

Understanding Cross-modal Interactions in V&L Models that Generate Scene Descriptions

Michele Cafagna¹ Kees van Deemter² Albert Gatt^{1,2}

¹University of Malta, Institute of Linguistics and Language Technology

²Universiteit Utrecht, Information and Computing Sciences

michele.cafagna@um.edu.mt

{a.gatt, c.j.vandeemter}@uu.nl

Abstract

Image captioning models tend to describe images in an object-centric way, emphasising visible objects. But image descriptions can also abstract away from objects and describe the type of scene depicted. In this paper, we explore the potential of a state of the art Vision and Language model, VinVL, to caption images at the scene level using (1) a novel dataset which pairs images with both object-centric and scene descriptions. Through (2) an in-depth analysis of the effect of the fine-tuning, we show (3) that a small amount of curated data suffices to generate scene descriptions without losing the capability to identify object-level concepts in the scene; the model acquires a more holistic view of the image compared to when object-centric descriptions are generated. We discuss the parallels between these results and insights from computational and cognitive science research on scene perception.

1 Introduction

When humans view images, they can quickly capture their ‘gist’. For example, it is immediately evident that Figure 1 is a kitchen. Such judgments are fast and are informed by expectations about which objects occur in typical scenes (‘scene semantics’) and their configuration (‘syntax’) (Malcolm et al., 2016; Võ, 2021; Self et al., 2019). This knowledge affects the deployment of attentional resources (Torralba et al., 2006; Oliva and Torralba, 2007; Wu et al., 2014; Henderson and Hayes, 2017). Scene understanding and object recognition constrain the selection of attended locations in human visual attention (Itti and Koch, 2001).

In this paper, we explore the implications of these findings for image captioning models. There are at least two levels at which an image can be appraised. An **object-centric** perspective focuses primarily on individual objects and actions (e.g. the example caption in Fig 1). This has dominated captioning models (see Hodosh et al., 2013, for an



Figure 1: Image from the MS-COCO 2014 validation set. One reference caption is: *a man in a chefs hat chopping food.*

early, influential statement of this view) and has informed the design of widely-used datasets, which pair images with captions that explicitly mention at least some of the objects in a picture (e.g. Young et al., 2014; Chen et al., 2015; Pont-Tuset et al., 2020; Gurari et al., 2019; Sharma et al., 2018; Agrawal et al., 2019). In contrast, a **scene-level** caption (e.g. ‘a kitchen’ for Figure 1) contains less object-specific detail. Such captions are less redundant with respect to the image they describe, but convey enough information to generate inferences about content and structure (e.g. kitchens typically contain cupboards, but not birds; etc).

Most image captioning datasets contain object-centric captions and no currently available resource pairs both scene-level and object-centric captions with images. In this paper, we address this gap and ask (i) whether captioning models can be adapted both for object-centric and scene-level captioning and (ii) whether the two strategies rely on different types of interplay between the visual and linguistic modalities. Addressing these questions can shed light on the ability of V&L models to reason about the relationship between scenes and their components. In addition, it is desirable for mod-

els to generate scene-level descriptions as well as object-centric ones. In many communicative contexts, scene-level captions are informative and non-redundant, recalling the quality and the quantity discourse maxims defined by Grice (1975).

We present a study of object-centric versus scene-level captioning. We focus on VinVL (Zhang et al., 2021), a BERT-based model in the OSCAR family (Li et al., 2020b) of models, which have recently dominated the state of the art in image captioning.¹ Our main contributions are:

- i) We introduce a novel dataset, HL-Scenes (Sec 3) extending part of the COCO dataset (Chen et al., 2015) with scene-level descriptions.
- ii) We perform an in-depth investigation of the impact of fine-tuning on the pre-trained model. The analysis is designed to thoroughly inspect object-scene relations by exploiting cross-modal attention (Sec 5), coupled with probing (Sec 7) and ablation studies (Sec 6).
- iii) We show that (i) VinVL’s pre-trained representations are rich enough to support scene-level captioning, but that (ii) fine-tuning results in a different deployment of attentional resources. This bears parallels to the findings in research on human scene perception.

2 Related work

Datasets Existing image-caption datasets emphasise object-centric captions (an early exception, using abstract scenes, is Ortiz et al., 2015). This is also true of web-sourced datasets such as Conceptual Captions (CC; Sharma et al., 2018). For example, the CC filtering pipeline explicitly checks for overlaps between caption tokens and objects identified in the image. The `nocaps` benchmark (Agrawal et al., 2019) tests models’ ability to generalise to out-of-domain objects. There are several V&L datasets and tasks which introduce knowledge-rich annotations and address models’ ability to reason with linguistic and visual cues (Zellers et al., 2019, 2018; Suhr et al., 2017, 2019; Park et al., 2020; Pezzelle et al., 2020). In this paper, we take this line of work further by introducing the novel HL-Scenes dataset, which pairs object-centric and scene-level captions to images.

¹At the time of this work, three OSCAR-based models (OSCAR, VinVL, LEMON) are among the top 5 in the [leaderboard](#) of the COCO image captioning task.

Models Transformer-based V&L models are usually divided into *single-stream* (Li et al., 2020a; Chen et al., 2020; Li et al., 2020b; Su et al., 2020) and *dual-stream* (Tan and Bansal, 2019; Lu et al., 2019; Radford et al., 2021) architectures. It has been shown that single- and dual- stream models perform roughly at par under the same training settings (Bugliarello et al., 2021). On the other hand Hendricks et al. (2021) showed that model performance is highly impacted by dataset curation, attention, and loss function definition.

Most V&L single-stream models are inspired by BERT (Devlin et al., 2019). They incorporate the visual modality in the form of features extracted using a visual backbone, typically a Faster-RCNN (Ren et al., 2015) pre-trained on an object labelling task such as ImageNet (Deng et al., 2009; Russakovsky et al., 2015). From the perspective of caption generation, the Oscar (Li et al., 2020b) single-stream architecture has emerged as an influential model. Oscar enforces grounding between image-caption pairs by using object labels as anchor points (a strategy also adopted by Hu et al., 2021). This makes it particularly suited to the goals of this paper, namely, in-depth analysis of the cross-modal interactions in the treatment of objects during generation. Oscar and its successors, VinVL (Zhang et al., 2021) and LEMON (Hu et al., 2022) achieved SOTA performance on captioning tasks such as COCO and `nocaps`.

Methods In this paper, we focus on three techniques for model analysis: attention analysis, multi-modal ablation and probing. Analyses of attention in pre-trained V&L models include both quantitative methods (e.g. Abnar and Zuidema, 2020) and qualitative analysis (e.g. Li et al., 2020a; Wei et al., 2021). We use both methods to study how VinVL deploys attention during the generation, of object-centric, versus scene-level captions (Section 5).

