large have similar performance, while ViT-Large performs slightly worse in indoors/outdoors classification. It is because when the ViT-Large is not right, it is more often the incorrect indoors/outdoors supercategory than in the case of the previous two mentioned models. The best results are achieved by Swin-Base in both 365 locations and binary indoors/outdoors recognition. It obtains 56% top-1 accuracy in recognizing 365 locations, which is 1.1% higher than that of the second-best ViT-Large. Therefore, this model is further used to extract novel visual representations for our VinVL+L.

Table 2
Performance evaluation of selected networks. We do the evaluation on the Places365-val dataset on all categories and Accuracy$_{\text{IO}}$ on their binary supercategories (Indoor/Outdoor).

Backbone	Accuracy	Top3	Accuracy$_{\text{IO}}$
MobileNetV3	47.9	70.8	94.6
InceptionV3	53.1	76.0	95.3
ResNeXt-50-D	54.2	77.0	95.6
EfficientNetV2	54.7	77.4	95.6
ViT-Large	54.9	77.7	95.5
Swin-Base	**56.0**	**78.7**	**96.1**

5.2. Visual Question Answering

Statistical significance of novel features We show the advantages of the new visual representations by comparing our method with the reproduced VinVL using the same training pipeline – see Table 5. The used scene tags as 365 location categories (C), or indoors/outdoors (IO), are denoted in subscripts of the model name. Besides, we compute the statistical significance [33] between the two models to show that recognizing the location categories truly brings benefits and it is not just a coincidence. For demonstration, we compare the reproduced VinVL with our VinVL+L$_{\text{C}}$ on the validation dataset. Our goal is to reject the null hypothesis defined as "there is no **difference** between system A and B". To do this, we shuffle the predictions between systems A and B with a probability of 50%, and we compare the performance with the initial one (all repeated 10,000 times). Consequently, we reject the null hypothesis at the 95% significance level, i.e., a

Table 3
Results of individual methods according to the official leaderboard. We show the prior State-of-the-Arts performance on GQA dataset, sorted by primary metric – Accuracy. The meaning of individual metrics is described in Section 3.

Method	**↑Accuracy**	**↑Binary**	**↑Open**	**↑Consist.**	**↑Plausib.**	**↑Valid.**	**↓Distrib.**
Bottom-Up [5]	49.74	66.64	34.83	78.71	84.57	96.18	5.98
MMN [4]	60.83	78.90	44.89	92.49	84.55	96.19	5.54
Oscar [3]	61.62	-	-	-	-	-	-
MDETR [31]	62.45	80.91	46.15	93.95	84.15	96.33	5.36
LXR955 [8]	62.71	79.79	47.64	93.10	85.21	96.36	6.42
NSM [32]	63.17	78.94	**49.25**	93.25	84.28	96.41	**3.71**
VinVL [6]	**64.65**	**82.63**	48.77	**94.35**	**84.98**	**96.62**	4.72

Table 4
Performance evaluation of individual scene tags. We compare the reproduced VinVL with additional scene tags as Indoors/Outdoors (IO), and/or 365 location category (C). The last row indicates improvement/deterioration as a difference between our best model and the reproduced VinVL.

Method	**↑Accuracy**	**↑Binary**	**↑Open**	**↑Consist.**	**↑Plausib.**	**↑Valid.**	**↓Distrib.**
VinVL (reproduced)	*64.53*	*82.36*	*48.79*	*94.14*	*84.77*	*96.55*	*4.72*
VinVL+L_{IO}	64.65	82.43	48.94	**94.17**	84.81	96.61	4.73
VinVL+L_{C+IO}	64.71	82.38	49.12	94.06	84.84	**96.65**	**4.55**
VinVL+L_{C}	**64.85**	**82.59**	**49.19**	94.00	**84.91**	96.62	4.59
$\Delta_{VinVL+L_C\,-\,VinVL}$	*+0.32*	*+0.23*	*+0.40*	*-0.14*	*+0.14*	*+0.07*	*-0.13*

Table 5
Accuracy of answers on the validation dataset. We evaluate the reproduced VinVL with our improved versions on the balanced validation GQA dataset.

Backbone	**Accuracy**	**Binary**	**Open**
VinVL (reproduced)	*63.2*	*52.5*	*82.3*
VinVL+L_{C+IO}	63.4	52.7	82.3
VinVL+L_{C}	63.8	53.0	**83.0**
VinVL+L_{IO}	**64.1**	**53.7**	82.6

threshold is equal to 0.05, with obtained $p\text{-}value = 0.03$. The same conclusion is reached for VinVL+L_{IO}. In the case of the VinVL+L_{C+IO}, the difference is not significant, so the null hypothesis cannot be rejected.

