Are Vision-Language Transformers Learning Multimodal Representations? A Probing Perspective.

Emmanuelle Salin[1], Badreddine Farah[2], Stéphane Ayache[1], Benoit Favre[1]
[1]Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
[2]École Sup Galilée, Université Sorbonne Paris Nord, France
{emmanuelle.salin,sephane.ayache,benoit.favre}@lis-lab.fr, badreddine.farah@edu.univ-paris13.fr

Abstract

In recent years, joint text-image embeddings have significantly improved thanks to the development of transformer-based Vision-Language models. Despite these advances, we still need to better understand the representations produced by those models. In this paper, we compare pre-trained and fine-tuned representations at a vision, language and multimodal level. To that end, we use a set of probing tasks to evaluate the performance of state-of-the-art Vision-Language models and introduce new datasets specifically for multimodal probing. These datasets are carefully designed to address a range of multimodal capabilities while minimizing the potential for models to rely on bias. Although the results confirm the ability of Vision-Language models to understand color at a multimodal level, the models seem to prefer relying on bias in text data for object position and size. On semantically adversarial examples, we find that those models are able to pinpoint fine-grained multimodal differences. Finally, we also notice that fine-tuning a Vision-Language model on multimodal tasks does not necessarily improve its multimodal ability. We make all datasets and code available to replicate experiments.

Introduction

Vision-Language (VL) tasks consist in jointly processing a picture and a text related to the picture. VL tasks, such as visual question answering, cross modal retrieval or generation, are notoriously difficult because of the necessity for models to build sensible multimodal representations that can relate fine-grained elements of the text and the picture. Following the success of pre-trained transformers for language modeling such as BERT (Devlin et al. 2018), the community has proposed various transformer-based models, such as VilBERT (Lu et al. 2019), LXMERT (Tan and Bansal 2019), VLBERT (Su et al. 2019), UNITER (Chen et al. 2020), OSCAR (Li et al. 2020b), VinVL (Zhang et al. 2021), ViLT (Kim, Son, and Kim 2021) or ERNIE-VIL (Yu et al. 2021), that combine representations from both the text and image modalities to reach state-of-the-art results in several multimodal tasks. Similar models have been developed in the field of video-language pre-training, such as ClipBERT (Lei et al. 2021) and HERO (Li et al. 2020a).

While the results are impressive, it is important to understand how multimodal information is encoded in the representations learned by those models, and how affected they are by various bias and properties of their training data. A few studies have been conducted to better understand those models and their representations. (Cao et al. 2020) have probed attention heads at various layers of the models, showing that textual modality is more important than visual modality for model decisions. This prevalence of language over vision in multimodal models is not specific to transformer-based representation models, as noticed by (Goyal et al. 2017a). (Li, Gan, and Liu 2020) have looked into the robustness of the representations to manipulations of the input, compared to more traditional models. (Hendricks and Nematzadeh 2021) relied on probing tasks to study verb understanding in pre-trained transformer-based models and determined that models learn less multimodal concepts associated to verbs than to subjects and objects. While these studies have shed light on some particular aspects of transformer-based VL models, they lack a more systematic analysis of monomodal biases that impede the nature of the learned representations.

In that light, we are interested in studying the multimodal capacity of VL representations, and in exploring what information is learned and forgotten between pre-training and fine-tuning, as we think this could show the current limits of the pre-training process. Inspired by probing tasks developed in the Natural Language Processing field, we probe three VL models: UNITER, LXMERT and ViLT to answer those questions. We probe both pre-trained and fine-tuned models. We propose probing tasks and collect associated datasets to evaluate the monomodal and multimodal capabilities of those models over a range of concepts. We find that UNITER reaches better overall results on the language modality, while ViLT reaches better results on the vision modality. Finally, we notice that while the models show their ability to identify colors, they do not yet have multimodal capacity to distinguish object size and position. We make the set of monomodal and multimodal probing tasks, as well as all software developed for this study, available for further research[1].

[1] https://github.com/ejsalin/vlm-probing

Related Work

Vision-Language models Transformer-based VL models are typically trained from image/caption pairs. The text is tokenized and projected to an embedding space with additive position encoding, similar to BERT. The image is usually encoded using a Faster RCNN (Anderson et al. 2018) to extract a sequence of object-region representations with additive position and shape embeddings. However some recent models, such as ViLT and SOHO (Huang et al. 2021) swap the object-based representations for grid-based representations trained from scratch.

After encoding, the representations from each modality are passed through a transformer following one of two architectures. The single-stream architecture, used in UNITER and ViLT, relies on a single transformer spanning inputs from both modalities. The dual-stream architecture, used in LXMERT, inputs each modality in its own transformer, the outputs of which are fed to a cross-modal transformer.

VL models are pre-trained using text-oriented, image-oriented and cross modal losses. While the text-oriented loss reflects that of language models (Masked Language Modeling, i.e. MLM), the image loss varies from feature regression tasks, to object class prediction tasks such as for VilBERT (Lu et al. 2019). Most models, like ViLT, UNITER and LXMERT adopt the Image Text Matching (ITM) task for cross-modal pre-training. Some models add other pre-training tasks to complete their multimodal knowledge. In addition, some models rely on other specific pre-training tasks such as Visual Question Answering (VQA) in LXMERT, word-region alignment in UNITER, scene graphs in ERNIE-VIL (Yu et al. 2021) or object semantics in OSCAR (Li et al. 2020b).

