

Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers

Stella Frank^{*, δ} Emanuele Bugliarello^{*, ϵ} Desmond Elliott ^{ϵ}

^{δ} University of Trento ^{ϵ} University of Copenhagen

stella.frank@unitn.it {emanuele, de}@di.ku.dk

Abstract

Pretrained vision-and-language BERTs aim to learn representations that combine information from both modalities. We propose a diagnostic method based on *cross-modal input ablation* to assess the extent to which these models actually integrate cross-modal information. This method involves ablating inputs from one modality, either entirely or selectively based on cross-modal grounding alignments, and evaluating the model prediction performance on the other modality. Model performance is measured by modality-specific tasks that mirror the model pretraining objectives (e.g. masked language modelling for text). Models that have learned to construct cross-modal representations using both modalities are expected to perform worse when inputs are missing from a modality. We find that recently proposed models have much greater relative difficulty predicting text when visual information is ablated, compared to predicting visual object categories when text is ablated, indicating that these models are not symmetrically cross-modal.

1 Introduction

Vision-and-language (V&L) BERT models extend the BERT architecture (Devlin et al., 2019) to produce cross-modal contextualised representations of multimodal inputs (Lu et al., 2019; Tan and Bansal, 2019; Chen et al., 2020; Su et al., 2020). These models have proven to be highly effective when fine-tuned for a range of downstream tasks. However, in spite of their versatility, little is known about how these models use cross-modal information: do their learned representations for language tasks include visual information (vision-for-language) and vice-versa (language-for-vision)? Su et al. (2020) claim that their model will use visual context for predicting masked words, thus aligning

^{*}Equal contribution.

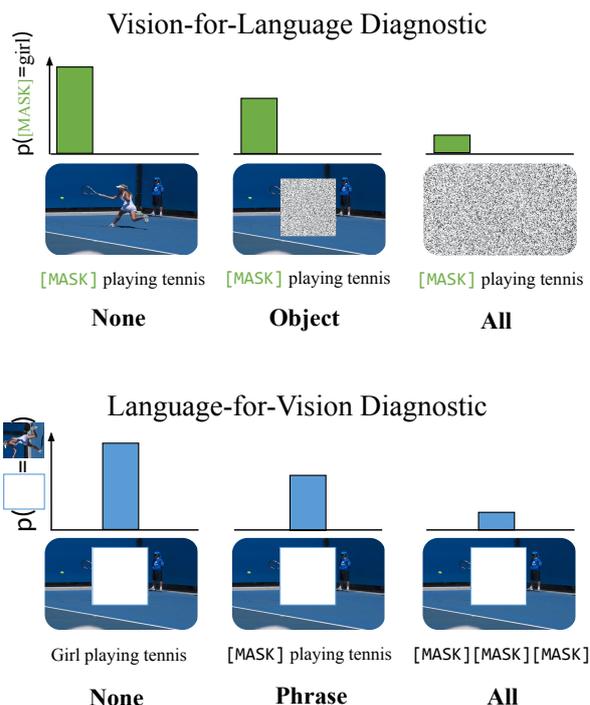


Figure 1: Cross-modal input ablation tests how well models predict masked data, given ablated inputs in the other modality. The Vision-for-Language Diagnostic (top) measures the effect of ablation, of either the aligned object or the full image, on masked token prediction, while the Language-for-Vision Diagnostic (bottom) measures the effect of ablating either the aligned Phrase or the entire sentence when predicting the properties of a masked image region.

visual and linguistic contexts, while Lu et al. (2020) hope to force their model to rely more heavily on language to predict image content. The extent to which these statements hold is unknown, in part because of the difficulty of analysing exactly how these models use information across modalities.

In this paper, we introduce a *cross-modal input ablation* method to quantify the degree to which a pretrained model has learned to use cross-modal information. This method, which requires no additional training, involves ablating all or some of the inputs from one modality when making a pre-

diction in the other modality, and measuring the change in performance. See Figure 1 for an illustrative sketch. Performance is measured using the same masked-target prediction tasks used during pretraining, in which a model must predict either a masked token given surrounding text and visual context (masked language modelling), or a masked visual object given the surrounding visual context and accompanying text (masked region classification). Cross-modal input ablation thus captures the degree to which a model depends on cross-modal inputs and activations when generating predictions.