Several methods have been proposed to study the extent to which V&L models exploit both visual and textual information (Shekhar et al., 2017; Parcalabescu et al., 2022; Gat et al., 2021; Hessel and Lee, 2020). Ablation methods analyse model behaviour when portions of the input are masked or deleted (Bugliarello et al., 2021; Cafagna et al., 2021). We use the ablation of diagnostic objects in scenes (Section 6), to study the reliance of VinVL on such objects during scene-level caption generation.

Probes are well-suited to test for the presence

of task-relevant information in model representations (Belinkov and Glass, 2019; Belinkov, 2022). Cao et al. (2020) develop a probe-based benchmark centred around different V&L tasks. Salin et al. (2022) analyse models’ reliance on text versus vision to capture colour information. Hendricks and Nematzadeh (2021) rely on probes to study lexical and syntactic understanding in V&L models. In our approach, similar in spirit, we develop probes to identify and measure the extent to which scene information is present in the model’s representations before and after fine-tuning on scene-level caption generation.

3 Data

We developed the new High Level Scenes (HL-scenes) dataset, which is explicitly designed to pair images with both object-centric and scene-level captions. To this end, we sampled 15k images from the 2014 COCO train split (Chen et al., 2015), with the constraint that each image depicts at least one person. Captions in COCO are highly object-centric (Lin et al., 2014). We crowd-sourced three scene-level annotations per image on Amazon Mechanical Turk², from workers with at least an 85% approval rating. Crowd workers saw an image and wrote a description in response to the question: *Where is the picture taken?* Annotators were encouraged to use their knowledge of typical scenes in writing their descriptions. Finally, we paired our scene-level HL captions with the previously available COCO (Lin et al., 2014) captions.

Figure 2 shows an example of an image with the two types of captions. See Appendix E for more examples. We collected a total of 14,997 image-caption pairs, and we reserve 11,999 for training and 1,499 each for validation and testing.

4 Model

VinVL (Zhang et al., 2021) is a single-stream BERT-based model with a Faster-RCNN (Ren et al., 2015) visual backbone. It is an extension of Oscar (Li et al., 2020b). VinVL implements a training strategy where object tags are used as anchor points between the visual and textual modality to facilitate cross-modal alignment. As pointed out by Li et al. (2020b), this strategy is motivated by the fact that in the datasets used to pre-train multimodal

²Workers were paid at the rate of €0.03 per item, an amount we consider equitable for the work involved, and in line with rates for similar tasks.



COCO

Reference: a close-up of a kitten looking at a dog laying in the background.

Generated: a cat and a dog sitting next to each other.

HL-scenes

Reference: in the home.

Generated: the picture is taken in a house.

Figure 2: Scene-level captions in HL-Scenes, with corresponding object-centric COCO caption. The generated captions are outputs from VinVL before and after fine-tuning (see Section 4).

models, between 1 and 3 of the objects detected by the visual backbone are mentioned in the caption. However, the object labels are provided by an off-the-self object detector separately trained on Visual Genome (Krishna et al., 2017). VinVL was pre-trained on a combination of COCO (Chen et al., 2015), Conceptual Captions (Sharma et al., 2018), SBU captions (Ordonez et al., 2011) and Flickr30k (Young et al., 2014), as well as additional VQA data.

VinVL has been shown to perform well on understanding tasks, including VQA, NLVR2, image-text and text-image retrieval (Goyal et al., 2017; Suhr et al., 2019; Lin et al., 2014), and on generative tasks, including COCO (Chen et al., 2015) and nocaps (Agrawal et al., 2019).

In the Oscar family of models, the use of labels as anchors makes the models ideal for our experiments, in that it explicitly enables us to study the interaction between object-level information (captured by labels and visual features) and scene-level description generation.

4.1 Fine-tuning

We first establish that VinVL can generate scene descriptions after fine-tuning, before turning to an in-depth analysis of the model’s attention and internal representations.

We note that since the HL-scenes dataset extends the COCO dataset, the model has been exposed to the images of the HL-scenes dataset during pre-

Epoch.	B4	M	RL	CIDEr	SPICE
2	49.3	29.3	67.1	161.8	32.6
4	49.7	30.1	68.1	168.5	34.0
6	48.5	29.8	67.3	164.9	33.5
8	48.9	30.2	67.6	165.8	33.9
10	49.1	30.4	67.7	168.0	34.4

Table 1: Automatic metrics computed over different epochs on the HL-Scenes validation set. B4: Bleu-4; M: METEOR; RL: ROUGE-L.

training on COCO. On the other hand, the scene descriptions are completely novel.

We fine-tune on scene-descriptions for 10 epochs. We use the standard configuration used by Zhang et al. (2021) for image captioning. At inference time, we fix the maximum generation length to 20 tokens and use a beam size of 5.

VinVL shows a quick adaptation to the scene-level descriptions from the first epoch. This adaptability recalls observations made for other transformer-based generative models (e.g. Brown et al., 2020). We show an example in Figure 2. For completeness, Table 1 reports the automatic evaluation metrics computed on the validation set over 10 epochs. For more details see Appendix A.

5 How does attention to objects change from object-centric to scene-level generation?

We first investigate the model’s self-attention before and after fine-tuning on the scene-level caption generation task.

Method We focus on the self-attention patterns in the first layer, as they are directly connected to the inputs and do not depend on higher-level interactions which might obscure the fundamental changes in attention across the two modalities (visual features and labels) in VinVL. A discussion of attention patterns at higher layers can be found in Appendix (B). We select 100 random samples from the HL-Scenes test-set and extract the attention matrices before and after fine-tuning on scene descriptions. We aggregate the attention values by taking the maximum across all the heads, as it allows us to observe where the model tends to assign a significant amount of attention, giving us a better view of the potential impact of fine-tuning on scene-level captions. VinVL prevents textual inputs from directly interacting with the other modalities during generation; therefore there is no interaction between caption tokens and visual features. On

the other hand, the model includes object tags as anchors and this allows us to study the multimodal interactions between the visual features and these object labels.

VinVL acquires a holistic view of the scene after pre-training Figure 3 is a representative example of self-attention matrices extracted from the pre-trained (3a) and fine-tuned (3b) model with the image in Figure 2. The pre-trained model, which generates an object-centric caption, focuses attention on individual input tokens in the **vision-to-vision**, **vision-to-label** and **label-to-vision** sub-blocks.