In order to reach state-of-the-art performances on downstream multimodal tasks such as VQA (Goyal et al. 2017b) and NLVR2 (Suhr et al. 2018), those models need fine-tuned on the target task.

Model probing Probing tasks have been first developed to analyse language models through benchmarks such as SentEval (Conneau and Kiela 2018). For example, (Hewitt and Manning 2019) showed that syntactic parse trees can be inferred from ELMO and BERT representations.

Although explainability has been largely explored in vision models, the use of probing tasks is more limited. Recently, Basaj et al. (2021) have developed a visual probing framework by constructing visual equivalents to words based on superpixels. It then translated language probing tasks such as sentence length and semantic odd man out to the vision modality.

Cao et al. (2020) study how the transformer architecture impacts the learning process of single-stream and dual stream models. They observe the role of each layer as well as the fusion of the vision and language modalities across layers. They in particular notice the prevalence of the language modality over the visual modality, which we study further in our work, by evaluating the multimodal nature of representations. Lindström et al. (2021) explore the use multimodal probing tasks such as object counting, and object identification to study Visual Semantic embeddings. They also notice the importance of linguistic information in multimodal tasks. However, they do not analyse VL transformer-based models in their study. Hendricks and Nematzadeh (2021) rely on probes to study verb understanding in pre-trained transformer-based models. They determine that models learn less multimodal concepts associated to verbs than to subjects and objects. Similar to our study, Shekhar et al. (2017) build a dataset to evaluate if the text and image information in VL models are both deeply integrated. Contrary to this dataset, we do not study the ability of a model to differentiate between objects from the same super-categories, but focus on multimodal concepts such as color, size and position.

Methodology

Framework In this paper, we aim at evaluating VL models at a language, vision and multimodal level through their text-image representations. We write VLM_{pre} a pre-trained transformer-based VL model, such as UNITER, LXMERT or ViLT. This model can be fine-tuned on a task T, for example VQA or NLVR2. Fine-tuning tasks are used to embed new knowledge in the model, and to evaluate more dedicated semantics or abilities related to a specific task. We write $VLM_{fine(T)}$ the VL model fine-tuned on task T.

We use *probing* to study the representations of VLM_{pre} and $VLM_{fine(T)}$ models. To evaluate the representations of VL models on a probing task p, we build a training dataset $S_p = \{(X_j, Y_j)\}_{j=1}^{n_p}$, drawn i.i.d. from $\mathcal{D}^{\mathrm{p}}_{\mathcal{X}\times\mathcal{Y}}$ where $\mathcal{X}$ and $\mathcal{Y}$ relate to the datasets needed to probe the models.

The first step of our method is to compute the final layer representations $VLM_{pre}(X)$ or $VLM_{fine(T)}(X)$ of an instance $X_j = (x_j^{image}, x_j^{caption})$ of S_p through the VL transformer-based model. If the probing task p studies the instance at a global level, we use the representation of the classification token $[CLS]$ as input for p. If p studies the representations of each word, the representation of $WORD$ tokens are used as input for p.

The second step of our method is to use the representations R_j of the $[CLS]$ or $WORD$ tokens as input of a linear probing model PM_p trained using the $\{(R_j, Y_j)\}_{j=1}^{n_p}$ dataset. As VLM_{pre} or $VLM_{fine(T)}$ are not trained on the probing task, the probing model PM_p can only rely on linearly separable information the model has already learned to extract during pre-training or fine-tuning. As a result, the performance of PM_p will reflect the capability of VLM_{pre} and $VLM_{fine(T)}$ models to extract the information needed for the probing task p.

The models VLM and the set of probing tasks P are described in the following sections. Figure 1 illustrates the methodology on the object counting task V-ObjCount of P.

Studying the impact of each modality We also want to study how much VL models rely on the language and vision modalities when building text-image representations. As a result, for each task p build on S_p, we create another corresponding task $p^{\clubsuit}$ with *mismatched* image and caption pairs. The dataset $S_{p^{\clubsuit}}$ is build using S_p by associating the label with the image (resp. caption) and selecting at random a