The significance may seem small from a general point of view. However, it should be considered that these results were achieved by simply adding locations to the system. To improve significance, scene features should be generated from the same model as region features. In addition, other global information such as weather may be included.

Comparison on the test set Although we followed the original training pipeline, on which the results of our models are based, it should be noted that the reproduced VinVL works worse than the original version. Therefore, we decided to select models after the 1st, 3rd, and 5th epochs from the pre-training on the unbalanced set. Then we fine-tuned these models for 2 epochs on the balanced set to slightly increase the final performance. We selected the best model with respect to overall accuracy on the validation set, and we pushed the results into the evaluation server. The performances of the models are listed in Table 4. The reproduced version of VinVL still has worse performance than the original one, but the difference is decreased with this modification of the training.

According to the results, all of our models answer more accurately and outperform the reproduced model in all metrics, except in some cases of Consistency and Distribution. For example, even the VinVL+L_C answers 0.40% better on open questions and 0.23% better on yes/no questions, resulting in 0.32% higher overall accuracy, it has 0.14% lower performance in Consistency metric. This means that when our model fails, the prediction is truly meaningless to the given question. However, this model shows the best performance compared with other versions of VinVL+L: VinVL+L_{IO} holds only the highest Consistency (+0.03% compared with reproduced VinVL and +0.17% compared with VinVL+L_C), and VinVL+L_{C+IO} outperforms all compared models in Validation and Distribution. We show the results of the prior State-of-the-Arts in Table 3. Even though, our VinVL+L method noticeably surpasses the original version in the primary metric: +0.20% of overall accuracy for VinVL+L_C.

5.3. Summary and Discussion

An improvement in the visual question answering is achieved by taking global information about the visual component into account. Table 4 and 5 confirm this fact

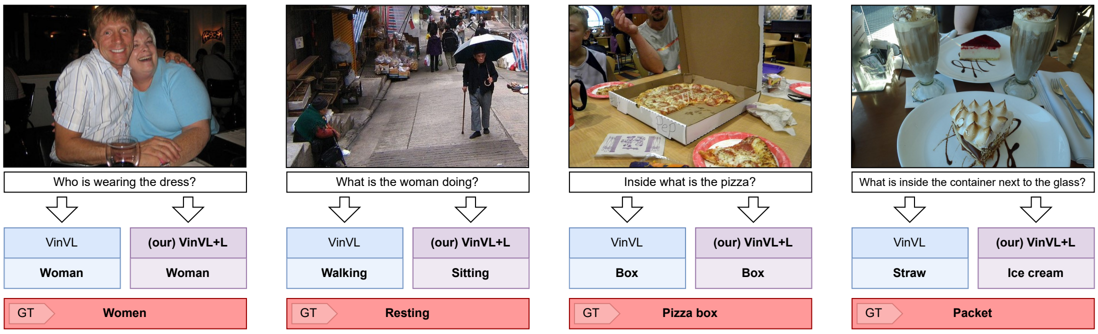

Figure 3: Wrong predictions w.r.t. Ground Truth labels (GT). We show the wrong predictions of two models (VinVL and VinVL+L) to randomly chosen image-question pairs from the validation set.

for all our VinVL+L models. In addition, we show the wrong predictions of our VinVL+L (along with predictions of reproduced VinVL) against the Ground Truth labels. The image-question pairs are randomly chosen from the validation set, see Figure 3. Even if our model answers are wrong in the given examples, it is worth saying that some of the answers are not truly wrong, e.g., in the second example, in which the woman is truly sitting and, in our opinion, there is missing additional information to say if she is really resting, instead of just sitting. Besides these examples, we show predictions from the test2019 set in Appendix A.

It is worth emphasizing that the listed models do not use scene features, only tags. A model using both scene tags and features did not achieve the expected results. This behavior was anticipated for two reasons. First, even if we follow the generating procedure of the scene features, the VL model obtains a vector with different semantics compared to region features. To solve this issue, the scene features must be generated from the same model to avoid subsequent confusion. Second, all image and text representations are passed to the modified BERT model, which is still a language model pre-trained on text corpora, with additional visual features added. Therefore, the words still have a higher weight than the visual features.

Regarding the performance of the reproduced VinVL, we used the original code including the pipeline presented in [6]. However, the network reproduced by us achieved worse performance in all metrics, e.g., 0.12% in overall accuracy. Since the main goal is to improve this method, we decided to primarily compare our models with the reproduced version, on which the benefits are best observed. All the listed models were trained using the same device, hyperparameters settings, only differ in the used novel visual representations. Therefore, our article only shows the effectiveness of incorporating global location information into a system that works only on the basis of objects.