After fine-tuning, as the model generates a scene-level caption, the self-attention appears to be more evenly distributed over the inputs (3b). This suggests that when generating scene-level captions, the model leverages a wider range of visual features with less exclusive focus on individual objects or labels.

We perform a quantitative analysis of the self-attention in the sub-blocks of the matrix involving visual regions and object labels, computing a kernel density estimate of the distributions of the standard deviations and attention masses for each of the 100 samples. The result is shown in Figure 4. It is clear that the fine-tuned model has overall a lower standard deviation than the pre-trained model. This confirms that a similar attention mass is distributed more evenly after fine-tuning. We take this as evidence that in the process of generating scene descriptions, the fine-tuned model acquires a more holistic view of the input image, in contrast to the highly object-centred deployment of attentional resources evident in the pre-trained model.

VinVL relies on diagnostic objects when generating scene-level captions VinVL redistributes self-attention over a wider range of visual features after fine-tuning. Nevertheless, previous work on scene perception (Self et al., 2019; Vö, 2021) leads us to expect that in describing a scene, the model needs to rely on highly diagnostic objects. We compute diagnosticity empirically, based on the occurrence of objects in scenes in our dataset. Let S be the set of the k most frequent scene types mentioned in scene-level captions in the HL-Scenes dataset.³ We proceed as follows:

1. $\forall s \in S$ we build $O_M^s = [o_1^s, o_2^s, \dots, o_n^s]$, the ranked list of the n most attended objects by

³Since our dataset consists of captions, we extract scene labels from these captions. See Appendix (B).

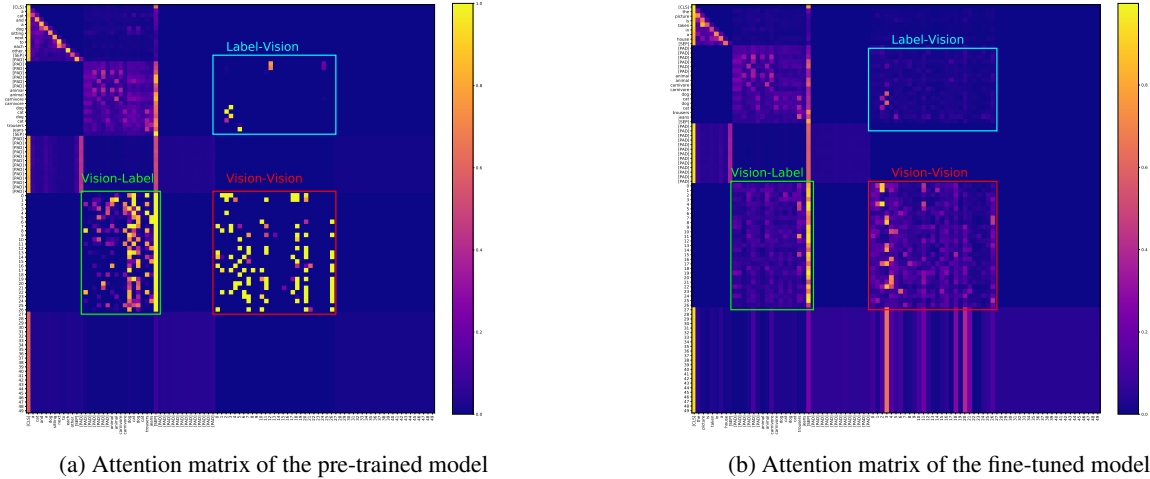


Figure 3: Attention matrices comparison for the image in Figure 2. We highlight the sub-blocks corresponding to **vision-to-vision**, **vision-to-label** and **label-to-vision**. In the pre-trained model, attention mass is sharply focused on individual portions of the input; after fine-tuning, a more even distribution is observed.

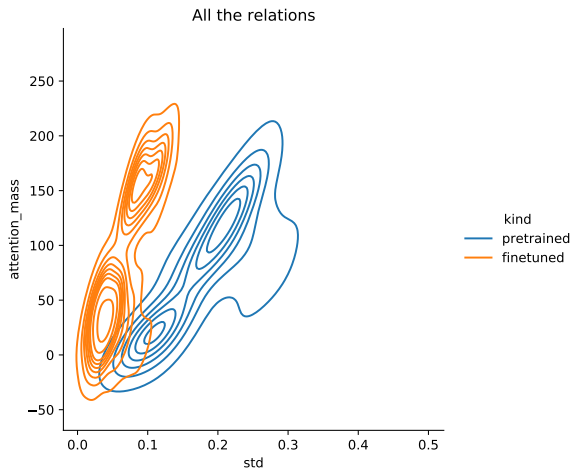


Figure 4: Kernel density estimate of distributions of standard deviations against attention mass for pre-trained and fine-tuned VinVL.

the model M when generating a description of a scene of type s .

2. Similarly, $\forall s \in S$ we collect $O_D^s = [o_1^s, o_2^s, \dots, o_n^s]$, the ranked list of the most frequent objects in images depicting scenes of type s in the dataset D .

We measure the overlap between O_M^s and O_D^s by computing their Intersection over Union (IoU), which is only sensitive to overlap in content, as well as their Rank Biased Overlap (RBO; Webber et al., 2010)⁴, which computes the similarity of two ranked lists. More details about this metric are given in Appendix B. Table 2 shows RBO

⁴<https://github.com/changyaochen/rbo>

Scene	RBO @			IoU @		
	3	5	7	3	5	7
station	0.88	0.84	0.87	0.5	0.66	1.0
road	1.0	0.9	0.91	1.0	0.66	1.0
room	0.27	0.25	0.24	0.2	0.11	0.18
sea	0.88	0.84	0.8	0.5	0.66	0.55
resort	0.72	0.7	0.7	0.5	0.42	0.55
house	0.38	0.5	0.53	0.5	0.42	0.55
restaurant	0.55	0.55	0.54	0.5	0.42	0.53

Table 2: Rank Biased Overlap (RBO) and Intersection over Union (IoU) of the most attended objects and the most frequent objects for the top seven common scenes. Both metrics range from 0 (no overlap) to 1 (perfect correspondence).

and IoU for the top 3, 5 and 7 objects in the lists. We observe that the two metrics correlate strongly ($r(19) = .81, p < .001$). From this we conclude that during generation of scene-level captions, the model attends more to diagnostic objects, i.e. those which are common in a scene of a given type. Moreover, we observe high scores for scene types such as *station*, *road*, *resort*, *sea*. In our dataset, these are characterised by frequently occurring objects, which are therefore highly diagnostic of scene type. In contrast, for scenes like *room*, *house*, *restaurant* we observe lower scores. We hypothesise that this is due to the fact that such scenes can contain a wider variety of objects, which individually have lower diagnosticity with respect to the scene type.