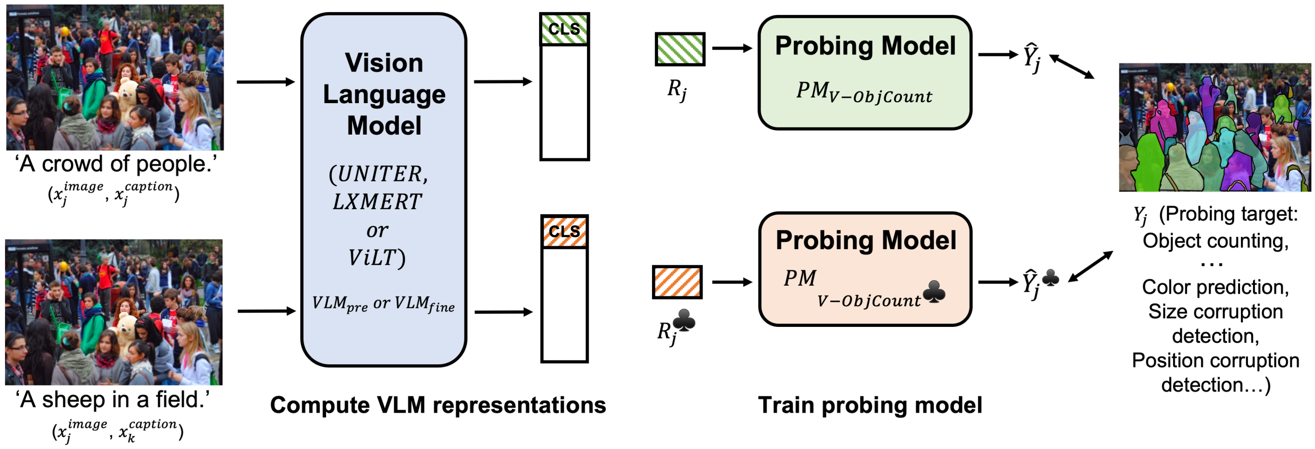

'A crowd of people.'
$(x_j^{image}, x_j^{caption})$
'A sheep in a field.'
$(x_j^{image}, x_k^{caption})$
Vision Language Model
(UNITER, LXMERT or ViLT)
VLM_{pre} or VLM_{fine}
CLS
CLS
Compute VLM representations
R_j
$R_j^{\clubsuit}$
Probing Model
$PM_{V-ObjCount}$
Probing Model
$PM_{V-ObjCount^{\clubsuit}}$
Train probing model
$\hat{Y}_j$
$\hat{Y}_j^{\clubsuit}$
Y_j (Probing target: Object counting,
…
Color prediction, Size corruption detection, Position corruption detection…)

Figure 1: Probing methodology: The first step is to compute the final layer representations of the image/caption input using the chosen Vision Language Model. Then, we use the final layer [CLS] or word token representations R_j to train a linear probing model on the probing task. This example illustrates the methodology using the visual probing task V-ObjCount and an image from MS-COCO. This task consists in counting the number of objects in an image using $(x_j^{image}, x_j^{caption})$ as input and Y_j as target. The notation $\clubsuit$ indicates the corresponding task (V-ObjCount) with mismatched instances (i.e. a caption that does not match the image and label), which uses $(x_j^{image}, x_k^{caption})$ as input and serves as a baseline.

mismatched caption (resp. image). For better comparison, all models use the same mismatched datasets.

If p is a language-oriented task, each instance of $p^{\clubsuit}$ will be $(x_k^{image}, x_j^{caption}, Y_j)$, with the caption corresponding to the label and a wrong image. If the performance of PM_p is similar to the performance of $PM_{p^{\clubsuit}}$, we can deduce that the representations $R_j^{\clubsuit}$ given by VLM are not affected by visual "bias".

Similarly, when p is a vision-oriented task, each instance of $p^{\clubsuit}$ will be $(x_j^{image}, x_k^{caption}, Y_j)$, with the image corresponding to the label and a wrong caption. If the performance of PM_p is similar to the performance of $PM_{p^{\clubsuit}}$, we can deduce the representations $R_j^{\clubsuit}$ given by VLM are not affected by linguistic "bias".

For a multimodal probing task p, as we want to study the presumed prevalence of language over vision in model decisions, each instance of $p^{\clubsuit}$ will be $(x_k^{image}, x_j^{caption}, Y_j)$, with the caption corresponding to the label and a mismatched image. If VLM extracts multimodal information rather than only linguistic information, PM_p should reach better performance than $PM_{p^{\clubsuit}}$. This is a way to control if $PM_{p^{\clubsuit}}$ only uses textual information or if it also uses multimodal information.

Probing Tasks

The set of probing tasks P, summarized in Table 1, is composed of language-oriented tasks L, vision-oriented tasks V, and multimodal tasks M. Each task of the set is built to evaluate the mono-modal or multi-modal performance of VLM on a specific capability. The tasks consist of regression, binary or multiclass classification problems. For each of them, a linear layer is trained as in (Hewitt and Manning 2019).

Ideally, one would probe VL models on all properties that have an impact on down-stream tasks or that help understand their behavior. However, in this paper we restrain ourselves to a few representative language, vision and multimodal properties. We choose and build tasks that are easy to implement on new datasets. For the language and vision properties, we use tasks well understood in past work. For multimodal properties, we create new tasks assessing multimodal properties we think are especially relevant for VL models. We explain the tasks and their choice in the following section.

Language probing tasks: L

For language-oriented probing tasks, we choose already existing language probing tasks and adapt them to a subset of 3,000 instances from Flickr30k (Young et al. 2014). We choose tasks appropriate for the relatively simple structure of captioning datasets, and easy to transfer to a new dataset.

- **Part of Speech Tagging (L-Tagging):** Part of Speech Tagging consists in associating a word with its corresponding part of speech label, such as *verb*. There are 34 categories. This task evaluates the syntactic knowledge present in the representation of individual word tokens. As a result, we train a linear classifier $PM_{\text{L-Tagging}}$ using word token representations given by VLM. To create a gold standard for this task, we annotate the Flickr30k dataset using the *en_core_web_sm* SpaCy tagger (Honnibal and Montani 2017), which performs at 97% accuracy on Ontonotes.
- **Bigram Shift (L-BShift):** Bigram Shift (Conneau and Kiela 2018) consists in determining whether two consecutive words in a sentence have been swapped. For exam-