Our use of input ablation to assess cross-modal recruitment is novel. Previous analyses of multimodal models have used diagnostic classifiers and attention analyses (Li et al., 2020; Cao et al., 2020; Parcalabescu et al., 2021). In comparison, cross-modal input ablation has the following advantages:

- It is straightforward to perform, and easy to interpret, requiring no intervention in the model and only minimal intervention on the data.
- As an intrinsic diagnostic, it examines the model directly, unlike methods that add learned parameters, such as classifier-based probing approaches (Hupkes et al., 2018).
- It does not require interpreting activations or attention, which can be difficult (Jain and Wallace, 2019).

We envision cross-modal input ablation as a useful diagnostic check to be performed during model development, to test the effect of changes in architecture and optimisation.

In this paper we perform a case study of cross-modal input ablation on existing models, to demonstrate its utility in understanding model behaviour. We test models that have different architectures but the same initialisation and training procedures (Bugliarello et al., 2021). Our cross-modal input ablation results show that these models do learn to use cross-modal information, resulting in multimodal representations, but this is not equally true across both modalities. In particular, the representations of text segments are strongly influenced by the visual input, while the representations of visual regions are much less influenced by the accompanying textual input. This indicates that the level of cross-modal information exchange is *not symmetrical*: the models have learned to use vision-for-language more than language-for-vision.

In subsequent analyses, we attempt to understand the lack of recruitment of language-for-

vision, in order to identify possible avenues for improvement. Our experiments investigate different loss functions, initialisation and pretraining strategies, and visual co-masking procedures. None of these factors changes model behaviour significantly. However, we find that the visual object annotations used in pretraining, which are automatically generated by an object detector, are much noisier than expected. We surmise that the ensuing models do not recruit text because it is not useful for predicting these noisy object features, and discuss implications for future V&L models.

2 Related Work

Significant efforts have been devoted to understanding what is learned by text-only pretrained language models (Rogers et al., 2020), which can be categorised as probing based on diagnostic classifier, analysing attention weights, or direct evaluations (Belinkov and Glass, 2019). However, much less work has looked at V&L pretraining. Two studies have examined learned attention weights: Li et al. (2020) find that there are attention heads specialising in entity phrase grounding in VisualBERT, but do not investigate text-to-vision attention. Cao et al. (2020) find that the textual modality dominates in UNITER and LXMERT, both in terms of overall attention and more specifically during visual co-reference resolution tasks. In a direct evaluation analysis, Parcalabescu et al. (2021) evaluate ViBERT and LXMERT on a counting task and find more evidence of dataset bias (predicting frequent numbers) than accurate image grounding. These results all indicate that cross-modal interaction within these models may not be as strong or as symmetric as often assumed, which is further confirmed by the methods introduced in this paper. In addition, we evaluate a larger set of models, covering both dual-stream and single-stream models.

The ablation method introduced here is related to concurrent work in language modelling (O’Connor and Andreas, 2021) and machine translation (Fernandes et al., 2021), in which the effect of removing context is measured. O’Connor and Andreas (2021) and Fernandes et al. (2021) use ablations at both training and evaluation, whereas we only study the effect of ablating inputs during evaluation.

3 Proposed Approach

We use ablations to determine whether pretrained vision-and-language models combine inputs from

both modalities when making predictions. In general, ablations are used to test the hypothesis that a certain model component (here, multimodal inputs) is responsible for certain model behaviour: ablated models are expected to perform differently because of the missing component. Importantly, this hypothesis is falsifiable (Popper, 1935) if the ablation leads to no change.

Multimodal models are hypothesised to use cross-modal activations, triggered by multimodal inputs, when making predictions. If a multimodal model relies on activations from certain input data, e.g. object category labels, to make predictions, the ablation of this input will lead to a change in performance, whereas if the model has not learned to use the input, removing it will have no effect. The predictions we evaluate are the same kinds of predictions that the models have learned to make during pretraining, i.e. predicting the identity of masked tokens or regions given (possibly ablated) multimodal text and image contexts.