6. Conclusion

This paper presents VinVL+L, an enriched version of the VinVL with location context as a novel visual representation. We generate the new representations as scene tags and features and we prepend them before the original embeddings of the architecture. Our version achieves higher overall accuracy than the original method on the GQA dataset, and we show that global information about the entire image influences the answers and thus should not be ignored. The best results of 64.85% overall accuracy are achieved with the model using 365 location categories as scene tags. Besides, we performed an Approximate Randomization test to verify that the achieved results are statistically significant. Similarly, weather recognition for outdoor scenes could be included in the concept to help the network with alignments of image-text pairs with respect to global information. All generated data and code are publicly available on our GitHub.

Acknowledgments

The work has been supported by the grant of the University of West Bohemia, project No. SGS-2022-017. Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA CZ LM2018140) supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

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

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[3] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al., Oscar: Object-semantics aligned pre-training for vision-language tasks, in: European Conference on Computer Vision, Springer, 2020, pp. 121–137.

[4] W. Chen, Z. Gan, L. Li, Y. Cheng, W. Wang, J. Liu, Meta module network for compositional visual reasoning, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 655–664.

[5] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, 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, 2018, pp. 6077–6086.

[6] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, J. Gao, Vinvl: Revisiting visual representations in vision-language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5579–5588.

[7] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, N. Carion, Mdetr-modulated detection for end-to-end multi-modal understanding, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1780–1790.

[8] H. Tan, M. Bansal, Lxmert: Learning cross-modality encoder representations from transformers, arXiv preprint arXiv:1908.07490 (2019).

[9] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 25, Curran Associates, Inc., 2012, pp. 1097–1105.

[10] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[11] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492–1500.

[12] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-first AAAI conference on artificial intelligence, 2017.

[13] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, M. Li, Bag of tricks for image classification with convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 558–567.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[16] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861 (2017).

[17] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.

[18] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al., Searching for mobilenetv3, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1314–1324.

[19] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.

[20] M. Tan, Q. Le, Efficientnetv2: Smaller models and faster training, in: International Conference on Machine Learning, PMLR, 2021, pp. 10096–10106.

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

[22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[23] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You

only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.
[24] A. Bochkovskiy, C.-Y. Wang, H.-Y. M. Liao, Yolov4: Optimal speed and accuracy of object detection, arXiv preprint arXiv:2004.10934 (2020).
[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[26] D. A. Hudson, C. D. Manning, Gqa: A new dataset for real-world visual reasoning and compositional question answering, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 6700–6709.
[27] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh, Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[28] M. Ren, R. Kiros, R. Zemel, Exploring models and data for image question answering, Advances in neural information processing systems 28 (2015).
[29] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: A 10 million image database for scene recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence (2017).
[30] R. Wightman, Pytorch image models, https://github.com/rwightman/pytorch-image-models, 2019. doi:`10.5281/zenodo.4414861`.
[31] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, N. Carion, Mdetr - modulated detection for end-to-end multi-modal understanding, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1780–1790.
[32] D. Hudson, C. D. Manning, Learning by abstraction: The neural state machine, Advances in Neural Information Processing Systems 32 (2019).
[33] S. Riezler, J. T. Maxwell III, On some pitfalls in automatic evaluation and significance testing for mt, in: Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 57–64.