Task	Description	Input Repr.	Type (Metric)	Dataset	Total Instances	Test Size	Maj. (%)
L-Tagging	Part-of-speech tagging	Word	Multiclass (acc.)	Flickr	3,000	1000	24.06
L-BShift	Bigram shift detection	[CLS]	Binary (acc.)	Flickr	3,000	1,000	50.20
V-Flower	Fine-grained classification	[CLS]	Multiclass (acc.)	Flower-102	7,169	1,020	0.98
V-ObjCount	Object counting	[CLS]	Regression (MSE)	MS-COCO	2,424	624	-
M-Col	Color prediction	[MASK]	Multiclass (acc.)	Flickr	3,000	1,000	25.30
M-Size	Size corruption detection	[CLS]	Binary (acc.)	Flickr	2,552	752	50.67
M-Pos	Position corruption detection	[CLS]	Binary (acc.)	Flickr	2,626	826	53.75
M-Adv	Adversarial captions	[CLS]	Binary (acc.)	MS-COCO	700	200	50.00

Table 1: List of probing tasks. *Repr.* is the representation vector used as input of the probing task. *Instances* indicates the number of image/caption instances used for the probing task. *Maj.* is the majority baseline.

ple, in the sentence “People *at relaxing* the park.”, tokens from the bigram (“relaxing”, “at”) have been swapped to create a negative example caption. As this evaluates the global correctness of a sentence, we use the [CLS] token representation given by VLM.

Vision probing tasks: V

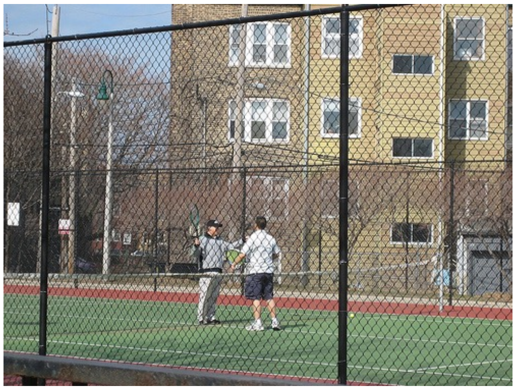
In order to probe the vision capability of the models, we selected two tasks: an object counting task to assess if information on the general structure of an image is present in the representation, and a fine-grained object classification task to evaluate whether the representations also retain information on fine details of objects.

- **Flower identification (V-Flower):** This is a fine-grained object classification task which consists in classifying flower pictures into 102 categories. We use the 102-Flower dataset (Nilsback and Zisserman 2008). As there is no caption available for this task, we use an empty caption. The linear classifier $PM_{\text{V-Flower}}$ uses the representation of the [CLS] token.
- **Object Counting (V-ObjCount):** We build this object counting task on a subset of 3,000 instances of the MS-COCO dataset (Lin et al. 2014). The labels are created by counting the number of objects in its manual annotations. The linear regression model $PM_{\text{V-ObjCount}}$ also uses the [CLS] token representation. As there can be clues in the caption indicating how many objects are in the image, some multimodal information present in the representation can be used for this task, which makes the use of a baseline important.

Multimodal probing tasks: M

To evaluate the multimodal information present in VLM representations, we focus on concepts which are used to describe objects, as those are inherently multimodal. However, as evaluating all those properties can be time-consuming, we restrain ourselves to a few important attributes that matter in many downstream applications: color, size and position. We create datasets to evaluate those attributes.

As we cannot evaluate all multimodal properties, we also assess the general multimodal competency of models, not linked to a specific property. To that aim, in addition to the three attribute-specific tasks, we create a task that assesses how well the model captures linguistically likely differences in multimodal concepts.



Task	Text Input
M-Col	Two men standing behind a tall [MASK] fence.
M-Size	Two men standing behind a short black fence.
M-Pos	Two men standing in front of a tall black fence.
M-Adv	Two men running behind a tall black fence.

Figure 2: Example of modified captions for the multimodal probing tasks, using the caption “Two men standing behind a tall black fence” as original (Flickr30k).

The creation of the four tasks consists in altering the caption of half of the instances to create negative examples and evaluating the performance of a model in distinguishing between positive and negative examples. The probing datasets are carefully designed to avoid textual bias. Figure 2 lists altered captions for the multimodal tasks with an example picture.

We leave as future work the evaluation through other multimodal tasks such as specific object properties as shape or texture, and global image properties such as focus, quality or emotion.

- **Color Identification (M-Col):** This task aims at evaluating precise color understanding, at a multimodal level. To this end, we select 8 common colors that are unlikely to be ambiguous: blue, red, black, white, yellow, orange, green, purple. A subset of 3,000 instances from Flickr30k that contain those colors is used for evaluation. We do not control for text bias, and use the text-only and

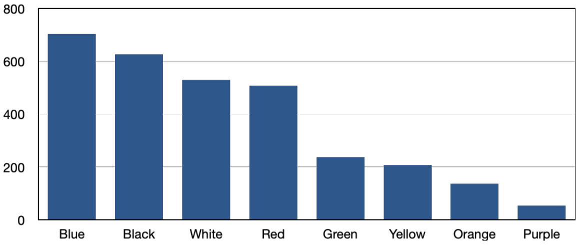


Figure 3: Colors distribution for task M-Col

mismatched baselines to analyse the results. For each instance, a color word is masked with [MASK] in the caption. The representation of this token by VLM is used as input of the linear classifier $PM_{\text{M-Col}}$ in order to predict the missing color as in the MLM pre-training task. The goal is to check whether VLM representations associate text and visual features to determine the masked color. Figure 3 represents the color distribution for the M-Col task.