Input data are in the form of an *image* paired with a *sentence* describing the image; within the sentence, *phrases* can refer to particular *objects* in the image. We expect such aligned, grounded, phrase-object pairs to elicit especially strong cross-modal activations at prediction time. By ablating the alignment link, we test models’ ability to create and use grounded alignments. We also test whether non-specific cross-modal information is used, by ablating whole modalities.

3.1 Vision-for-Language Diagnostic

The language task consists of predicting masked tokens, possibly using visual inputs. For visual input ablation, we compare the following setups:

None: None of the visual features are ablated, i.e. the model has access to the full image. This is the original multimodal setting and thus where a model that uses multimodal information effectively should perform best.

Object: Here we remove only the image regions that correspond to the aligned textual phrase. This tests a model’s ability to ground text to specific regions of the image, by breaking any possible alignment. However, the model can still use surrounding visual context features.

All: All the visual features are ablated and the model needs to predict masked textual tokens from its textual context only. Models that depend on multimodal inputs should suffer.

Figure 1 (top, middle) shows an example of **Object** ablation, where the target is to predict “*girl*” and the ablated visual input is the image region corresponding to the tennis player.

3.2 Language-for-Vision Diagnostic

The vision task is to predict the target object category within a specific region of the image, possibly aided by the text caption. In this case, a single region, corresponding to an object aligned to a phrase in the sentence, is selected as the target. The textual input is ablated analogously to the visual input:

None: None of the text is ablated, i.e. the model sees the entire sentence.

Phrase: Here we only ablate the tokens in the phrase aligned to the target object.

All: All tokens are ablated and the model is required to predict the masked visual region from its visual context only.

Figure 1 (bottom, middle) shows an example of **Phrase** ablation, which tests the extent to which the model relies on grounding information when predicting the target tokens. If ablating the phrase does not change performance compared to ablating the entire sentence, then the model has not learned to use phrase-to-object alignments to infer what is missing in the image given the sentence.

4 Experimental Setup

In this section, we describe our setup to evaluate cross-modal influence in pretrained vision-and-language BERTs. Our code is available online.¹

4.1 Evaluation data

We use the validation split of the Flickr30k Entities dataset (Plummer et al., 2015), a subset of Flickr30k (Young et al., 2014). This dataset contains human-annotated alignments between gold image regions and phrases in image captions.

The validation set contains 1,000 images, each associated to 5 English sentences. A sentence contains on average 2.89 phrases. Each phrase is aligned with 1.53 objects on average; 78% of phrases are linked to a single object. In total, this results in 14,433 phrase-object(s) data points.

4.2 Models

We evaluate five different architectures: LXMERT (Tan and Bansal, 2019) and ViL-

¹<https://github.com/e-bug/cross-modal-ablation>.

BERT (Lu et al., 2019) (dual-stream); VL-BERT (Su et al., 2020), VisualBERT (Li et al., 2019) and UNITER (Chen et al., 2020) (single-stream). These models extend the BERT architecture to multimodal data and tasks by supplementing the text input with image features. Single-stream models process image and text jointly with a single Transformer encoder, while dual-stream ones first encode each modality separately and then make them interact. We reuse the model weights from our previous *controlled* study (Bugliarello et al., 2021).²

Starting from BERT, each model was pretrained on Conceptual Captions (CC; Sharma et al. 2018) for 10 epochs. The input images were provided as 36 regions of interest extracted using a Faster R-CNN (Ren et al., 2015) pretrained on Visual Genome (Anderson et al., 2018). Each model was pretrained on three objectives: masked language modelling (MLM), masked region classification with KL divergence (MRC-KL), and image–text matching (Chen et al., 2020). We note that since each model used the same visual and textual vocabularies, as well as the same context length, cross-entropies and other entropy-based measures are comparable between model architectures.

In addition, as a control model, we adapt BERT to the same multimodal distribution by continued pretraining with MLM on Conceptual Captions for 5 epochs.³ We denote this version as BERT_{CC}.