6 How reliant is the model on diagnostic objects?

The results from the previous sections established that, following fine-tuning on scene-level descrip-

tions, VinVL distributes attention more evenly over objects in a scene. Nevertheless, the objects which are most likely to be present in a scene attract the highest proportion of the attention mass. This raises the question whether, by removing highly diagnostic objects from an image, the model representations are still informative enough to detect what type of scene is represented in an image.

We first address this issue from the perspective of generation: does a model fine-tuned on scene descriptions still manage to correctly describe a picture at the scene level, when highly diagnostic objects are unavailable? Given the more even distribution of attention observed across scene components in the fine-tuned model, our hypothesis would be that even in the absence of such highly diagnostic objects, the model can rely on other information to detect the scene type. Hence, we expect the fine-tuned model to be more robust to object ablation in the visual modality, compared to the model pre-trained on object-level captions.

6.1 Method

As explained in Section 4, in VinVL, two separate models are used to (i) extract visual features corresponding to regions via the model’s visual backbone; and (ii) to determine the object labels that function as anchors between the visual and textual modalities. This means we do not have an exact correspondence between object labels and visual features.

Visual feature tagging For simplicity we will refer to vf as the bounding box a visual feature corresponds to, and ot as the bounding box an object label corresponds to. To perform an ablation, we first establish an approximate correspondence between ot and vf , using ot as reference to assign an object label to the visual features.

We compute the IoU⁵ between vf and ot and empirically assign a label to a visual feature if $IoU(vf, ot) \geq 0.6$. Moreover, if vf is contained by or overlaps with ot by at least 80% of its area, we assign to vf the label of ot . With this heuristic we cover 74% of the visual features of every image of our sample.

Computing object diagnosticity We use the scene labels extracted from captions in Section 5,

⁵Note that in this section we refer to the Intersection Over Union to compute the overlap between two bounding boxes, not the metric used to compute the overlap between two sets of items as done in Section 5.

Scene	Top informative objects
restaurant	french fries, fork, submarine sandwich
road	vehicle number plate, traffic sign, traffic light
sea	surfboard, watercraft, boat
room	computer mouse, nightstand, tablet computer
station	train, suitcase, luggage and bags

Table 3: Most informative objects for some scenes ranked using PMI.

the picture is **shot in a ski resort** → the picture is **taken in a snowfield** (*jacket, tree, footwear*)
 the picture is **shot in a baseball field** → the picture is **taken in a ground** (*sports uniform, man, boy*)
 in **a kitchen** → in **the kitchen** (*kitchen appliance, countertop, cabinetry*)

Figure 5: Changes to scene-level captions generated by the fine-tuned model after ablation of three diagnostic objects. Ablated objects are shown in parentheses.

and compute the Pointwise Mutual Information (PMI) between scene types and object labels. Examples of the most informative objects for some scenes are shown in Table 3.

Ablation Ablation of an object is performed similarly to (Frank et al., 2021), by removing its corresponding label from the list of object tags, along with every visual feature assigned to that object. We replace them with a [PAD] token. We compare captions generated by both the pre-trained and fine-tuned model with and without ablation of the top 1, 2 and 3 most informative objects for a given scene in the test-set. For more details on the sample sizes see Appendix C.

6.2 Results

We expect to observe some differences in the generations when ablation is applied, especially in the pre-trained model, as the ablation removes information which is explicitly verbalised in object-centric captions. For the pre-trained model, object-centric captions change 41% of the time after ablation, compared to 13% of the time for the scene-level captions by the fine-tuned model.

A manual inspection on a sample of items suggested that the changes in the captions involve minimal semantic shifts, often due to minor function word changes or a more generic term being generated for the noun denoting the scene type. Some examples are shown in Figure 5.

In summary, the model is resilient to ablation in the visual modality, suggesting that its representations are robust for both types of generation task,

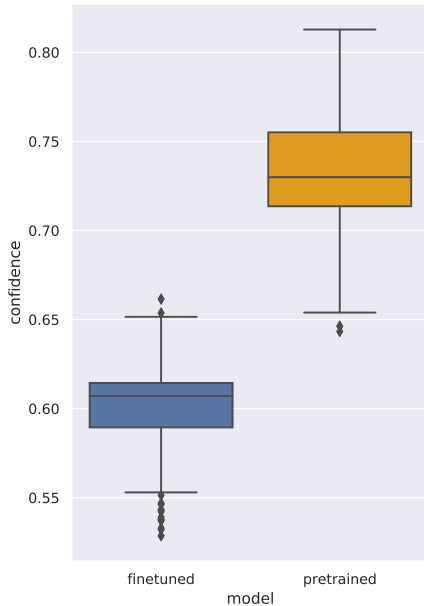


Figure 6: Confidence scores of the unchanged caption after ablation. On the left, the model generating scene-level descriptions (fine-tuned); on the right, the model generating objective descriptions (pre-trained).

but more so for scene-level captioning. We study robustness of representations in more detail using probes, in Section 7.

Confidence scores We also analyse the confidence score produced at generation time by the model for those captions which do not change after ablation. As shown in Figure 6, after ablation pre-trained VinVL generates object-centric descriptions with higher confidence than fine-tuned VinVL does with scene-level descriptions. However, the variance in the confidence score after ablation is lower for the fine-tuned model generating scene-level captions (Figure 7), suggesting greater robustness to ablation during scene-level caption generation.

7 Can we disentangle the role of attention and model representation?

The results so far suggest that there are significant changes in the model’s self-attention, though it relies on diagnostic objects to generate scene-level captions. It is also somewhat more robust to object ablation, especially in the fine-tuned case. At this point, we probe the model’s representations to address to what extent the knowledge required for scene-level caption generation is already present after pre-training. This would imply that the primary change to the model after fine-tuning is in the

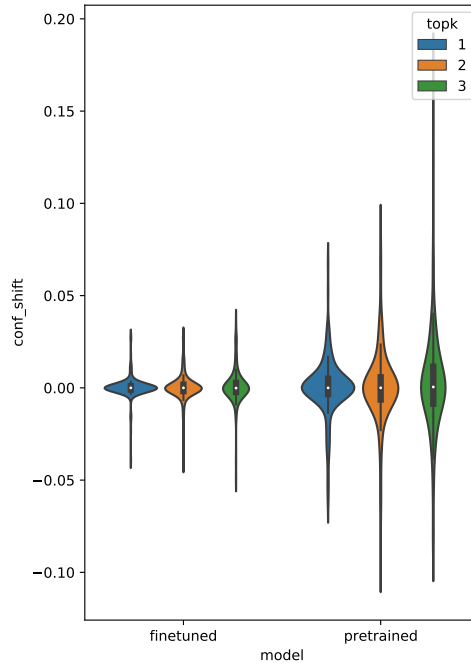


Figure 7: Confidence shift of the unchanged captions when ablating the top 1, 2 and 3 most informative objects from the scene. A negative shift means that the caption was generated with higher confidence after ablation. On the left, the model generating scene-descriptions (fine-tuned); on the right, the model generating object-centric descriptions (pre-trained).

self-attention mechanism.