- **Size Identification (M-Size):** This task aims at assessing if object size is a multimodal concept included in VLM representations. We want to force the probing model to use multimodal cues instead of textual bias for this task, so we build the dataset to minimize the possibility of using only linguistic cues. Instances are selected if their caption contain size adjectives (large, big, long, tall or small, little, short, narrow). Then, we select among those captions 56 concrete object categories that are present in the test set with opposite size adjectives (i.e. large vs small) subject to a relatively balanced prior. To ensure this balance, the least frequent variant represents at least 10% of the occurrences in the subset. For example, there are no examples in the dataset describing a rock as “small” while more than 100 describe one as “large”, leading this category to be left out. By comparison, if we compare “small” and “large” dogs, 37% of dogs are “large”. This ensures that the model has limited ability to exploit the object category bias to determine its size. We then manually create negative instances by switching the adjective with its opposite. The resulting dataset is a subset of 2,552 instances of Flickr30k. A linear binary classifier $PM_{\text{M-Size}}$ is trained to determine if the caption has been modified, using the [CLS] token representation.
- **Position Identification (M-Pos):** This task aims at assessing if object position is a multimodal concept present in VLM representations. For this task, we also minimize the possibility of exploiting linguistic bias. Captions are selected based on their use of positional expressions (bottom, top, inside, outside, left, right, up, down, towards, away from, over, under, behind, in front of). Then, we select among those captions 16 different contexts where an expression and its opposite are both present in the dataset in similar proportions. For example, the top/bottom pair is unbalanced since there are 2,362 occurrences of top and 161 occurrences bottom in the dataset, while there are 177 occurrences of “at the top” and 78 occurrences of “at the bottom” which is more balanced. To ensure relative balance, we select expressions where the least frequent variant represents at least 30% of the occurrences in the subset. The negative instances are created by switching an expression with its opposite. The resulting dataset is a subset of 2,626 instances from Flickr30k. We train a linear binary classifier $PM_{\text{M-Size}}$ to determine if the caption has been modified, using the [CLS] token representation.
- **Adversarial captions (M-Adv):** This task evaluates the general multimodal information present in VLM representations. It consists in determining if a caption matches an image, except that the examples are crafted in order to be challenging. For each caption from an MS-COCO subset, we select words corresponding to visually relevant grammatical categories (nouns, verbs, adjectives, numbers). For each target word, a likely replacement is selected from the top of the distribution output by the text-only BERT model. This means that the created captions, although wrong, are believable for a language model, which minimizes the possibility for multimodal models to rely on text bias. The adversarial instances are manually screened for semantic and syntactic correctness prior to inclusion. As a result, the words replaced in the test set are mainly object related, either related to people (15%), or from the 79 other MS-COCO categories (35%) or referring to other objects (26%), as well as noun and adjectives qualifying objects (10%), verbs (6%), words expressing quantity (6%) and others (2%). Contrary to the other tasks, as BERT is used to generate the adversarial captions, the multimodal concepts that are evaluated are diverse. We train a linear binary classifier $PM_{\text{M-Adv}}$ to determine if the caption has been altered, using the [CLS] token representation.

Experimental Setup

We choose three state-of-the-art VL models that differ in transformer architecture and pre-training tasks: UNITER (single-stream with Faster RCNN visual features), LXMERT (dual-stream with Faster RCNN visual features), and ViLT (single-stream which does not use Faster RCNN visual features). Although there are alternatives, we choose those models as they are representative of different types of architectures. We list the pre-training tasks of all three models in Table 2. The training protocol of those models vary, and they are not pre-trained on the same datasets.

	UNITER	LXMERT	ViLT
Language task	Masked Language Modeling		
Vision tasks	Region Classification		n/a
	Feature Regression		
Multimodal tasks	Image-Text Matching		
	WRA	VQA	n/a

Table 2: Pre-training tasks used by UNITER, LXMERT and ViLT. Abbreviations are Word-Region Alignment (WRA) and Visual Question Answering (VQA).

Each VLM is studied as a pre-trained model and as a fine-tuned model on fine-tuning tasks $T = VQA$ and $T = NLVR$. We choose these tasks as they differ from the tasks used for pre-training and therefore require non-trivial model fine-tuning. They are also very popular when evaluating VL models and necessitate fine-grained multimodal understanding. VQA is a visual question answering task while NLVR2 consists in determining whether a sentence is correct using a pair of images as input. Our goal is to explore the effect of the tasks T on probing task performances.

In addition, we compare the performance of VLM to monomodal baselines BERT, ResNet (He et al. 2016) and ViT (Dosovitskiy et al. 2020) for a better understanding of the performance that can be reached using a single modality.

For all three models, we use the available checkpoints for VLM_{pre}. For UNITER and LXMERT, we fine-tune the models using authors' instructions, to obtain $VLM_{fine(VQA)}$ and $VLM_{fine(NLVR)}$. For ViLT, we use the available checkpoints. For the BERT and ViT baselines, we follow the same instructions as the VLM models for the representations, which are of dimension 768. For the ResNet baseline, we use the whole final layer representation, which is of dimension 2048. We use the pre-trained models from Pytorch (Paszke et al. 2019) and Hugging Face (Wolf et al. 2020) for the experiments.

The probing model PM is a linear model trained over 30 epochs for M, V and L-BShift tasks and 50 epochs for L-Tagging, with a learning rate of 0.001. We used MSE loss to train $PM_{\text{V-ObjCount}}$ and report RMSE as a metric to evaluate V-ObjCount, and the cross entropy loss for all other probing tasks, with accuracy as metric. The results of each probing task are averaged over 5 runs. We trained the models on a cuda75-capable GPU.