A. Additional prediction examples

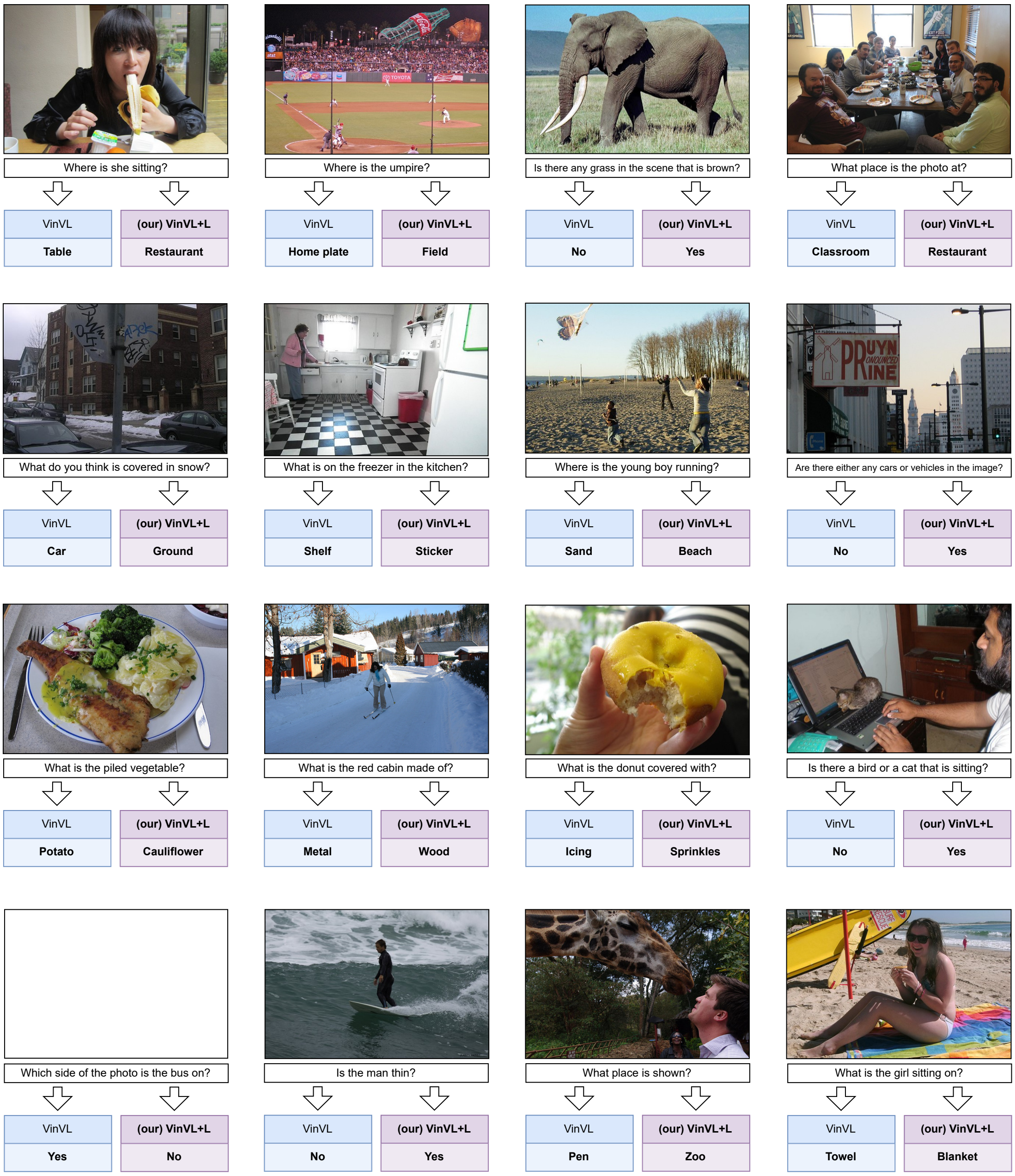
Where is she sitting?
VinVL
Table
(our) VinVL+L
Restaurant
Where is the umpire?
VinVL
Home plate
(our) VinVL+L
Field
Is there any grass in the scene that is brown?
VinVL
No
(our) VinVL+L
Yes
What place is the photo at?
VinVL
Classroom
(our) VinVL+L
Restaurant
What do you think is covered in snow?
VinVL
Car
(our) VinVL+L
Ground
What is on the freezer in the kitchen?
VinVL
Shelf
(our) VinVL+L
Sticker
Where is the young boy running?
VinVL
Sand
(our) VinVL+L
Beach
Are there either any cars or vehicles in the image?
VinVL
No
(our) VinVL+L
Yes
What is the piled vegetable?
VinVL
Potato
(our) VinVL+L
Cauliflower
What is the red cabin made of?
VinVL
Metal
(our) VinVL+L
Wood
What is the donut covered with?
VinVL
Icing
(our) VinVL+L
Sprinkles
Is there a bird or a cat that is sitting?
VinVL
No
(our) VinVL+L
Yes
Which side of the photo is the bus on?
VinVL
Yes
(our) VinVL+L
No
Is the man thin?
VinVL
No
(our) VinVL+L
Yes
What place is shown?
VinVL
Pen
(our) VinVL+L
Zoo
What is the girl sitting on?
VinVL
Towel
(our) VinVL+L
Blanket

Figure 4: Randomly selected predictions; VinVL+L and VinVL methods evaluated over GQA test2019 set. The VinVL+L method impacts a decision based on newly included binary location (i.e. indoor and outdoor) metadata. In most cases where VinVL+L prediction differs from VinVL, the VinVL+L produced a subjectively more reasonable prediction.