Method Given a pair (V, L) consisting of visual features V and object labels L , we train a probe to classify scene type based on VinVL encodings, before and after fine-tuning. We also repeat the procedure on inputs ablated as described in Section 6. For this experiment, we identify 1426 images from HL-scenes, representing 8 types of scene, downsampling the more frequent classes (see Appendix D for details). The class distribution is shown in Figure 8. For every image in the probing dataset we extract the model’s feature representations from the last layer and we average across the inputs, obtaining a single vector.

We train both a neural and a random forest probe. We report results from the latter which is the best performing; full details of the neural probe are in Appendix D.

Results Probes are tested on different train/test proportions, up to a 50/50 split. In Figure 9 we report results for the 50/50 train/test split, which is also the most challenging (for results on other splits see Appendix D). The baseline performs a

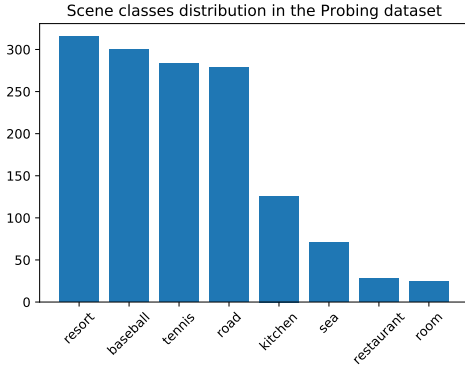


Figure 8: Scene distribution in the probing dataset

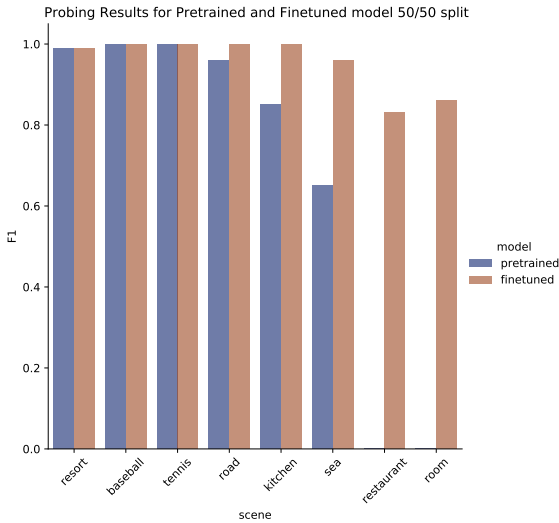


Figure 9: F1-scores of the scene classification task for the pre-trained in (blue) and the fine-tuned model (orange).

random assignment of the labels to the features. For both pre-trained and fine-tuned models, probes perform at ceiling for scenes with a high support (cf. Figure 8). For scene types with a very low frequency, like *restaurant* and *room*, the probe trained on features from the pre-trained model fails. In contrast, probing features from the fine-tuned model still performs at ceiling. These results suggest that the information to detect the scene type is already present to some extent in the pre-trained model. Nevertheless, fine-tuning proves effective in closing the gap for low-support scenes.

When trained on features extracted from ablated inputs in Table 4, the probe is not particularly affected by the ablation, confirming the robustness of the model’s representations as observed in the ablation study (Section 6).

Model	micro-F1	macro-F1	weighted-F1
Random	0.16	0.12	0.16
Pretrained	0.94	0.67	0.92
Finetuned	0.99	0.96	0.99
Pretrained (A)	0.92	0.66	0.90
Finetuned (A)	0.98	0.88	0.97

Table 4: F1-scores for the scene classification task in the 50/50 split using a random forest. The first row (Random) corresponds to the performance of random baseline while (A) is the performance on the features obtained by the ablating the input.

8 Conclusion

In this paper, we addressed scene-level caption generation. Taking a cue from prior work on scene semantics and syntax, our goal was to assess V&L models’ ability to reason about the link between scenes and their components and exploit this to generate informative captions with less redundancy.

Findings and Contributions We contributed a new dataset pairing object-centric and scene-level captions, and showed that VinVL is able to generate scene-level descriptions with minimal fine-tuning.

Our analysis showed that the fine-tuning results in a more even distribution of attention mass over the image, suggesting a more ‘holistic’ view of the scene which nevertheless makes use of diagnostic object information. Using a combination of ablation and probing methods, we also show that much of the relevant information for scene-level captioning is present after pre-training. Hence, the model’s ability to generate scene-level captions is primarily acquired through a change in its self-attention.

Limitations In this work we draw conclusions from an analysis of a single model, this can be considered a limitation. Nevertheless, VinVL is representative of a larger family of SOTA models in the field, based on Oscar, which are dominating the scene in V&L tasks. Moreover, Oscar pretraining using object tags makes the model well-suited to an in-depth analysis of cross-modal interactions in a generative context.

We acknowledge also that the results of the ablation analysis (Section 6) could in part be affected by the approximate nature of our tagging method. Furthermore, as noted by Frank et al. (2021), visual feature deletion may still leave relevant contextual information in the remaining feature vectors, due to the Faster-RCNN’s wide field of view.

Acknowledgements

Contribution from the ITN project NL4XAI (*Natural Language for Explainable AI*). This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860621. This document reflects the views of the author(s) and does not necessarily reflect the views or policy of the European Commission. The REA cannot be held responsible for any use that may be made of the information this document contains.