Results

This section is organized according to the model analysed and to the modality of the probing task.

Pre-trained models

	UNITER	LXMERT	ViLT	BERT
L-Tagging	94.66	95.13	**96.27**	95.57
L-Tagging♣	90.86	95.36	**96.16**	-
L-BShift	80.89	70.65	72.08	**86.33**
L-BShift♣	**76.25**	72.22	71.05	-

Table 3: Language probing: Accuracy of the pre-trained VL models. ♣ indicates mismatched instances.

Language (pre-trained) Table 3 shows the results of the VLM_{pre} representations for L probing tasks: part-of-speech tagging (L-Tagging) and bigram shift (L-BShift). Results for the L-Tagging task are close for all models. For L-BShift, BERT reaches the best results with an accuracy of 86.33, and UNITER has a higher performance than the others VLMs. We notice that using wrong images as input for these tasks impacts negatively UNITER.

	UNITER	LXMERT	ViLT	ResNet	VIT
V-Flower	71.82	75.56	91.34	86.83	**99.66**

Table 4: Vision probing: Accuracy of pre-trained VL models for the fine-grained classification task (V-Flower).

Vision (pre-trained) Tables 4 and 5 show the results of VLM_{pre} representations for V probing tasks: fine-grained classification (V-Flower) and object counting (V-ObjCount). We notice that the ViLT reaches significantly better results than both UNITER and LXMERT models. V-Flower is an image-only task, so we use an empty caption. On the V-Flower task, ViLT is better than the ResNet baseline.

On the V-ObjCount task, the metric symbolises the average object count error for different models. The results show that using the associated caption significantly improves the object counting results. It shows that VLM_{pre} models use linguistic cues for V-ObjCount. The performance of ViLT drops using when using mismatched captions, but it remains better than the vision-only baseline VIT. UNITER and LXMERT, however, barely reach this performance using the right caption.

Multimodality (pre-trained) Table 6 shows the results for the multimodal probing tasks.

On the color prediction (M-Col) and adversarial examples (M-Adv) tasks, VL models reach much higher results than the monomodal baselines. UNITER and ViLT have better performance than LXMERT for M-Col, with an accuracy of 86.27 and 85.97, while LXMERT reaches 71.21. UNITER also has significantly better results than the other two models for the M-Adv task. We notice that LXMERT has better results when using the wrong image, on the M-Adv♣ and M-Col♣ tasks. It means that the LXMERT performances are both lower and more dependent on linguistic cues than UNITER and ViLT, which extract more visual information.

For the M-Size and M-Pos tasks, UNITER and LXMERT yield similar results while ViLT shows the worst results on those tasks. However, all results are close to the monomodal baselines. It seems to show that VL models have a hard time extracting visual information related to size and position. On these tasks, it seems that bias in text data is linked to the performances of the models. Thus, the concepts of size and position seem to not be very well understood at a multimodal level by VLM_{pre} models.

Fine-tuned models

Language (fine-tuned) Table 7 shows the results of the $VLM_{fine(VQA)}$ and $VLM_{fine(NLVR)}$ representations for L probing tasks.

We notice that fine-tuning negatively impacts model performance on L-BShift, and especially for LXMERT. For L-Tagging, all fine-tuned models except LXMERT$_{fine(NLVR)}$ have similar performances to pre-trained models.

The performance of UNITER for using wrong images are the only ones which show an improvement, reaching the level of the their respective "normal" task. It seems to show

	UNITER	LXMERT	ViLT	BERT	ResNet	VIT
V-ObjCount	5.49	5.49	**4.90**	6.27	4.96	5.67
V-ObjCount♣	7.20	7.31	5.44	-	**4.96**	5.67

Table 5: Vision probing: Square Root of the Mean Square Error (RMSE) for pre-trained VL models on the V-ObjCount task (lower is better). ♣ indicates mismatched instances.

	UNITER	LXMERT	ViLT	BERT	ViT
M-Col	**86.27**	71.21	85.97	37.02	41.19
M-Col♣	34.80	39.33	35.69	-	**41.19**
M-Size	57.15	**58.43**	55.45	55.66	51.76
M-Size♣	**56.06**	55.05	52.10	-	51.76
M-Pos	**55.92**	54.62	48.95	56.52	52.78
M-Pos♣	54.06	**54.68**	52.37	-	52.78
M-Adv	**79.71**	72.60	73.4	53.46	-
M-Adv♣	51.92	**61.25**	56.4	-	-

Table 6: Multimodal probing: Accuracy of pre-trained VL models. ♣ indicates mismatched instances.

		UNITER	LXMERT	ViLT
VQA	L-Tagging	93.84	94.14	**94.79**
	L-Tagging♣	93.80	94.73	**94.89**
	L-BShift	**79.48**	65.40	69.43
	L-BShift♣	**76.92**	62.74	68.32
NLVR	L-Tagging	94.37	88.44	**95.60**
	L-Tagging♣	94.38	88.49	**95.50**
	L-BShift	**72.74**	57.10	67.18
	L-BShift♣	**72.34**	57.82	67.10

Table 7: Language probing: Accuracy of the fine-tuned VL models. ♣ indicates mismatched instances, gray cells show better performance than their VLM_{pre} counterpart.

that the gap in performance of UNITER$_{pre}$ for mismatched instances is due to a specificity of its the pre-training protocol.