Compute infrastructure Our experiments were run on a shared computing infrastructure. Ablation setups were measured on CPU, with an average runtime of 9 minutes. When we pretrain V&L BERTs, we use the same hyperparameters as in our controlled study (Bugliarello et al., 2021). Pretraining takes 5 days, corresponding to 10 epochs of CC, on one NVIDIA Titan RTX GPU card.

4.3 Prediction Tasks

The prediction tasks mirror two of the training tasks, MLM and MRC-KL, with a few subtle but crucial differences in masking and ablation. In the following, we denote the input textual tokens as $\mathbf{w} = \langle w_1, \dots, w_T \rangle$ and the input image regions as $\mathbf{v} = \langle \mathbf{v}_1, \dots, \mathbf{v}_K \rangle$. In addition, we use \mathbf{m} to refer to the set of masked indices, either in the textual or visual modality, and $\setminus \mathbf{m}$ for its complement.

²<https://github.com/e-bug/volta>.

³This results in the same number of MLM updates observed by vision-and-language models in their pretraining.

Masked language modelling The MLM task is to predict the identity of a set of masked tokens $\mathbf{w}_{\mathbf{m}}$, given unmasked tokens $\mathbf{w}_{\setminus \mathbf{m}}$ and visual context \mathbf{v} :

$$\text{MLM}(\mathbf{m}, \mathbf{w}, \mathbf{v}; \theta) = - \sum_{i \in \mathbf{m}} \log_2 \mathbb{P}_{\theta}(\mathbf{w}_i | \mathbf{w}_{\setminus \mathbf{m}}, \mathbf{v}), \quad (1)$$

where θ denotes a model’s parameters.

During pretraining, tokens are masked at random, using the special [MASK] token. During evaluation, only tokens corresponding to a specific grounded phrase are masked for prediction.

When ablating visual inputs (as opposed to masking), the ablated regions are replaced with the average feature vector (calculated over the evaluation dataset). This keeps the visual inputs indistribution, but renders them uninformative. During **Object** ablation, all regions that overlap with the gold region are ablated, using the intersection over target (IoT) function defined in Eq. (4) below, with an overlap threshold of $\tau = 0.5$.

Masked region classification The MRC task is to predict the object class of a masked visual region \mathbf{v}_i given unmasked visual context $\mathbf{v}_{\setminus \mathbf{m}}$ and tokens \mathbf{w} . The MRC-KL variant (Li et al., 2019) measures the KL-divergence of the predicted distribution rather than the cross-entropy against a single object class. For each masked region \mathbf{v}_i linked to a phrase, MRC-KL is computed as follows:

$$\text{MRC-KL}(\mathbf{w}, \mathbf{v}_i; \theta) = \text{KL}(\mathbb{P}_g(\mathbf{v}_i) || \mathbb{P}_{\theta}(\mathbf{v}_i | \mathbf{w}, \mathbf{v}_{\setminus \mathbf{m}})), \quad (2)$$

where \mathbb{P}_g is the target object distribution and \mathbb{P}_{θ} is the distribution predicted by the model. During training, the visual annotations of object classes \mathbb{P}_g are given by the Faster R-CNN object detector, which has a set of 1,601 object classes. Chen et al. (2020) showed that MRC-KL leads to better downstream performance compared to MRC with cross-entropy (MRC-XE).

When a given region is masked for prediction, other regions that overlap with it are also masked out, according to the IOT function (Eq. (4)). In practice, masking for prediction involves replacing the input region vector with a vector of zeros. When performing textual ablation, the ablated tokens are replaced with [MASK] tokens.⁴

During our evaluation procedure, we only evaluate region predictions for the target objects that are

⁴For the **All** ablation setup, we also investigated using one [MASK] token but found no difference in model behaviour.

mentioned in the text. As the target region, among the regions masked by the IoT function, we use the (Faster R-CNN-generated) input region that has the highest overlap with the gold one, according to intersection over union (IoU; Eq. (3)).