References

- Samira Abnar and Willem Zuidema. 2020. [Quantifying attention flow in transformers](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.
- Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. No-caps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8948–8957.
- Yonatan Belinkov. 2022. [Probing classifiers: Promises, shortcomings, and advances](#). *Computational Linguistics*, 48(1):207–219.
- Yonatan Belinkov and James Glass. 2019. [Analysis Methods in Neural Language Processing: A Survey](#). *Transactions of the Association for Computational Linguistics*, 7:49–72.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language bert. *Transactions of the Association for Computational Linguistics*, 9:978–994.
- Michele Cafagna, Kees van Deemter, Albert Gatt, et al. 2021. [What vision-language models ‘see’ when they see scenes](#). *ArXiv preprint 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. In *European Conference on Computer Vision*, pages 565–580. Springer.
- Xinlei Chen, Hao Fang, Tsung-yi Lin, Ramakrishna Vedantam, C Lawrence Zitnick, Saurabh Gupta, and Piotr Doll. 2015. [Microsoft COCO Captions : Data Collection and Evaluation Server](#). *arXiv preprint 1504.00325*, pages 1–7.
- 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. In *Computer Vision – ECCV 2020*, pages 104–120, Cham. Springer International Publishing.
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. [What does BERT look at? an analysis of BERT’s attention](#). In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Minneapolis, Minnesota. Association for Computational Linguistics.
- Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. [Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Itai Gat, Idan Schwartz, and Alexander Schwing. 2021. [Perceptual Score: What Data Modalities Does Your Model Perceive?](#) In *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, Sydney, Australia.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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*, pages 6904–6913.
- Herbert P Grice. 1975. Logic and conversation. In Peter Cole and Jerry Morgan (eds), editors, *Speech acts*, pages 41–58. New York: Academic Press.
- Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P. Bigham. 2019.

- Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 939–948.
- John M. Henderson and Taylor R. Hayes. 2017. [Meaning-based guidance of attention in scenes as revealed by meaning maps](#). *Nature Human Behaviour*, 1:743–747.
- Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. 2021. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*, 9:570–585.
- Lisa Anne Hendricks and Aida Nematzadeh. 2021. [Probing image-language transformers for verb understanding](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, Online. Association for Computational Linguistics.
- Jack Hessel and Lillian Lee. 2020. [Does my multimodal model learn cross-modal interactions? It’s harder to tell than you might think!](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP’20)*, pages 861–877, Online. Association for Computational Linguistics.
- Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. [Framing image description as a ranking task: Data, models and evaluation metrics](#). *Journal of Artificial Intelligence Research*, 47:853–899.
- Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17980–17989.
- Xiaowei Hu, Xi Yin, Kevin Lin, Lei Zhang, Jianfeng Gao, Lijuan Wang, and Zicheng Liu. 2021. [Vivo: Visual vocabulary pre-training for novel object captioning](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1575–1583.
- Laurent Itti and Christof Koch. 2001. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020a. [What does BERT with vision look at?](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, Online. Association for Computational Linguistics.
- Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020b. Oscar: Object-semantic aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.
- George L. Malcolm, Iris I.A. Groen, and Chris I. Baker. 2016. [Making Sense of Real-World Scenes](#). *Trends in Cognitive Sciences*, 20(11):843–856.
- Aude Oliva and Antonio Torralba. 2007. [The role of context in object recognition](#). *Trends in cognitive sciences*, 11(12):520–527.
- Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. [Im2text: Describing images using 1 million captioned photographs](#). In *Proceedings of the 2011 Conference on Advances in Neural Information Processing Systems (NIPS’11)*, pages 1143–1151, Granada, Spain. Curran Associates Ltd.
- Luis Gilberto Mateos Ortiz, Clemens Wolff, and Mirella Lapata. 2015. Learning to Interpret and Describe Abstract Scenes. In *Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL (NAACL’15)*, pages 1505–1515, Denver, Colorado. Association for Computational Linguistics.
- Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. [VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland. Association for Computational Linguistics.
- Jae Sung Park, Chandra Bhagavatula, Roozbeh Motlaghi, Ali Farhadi, and Yejin Choi. 2020. [Visual-COMET: Reasoning About the Dynamic Context of a Still Image](#). In *Proceedings of the European Conference on Computer Vision*, pages 508–524, Berlin and Heidelberg. Springer.
- Sandro Pezzelle, Claudio Greco, Greta Gandolfi, Eleonora Gualdoni, and Raffaella Bernardi. 2020. [Be Different to Be Better! A Benchmark to Leverage the Complementarity of Language and Vision](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2751–2767, Online. Association for Computational Linguistics.

- Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *European Conference on Computer Vision*, pages 647–664. Springer.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.
- Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. 2022. Are Vision-Language Transformers Learning Multimodal Representations? A probing perspective. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, Vancouver, BC. Association for the Advancement of Artificial Intelligence.
- Julie S. Self, Jamie Siegart, Munashe Machoko, Enton Lam, and Michelle R Greene. 2019. Diagnostic Objects Contribute to Late – But Not Early– Visual Scene Processing. *Journal of Vision*, 19:227.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. **Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning**. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL’18)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.
- Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. **FOIL it! Find One mismatch between Image and Language caption**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL’17)*, pages 255–265, Vancouver, BC. Association for Computational Linguistics.
- 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**. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. **A Corpus of Natural Language for Visual Reasoning**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL’17)*, pages 217–223, Vancouver, BC. Association for Computational Linguistics.
- Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. **A corpus for reasoning about natural language grounded in photographs**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.
- Hao Tan and Mohit Bansal. 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, Hong Kong, China. Association for Computational Linguistics.
- Antonio Torralba, Aude Oliva, Monica S. Castelhana, and John M. Henderson. 2006. **Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search**. *Psychological Review*, 113(4):766–786.
- Melissa Le Hoa Võ. 2021. **The meaning and structure of scenes**. *Vision Research*, 181:10–20.
- William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38.
- Haiyang Wei, Zhixin Li, Feicheng Huang, Canlong Zhang, Huifang Ma, and Zhongzhi Shi. 2021. Integrating scene semantic knowledge into image captioning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(2):1–22.
- Chia Chien Wu, Farahnaz Ahmed Wick, and Marc Pomplun. 2014. **Guidance of visual attention by semantic information in real-world scenes**. *Frontiers in Psychology*, 5.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. **From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions**. *Transactions of the Association for Computational Linguistics*, 2:67–78.
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6720–6731.
- Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. **SWAG: A large-scale adversarial dataset for grounded commonsense inference**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.

Appendix

A Fine-tuning Details

We fine-tune the VinVL pre-trained base version⁶ using the original configuration for 10 epochs on scene descriptions. We refer to it as the *fine-tuned* model. Since the HL-scenes dataset images are included in COCO, we use the pre-computed visual features and labels provided in the original VinVL implementation.

We refer to the *pre-trained* model, as the base model trained on the image captioning task on COCO captions optimized using cross-entropy. All the experiments involving the pre-trained model are performed using the original configuration used in Li et al. (2020b). The fine-tuning is carried out with batch size 32 on a NVIDIA GTX 2080 TI 11 GB.

B Self-attention Details

Attention beyond Layer 1 At higher layers the attention converges on the special token [SEP], used to separate the *text + object tags* from the *visual* input, as shown in Figure 10. A similar behaviour has been observed analysing BERT’s attention (Clark et al., 2019).