The lower performances for the NLVR fine-tuned models could be due to the fact that the NLVR task is used to having two images as input, contrary to pre-training and probing tasks. The lower performance of fine-tuned LXMERT models could show that LXMERT forgets more easily than other models the linguistic knowledge it has learned through pre-training.

		UNITER	LXMERT	ViLT
V-Flower	VQA	82.91	78.80	**93.11**
	NLVR	82.78	74.23	**91.23**

Table 8: Vision probing: Accuracy of fine-tuned VL models for the V-Flower task. Gray cells show better performance than their VLM_{pre} counterpart.

		UNITER	LXMERT	ViLT
VQA	V-ObjCount	**4.98**	5.13	5.20
	V-ObjCount♣	6.49	6.85	**5.87**
NLVR	V-ObjCount	4.95	5.65	**4.92**
	V-ObjCount♣	6.22	6.94	**5.50**

Table 9: Vision probing: RMSE for fine-tuned VL models on the V-ObjCount task. ♣ indicates mismatched instances, gray cells show better performance than their VLM_{pre} counterpart.

Vision (fine-tuned) Tables 8 and 9 show the results of $VLM_{fine(VQA)}$ and $VLM_{fine(NLVR)}$ for the V probing tasks. For the V-Flower task, we notice an improvement of the fine-tuned UNITER models compared to the pre-trained models. ViLT performances were already high, and decreased slightly. However, LXMERT only improves with VQA fine-tuning.

On the V-ObjCount task, UNITER and LXMERT also show improvements. UNITER fine-tuned models reach the performance of the ResNet baseline with 4.98 for UNITER$_{fine(VQA)}$. However, LXMERT$_{fine(NLVR)}$ is also worse than the pre-trained model for this task. Additionally, the results using the wrong caption also improve, showing that the increase in performance relies partly on a better extraction of visual information.

Fine-tuning improves the vision performance of UNITER and, to a lesser extent, LXMERT. This seems to show that VQA and NLVR rely on visual information that is not linearly accessible within the pre-trained models. On the the other hand, it seems that fine-tuning does not improve the visual capacity of ViLT, which was already similar in term of performance to the visual baselines for the pre-trained model. It shows that the vision performances of UNITER and LXMERT pre-trained models seem to be lacking, which could point out that the visual pre-training of the those models is a limiting factor. Our hypothesis is that it is easier to extract information from the textual input than the Faster RCNN features, making UNITER and LXMERT rely more on text than image.

Multimodality (fine-tuned) Table 10 shows the results for the multimodal probing tasks. On the color (M-Col) and adversarial (M-Adv) tasks, we notice that fine-tuned UNITER and ViLT models have slightly lower performances than their pre-trained counterpart, while LXMERT shows generally an increase in performance, except LXMERT$_{fine(NLVR)}$ for the M-Adv task. Indeed, for LXMERT especially, VQA fine-tuning leads to better performances than NLVR fine-tuning. UNITER remains the overall best model for those tasks, despite the improvement of LXMERT.

For the size (M-Size) and position (M-Pos) tasks, we notice a slight increase in performance for all models. This is

		UNITER	LXMERT	ViLT
VQA	M-Col	**83.39**	82.60	81.23
	M-Col$^\clubsuit$	35.75	**37.18**	33.49
	M-Size	**64.23**	60.85	58.62
	M-Size$^\clubsuit$	55.96	**56.60**	55.24
	M-Pos	**57.55**	56.25	53.43
	M-Pos$^\clubsuit$	**55.27**	54.51	52.84
	M-Adv	**78.37**	74.90	70.07
	M-Adv$^\clubsuit$	49.42	**61.06**	52.27
NLVR	M-Col	82.52	78.00	**83.18**
	M-Col$^\clubsuit$	36.17	**37.02**	33.83
	M-Size	**63.19**	59.28	54.20
	M-Size$^\clubsuit$	**59.28**	54.57	53.17
	M-Pos	**56.99**	56.23	52.80
	M-Pos$^\clubsuit$	54.42	**55.09**	53.31
	M-Adv	**77.50**	68.46	68.12
	M-Adv$^\clubsuit$	53.17	53.46	**57.42**

Table 10: Multimodal probing: Accuracy of fine-tuned VL models. $\clubsuit$ indicates mismatched instances, gray cells show better performance than their VLM_{pre} counterpart.

more noticeable for the M-Size task, while M-Pos results remain close to the mismatched image baseline. ViLT has the worst results on those tasks. The improvement on these tasks could be due to the fact that fine-tuning datasets are more focused on the concepts of size and position than pre-training datasets. These results seem to show that the models, and UNITER in particular, manage to extract additional visual information relevant to size, while they keep using linguistic clues for position.

Discussion

Language-oriented probing seems to show that VL models have slightly worse syntactic understanding than language-only models such as BERT. This could be due to the less varied syntactic structure of the captioning datasets used for pre-training. UNITER shows overall better performances.

Vision-oriented probing seems to show that visual pre-training is a limiting factor for VL models based on Faster-RCNN features as UNITER and LXMERT show significantly worse performance than ViLT. We think that the models rely on textual information because they cannot extract accurate visual information from the representation. This limiting factor is consistent with what has been found in other studies, such as VinVL (Zhang et al. 2021), which shows that a better object detection model leads to better downstream tasks results.