4.4 Region Overlap Calculations

Object detectors often output regions that overlap with each other. These regions may leak visual information and hence need to be masked in our diagnostic procedure. In particular, there are two points where this issue arises: first, in language-for-vision, when deciding which regions overlap with the target region of prediction; secondly, in vision-for-language **Object** ablation, when deciding which regions overlap with the gold region.

Intersection over union Visual region overlap is most commonly calculated using intersection over union (IoU; Everingham et al. 2010). Let \mathbf{b}_i and \mathbf{b}_j be the bounding boxes corresponding to regions \mathbf{v}_i and \mathbf{v}_j , respectively. IoU is computed as:

$$\text{IoU}(\mathbf{b}_i, \mathbf{b}_j) = \frac{|\mathbf{b}_i \cap \mathbf{b}_j|}{|\mathbf{b}_i \cup \mathbf{b}_j|}. \quad (3)$$

Our models were pretrained by co-masking any region that overlapped with the randomly chosen one by more than the threshold of $\tau = 0.4$.

Intersection over target A shortcoming of IoU is that it may not catch large regions that largely overlap small target regions, potentially leaking information. To avoid this, we define the more aggressive intersection over target (IoT) measure:

$$\text{IoT}(\mathbf{b}_i, \mathbf{b}_j) = \frac{|\mathbf{b}_i \cap \mathbf{b}_j|}{|\mathbf{b}_i|}. \quad (4)$$

To find co-ablated regions during **Object** ablation, we use IoT with a threshold $\tau = 0.5$ (as originally recommended by Everingham et al. 2010 for IoU).

5 Cross-Modal Input Ablation Results

Vision-for-Language diagnostic Figure 2 (left) shows the performance of the five models in different types of visual input ablation. The models perform best when **None** of the visual input is ablated, and perform worse when the **Object** is ablated, indicating that the models are missing information about the grounding object region. Models are most affected when removing **All** of the visual input, which is unsurprising. However, the effect of **Object** ablation is relatively small, compared to the

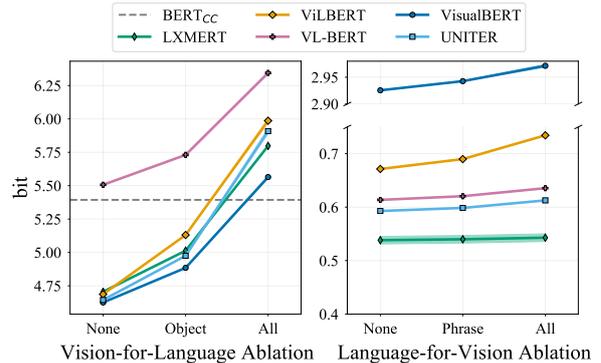


Figure 2: Left: Masked language modelling loss for different ablation settings: no ablation, ablation of the grounding object, ablation of the full image. Right: Masked region classification with KL divergence loss for different ablation settings: no ablation, ablation of the region-aligned phrase, ablation of the full text. The mean performance of 10 different initial seeds is shown for each model; variance around the mean is too small to be seen for most models. Models suffer when visual input, but not textual input, is removed.

effect of **All** ablation, which is counter to the expectation that models should use grounding object information rather than general visual context.

In Figure 3, we check whether these results are due to visual leakage of object information into the general visual context. We examine the effect of masking the visual object more or less aggressively, by changing the overlap threshold τ that triggers co-masking. Increased masking leads to worse performance on **Object** ablation, and at lower thresholds for masking, the relative contribution of **Object** information is higher than that of the additional general visual context in **All** ablation.

Comparing models, the single-stream VisualBERT is least affected by ablating the visual input, while the dual-stream ViLBERT is most affected. The behaviour of VL-BERT is an unexpected outlier that warrants future investigation, given that it performs similarly to other models in downstream tasks (Su et al., 2020; Bugliarello et al., 2021).