Figure 11 shows how this pattern becomes more pronounced as we move further across the layers, preventing from observing any kind of input interplay. Although the *text*, *object tags* and *visual* sequences can be of different lengths, the [SEP] token sits always in the same position among the inputs, as the padding is always applied to keep the *text + object tags* sequence of the same length. We believe that this regularity is used by the model as a sort of pivot among the inputs. This can cause the a high accumulation of attentional resources by the model.

Scene label extraction As described in Section 3, during the data collection, the annotators where asked to answer the direct question: *Where is the picture taken?* As a consequence, the scene-captions often have a regular structure, captured by the following three representative examples:

- the picture has been taken in a *restaurant*
- on a *beach*

⁶https://github.com/microsoft/Oscar/blob/master/VinVL_MODEL_ZOO.md#Oscarplus-pretraining

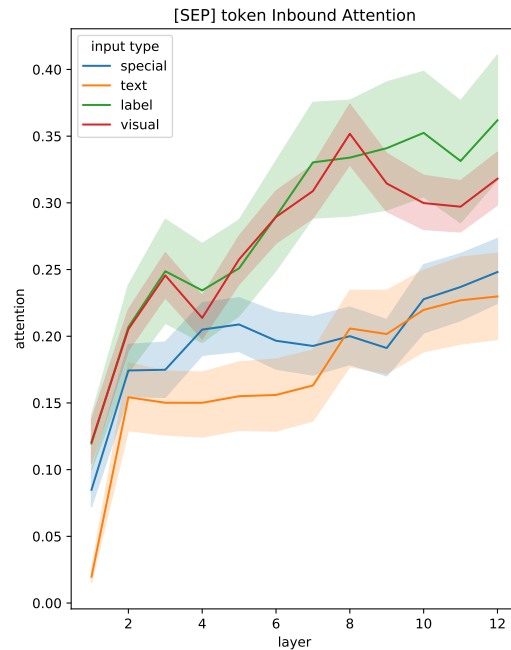


Figure 10: Inbound attention of the [SEP] per input type token across the layers. Special tokens correspond to [CLS], [PAD] and [SEP].

- this is in an *airport*

To extract the scene labels, we tokenize the scene-captions and we remove punctuation and stop-words (we add *picture* to the list of the standard stop-words). Among the remaining tokens, we extract all the nouns and we reduce them to lemmas, then we compute the frequencies of the remaining tokens. This allow us to extract the scene-types (*restaurant*, *beach* and *airport*) from the captions, such as those shown in the examples above. The whole procedure is performed using spaCy.⁷

Rank Biased Overlap RBO (Webber et al., 2010) computes the similarity of two ranked lists, as follows:

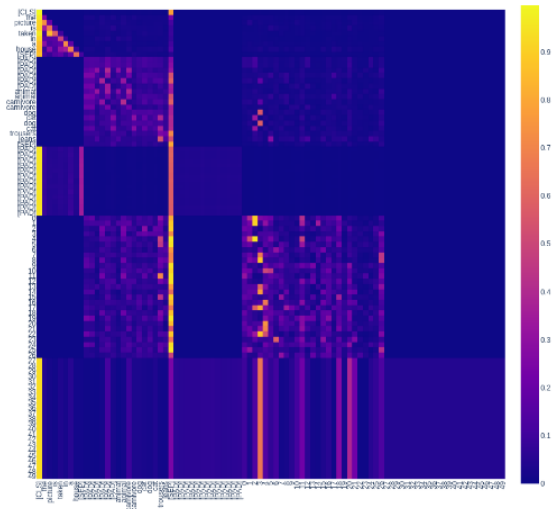
$$RBO(S, T, p) = (1 - p) \sum p^{d-1} A_d \quad (1)$$

where d is the depth of the ranking being examined, A_d is the agreement between S and T given by the proportion of the size of the overlap up to d , and p determines the contribution of the top d ranks to the final RBO measure. We use the standard value of $p = 1$.

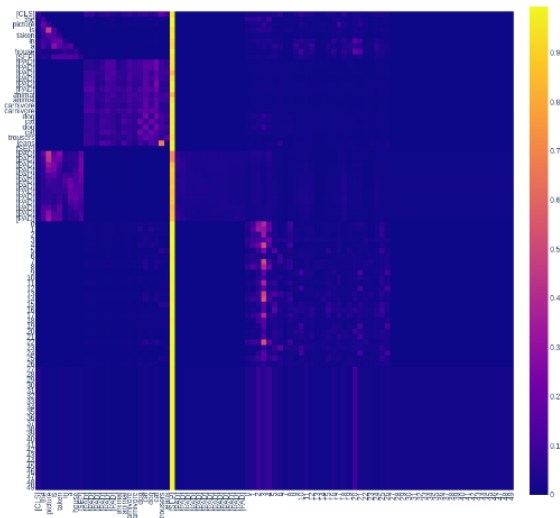
C Ablation Details

As described in Section 6 the ablation is performed by removing the most informative objects from

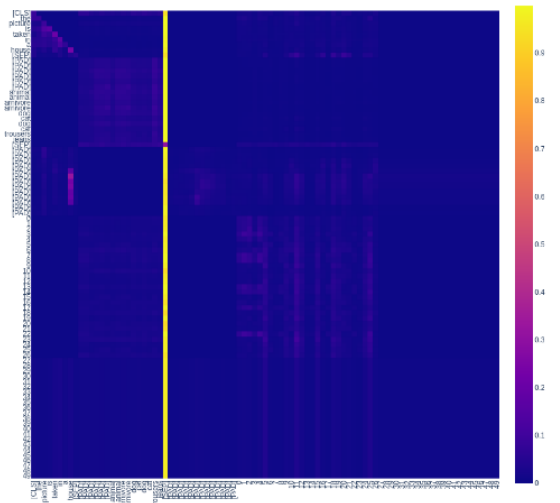
⁷<https://pypi.org/project/spacy/>



(a) Layer 1



(b) Layer 6



(c) Layer 12

Figure 11: Attention matrices for layers 1, 6 and 12. The attention weights progressively gather on the `[SEP]` token.

# Ablation	Train-Val	Test
no ablation	13498	1499
1	4269	469
2	2565	274
3	1554	170

Table 5: Sample size of the Train-Val and Test split after ablation of the top 1,2 and 3 most informative objects in the most frequent scenes. The top row corresponds to the original dataset split sizes.

images depicting the most frequent scene types. As a result, an image is included in the ablation study if (i) it belongs to the set of most frequent scenes; and (ii) it contains the objects we want to ablate. This means that the higher the number of objects ablated, the smaller the sample of images matching these constraints. As shown in Table 5, with 3 objects ablated in the test-set we obtain 170 valid images.