Multimodal probing shows that pre-trained VL models are able to capture some multimodal information, with UNITER reaching the best performances. While ViLT has shown better results on vision probing than UNITER, this has not translated to the multimodal probing tasks. In particular, the weaker performance in the M-Adv task could be due to the the absence of object prediction task, which could limit the semantic understanding of objects for ViLT. However, concepts related to object size and position are still not well understood by those models. These are harder to grasp because they are relatively subjective and depend on the context and annotator. For those concepts, the models still almost exclusively rely on linguistic cues, resulting in a performance drop when they cannot rely on textual bias. In additional ablation studies, we use non-curated size and position datasets to see how the models perform when there are more linguistic clues. We notice that on this dataset, UNITER pre-trained representations reach an accuracy of 71.66 on the M-Size probing task, and of 65.69 when using wrong images. For the M-Pos probing task, the model reaches 73.18 using the right images and 72.68 using the mismatched images. This shows that using linguistic cues is helpful on these tasks on less controlled datasets. The performance of the position task seem to show that visual information regarding this concept is even less accessible in representations than size-related information. It could show that the current visual pre-training is not enough to understand the positional relationship between objects at a multimodal level. This is especially true for ViLT, which shows the worst performances on those tasks.

Contrary to our expectations, fine-tuning does not necessarily lead to better cross modal probing performance. The improvements in performance on probing tasks are specific and not consistent from one model to another. This seems to point out that architecture and model pre-training are particularly important to understand multimodal concepts, and that concepts that are not well understood by a pre-trained model will generally not have much improvement with fine-tuned models.

Finally, our results seem to show that for some concepts, multimodal performance is dependent on the presence of textual biases in the dataset, which makes creating controlled datasets especially important. However, the reliance of a model on linguistic clues for training does not always help improve multimodal performance. On the contrary, LXMERT models which rely the most on linguistic clues for the M-Adv task will not necessarily show the best performance for this task.

Conclusion

We evaluate Vision-Language models: UNITER, LXMERT and ViLT using probing tasks. We find that although they extract slightly less syntactic information than language-only models. Additionally, we find that Faster-RCNN features seem to be a limiting factor for visual performances. As for their multimodal capability, UNITER manages to extract better multimodal information on some concepts, such as color. However, all models have trouble understanding less objective concepts, such as position and size. We notice for those tasks an over-reliance of VL models on linguistic clues. This highlights the importance of using more controlled datasets to evaluate multimodal performance, without allowing the models to learn linguistic bias for visual information. For future work, it would be interesting to adapt VL pre-training for better multimodal performance on fine-grained multimodal concepts such as position and size. We make available the datasets for further experiments.

Acknowledgments

This work was performed using HPC resources from GENCI–IDRIS (Grant 2021-[101693]).

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*, 6077–6086.

Basaj, D.; Oleszkiewicz, W.; Sieradzki, I.; Górszczak, M.; Rychalska, B.; Trzcinski, T.; and Zielinski, B. 2021. Explaining Self-Supervised Image Representations with Visual Probing. In *International Joint Conference on Artificial Intelligence*.

Cao, J.; Gan, Z.; Cheng, Y.; Yu, L.; Chen, Y.-C.; and Liu, J. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *European Conference on Computer Vision*, 565–580. Springer.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, 104–120. Springer.

Conneau, A.; and Kiela, D. 2018. SentEval: An Evaluation Toolkit for Universal Sentence Representations. *arXiv preprint arXiv:1803.05449*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017a. 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*, 6904–6913.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017b. 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)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hendricks, L. A.; and Nematzadeh, A. 2021. Probing Image-Language Transformers for Verb Understanding. *arXiv preprint arXiv:2106.09141*.

Hewitt, J.; and Manning, C. D. 2019. A structural probe for finding syntax in word representations. 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)*, 4129–4138.

Honnibal, M.; and Montani, I. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Huang, Z.; Zeng, Z.; Huang, Y.; Liu, B.; Fu, D.; and Fu, J. 2021. Seeing Out of tHe bOx: End-to-End Pre-training for Vision-Language Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12976–12985.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*.

Lei, J.; Li, L.; Zhou, L.; Gan, Z.; Berg, T. L.; Bansal, M.; and Liu, J. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7331–7341.

Li, L.; Chen, Y.-C.; Cheng, Y.; Gan, Z.; Yu, L.; and Liu, J. 2020a. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*.

Li, L.; Gan, Z.; and Liu, J. 2020. A closer look at the robustness of vision-and-language pre-trained models. *arXiv preprint arXiv:2012.08673*.

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

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Lindström, A. D.; Bensch, S.; Björklund, J.; and Drewes, F. 2021. Probing Multimodal Embeddings for Linguistic Properties: the Visual-Semantic Case. *arXiv preprint arXiv:2102.11115*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, 722–729. IEEE.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*, 8024–8035. Curran Associates, Inc.

Shekhar, R.; Pezzelle, S.; Klimovich, Y.; Herbelot, A.; Nabi, M.; Sangineto, E.; and Bernardi, R. 2017. Foil it! find one mismatch between image and language caption. *arXiv preprint arXiv:1705.01359*.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Suhr, A.; Zhou, S.; Zhang, A.; Zhang, I.; Bai, H.; and Artzi, Y. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.

Tan, H.; and Bansal, M. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2: 67–78.

Yu, F.; Tang, J.; Yin, W.; Sun, Y.; Tian, H.; Wu, H.; and Wang, H. 2021. ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph. In *AAAI*.

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