We repeat the ablation experiment on both the test and the train-val split. The results obtained on the latter mirror those reported in Section 6 with the test-split only. In Figure 12 we show the comparison of the distributions of the unchanged confidence scores after ablation for the test and train-val split. Moreover, there is no statistically significant difference between the distributions of confidence score shifts of the test set (shown in Figure 7) and the train-val set ($z = 0.13$ with $p = 0.89$ and $\alpha = 0.05$).

D Probing details

Model selection We test two probing models: a multi-layer perceptron and a random forest. We perform hyperparameter tuning of the neural probe by carrying out a random search followed by a probabilistic search. The tuned neural probe is a three-layer feed-forward network with *hidden size* 16, optimized using LBFGS with adaptive learning rate and $\alpha = 1$. Note that no parameter tuning is required for the random forest. As reported in Table 6, the random forest performs better or on a par with the neural probe. Therefore we report the performance of the random forest in the main results in Section 7.

Challenging the probe The probing model performs at ceiling with the more typical 90/10 split, especially when trained on the fine-tuned features (Figure 13). Therefore, we perform multiple experiments for different train/test splits namely, 90/10,

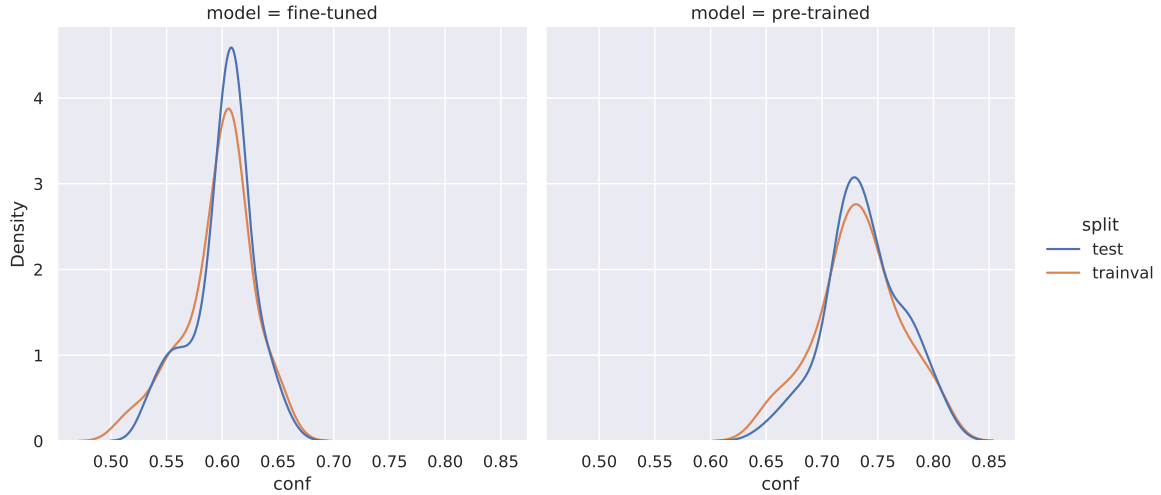


Figure 12: Kernel density estimate of the confidence scores distributions of unchanged captions after ablation for the test (blue) and train-val (orange) split.

Probe	Model	micro-F1	macro-F1	weighted-F1
RB		0.16	0.12	0.16
RF	PRE	0.94	0.67	0.92
	FT	0.99	0.96	0.99
	PRE (A)	0.92	0.66	0.90
	FT (A)	0.98	0.88	0.97
MLP	PRE	0.94	0.67	0.91
	FT	0.98	0.91	0.98
	PRE (A)	0.92	0.66	0.90
	FT (A)	0.98	0.85	0.97

Table 6: F1-scores of scene classification task in the 50/50 split, for Random Baseline (RB), Random Forest (RF) and Multilayer perception (MLP) trained on encodings extracted from the pre-trained (PRE) and fine-tuned (FT) model without and with ablation (A). In bold the best result for each setting.

70/30 and 50/50. The 50/50 is the most challenging for the probe and it allows us to highlight the performance gap across different settings. Results from all the splits are shown in Table 7.

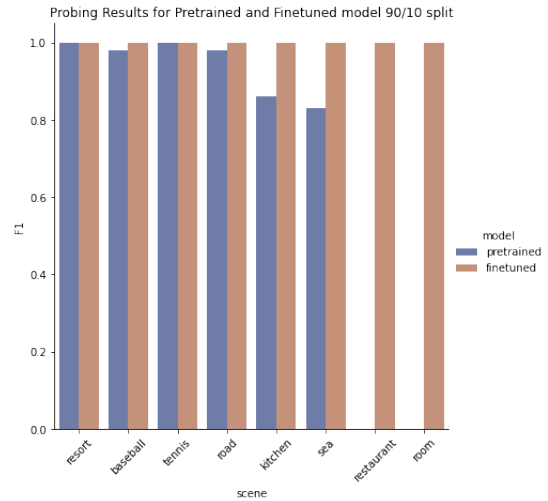


Figure 13: F1-scores of the scene classification task for the pre-trained (blue) and the fine-tuned model (orange) for the 90/10 split.

E HL-Scenes examples

Split	Model	micro-F1	macro-F1	weighted-F1
90/10	PRE	0.96	0.71	0.94
	FT	1.0	1.0	1.0
	PRE (A)	0.95	0.69	0.94
	FT (A)	0.99	0.99	0.99
70/30	PRE	0.94	0.67	0.92
	FT	0.99	0.97	0.99
	PRE (A)	0.93	0.66	0.91
	FT (A)	0.98	0.94	0.98
50/50	PRE	0.94	0.67	0.92
	FT	0.99	0.96	0.99
	PRE (A)	0.92	0.66	0.90
	FT (A)	0.98	0.88	0.97

Table 7: F1-scores for scene classification task the random forest in different train/tes splits. The random forest is trained on encodings extracted from the pre-trained (PRE) and fine-tuned (FT) model without and with ablation (A).



Image	Object (COCO)	description	Scene description (HL-Scenes)
		a woman and a boy sitting in the snow outside of a cabin.	the picture is shot in a ski resort
		a airplane with a group of people standing next to it.	the picture is shot in an airport
		a man holds his hands up as he stands over a trash can.	the picture is taken in front of a roadside toilet
		a couple of people that are skateboarding on a ramp	it is at the park.

Table 8: Randomly selected images from the HL-scenes dataset. For both COCO and HL-Scenes we show a randomly picked caption among the the available ones for the image.