The language-only BERT model has a loss of 7.44 bits on this evaluation set, which drops to 5.39 bits after pretraining on the Conceptual Captions dataset (BERT_{CC}): there is a strong effect of the language domain. Overall, when visual input is provided, the V&L BERTs show better performance than the language-only baselines, but their performance degrades below domain-adapted unimodal performance when visual input is with-

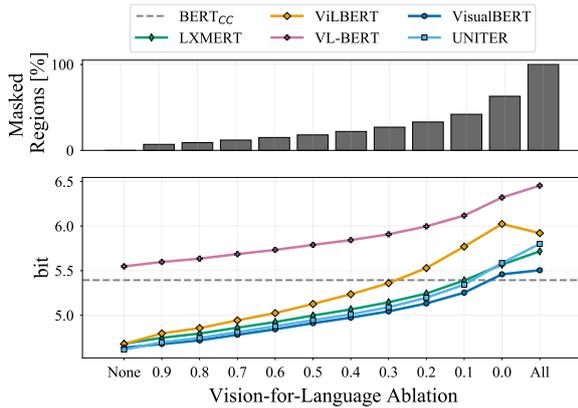


Figure 3: MLM performance across different visual masking thresholds, leading to increased co-masking of regions overlapping with the ablated **Object**. $\tau = 0.5$ corresponds to **Object** ablation in Figure 2.

held. We conclude that these models have indeed learned to use vision-for-language.

Language-for-Vision diagnostic On the right of Figure 2 are the results of ablating language when predicting visual object categories. Compared to the effect of ablating visual inputs for language modelling, where the performance drops by up to 25% during **All** ablation, we see only minimal effects of ablating text. According to paired-difference t-tests, all models’ ablated performances (both **Phrase** and **All** ablation) differ significantly from the standard model (**None** ablation), with $p < 0.05$. However, the practical effects of ablation are very small. Ablating the reference phrase leads to a increase in KL-divergence of 0.5%–3%, while ablating **All** text increases the KL-divergence between 1%–10%, relative to **None** ablation. In both cases, the least affected model is LXMERT, and the most affected model is ViLBERT. In summary, this analysis shows that these V&L models use language-for-vision much less than they use vision-for-language.

6 Diving into Language-for-Vision

The previous section showed that V&L BERTs are more sensitive to ablated visual inputs than language inputs. This behaviour could be due to several factors, including differences in model design and initialisation, pretraining objectives, and issues with the quality of the silver labels provided by Faster R-CNN. Here, we analyse how these factors affect language-for-vision recruitment.

6.1 Model Design

In theory, particular model architectures could favour the transfer and cross-modal use of features in one direction more than the other. Overall, our results show all of the evaluated model designs resulting in quite similar, asymmetric, behaviour.

Interestingly, however, the two dual-stream architectures, LXMERT and ViLBERT, behave very differently. LXMERT achieves the lowest MRC-KL value, and it is not affected by the presence of textual inputs. ViLBERT, on the other hand, is the model that most relies on language when predicting visual features. We believe that the main reason behind the difference in performance might be due to their different vision streams: ViLBERT directly maps visual features onto cross-modal layers, while LXMERT first processes them in a 5-layer vision-only stream. We hypothesise that these five layers are enough for LXMERT to learn visual representations to solve the MRC-KL task, without language. Secondly, the lack of a vision-only stream in ViLBERT makes it asymmetric, which might explain its greater recruitment of language for vision. Conversely, all the other models have approximately the same number of parameters for both modalities.

Finally, VisualBERT performs much worse than all the other models on MRC-KL. We hypothesise this is because it is the only model that does not encode spatial information from the bounding boxes corresponding to the input visual features.

6.2 Pretraining Loss Function

We investigate whether pretraining with a different vision objective leads to different behaviour in language-for-vision recruitment. In particular, predicting object class distributions, using MRC-KL, may be easier if the long-tail of object classes is relatively similar between (non-masked) regions within an image, since this would make visual contextual features highly informative. Replacing the KL divergence objective with a cross-entropy objective forces the model to predict only the most likely object class for each region, which is expected to be more dissimilar across the image.

We pretrain a UNITER encoder using a weighted masked region classification with cross-entropy (MRC-XE) objective (Tan and Bansal, 2019), where the weights are given by the probability assigned by the object detector to the most likely class. We find that pretraining with this vision objective does *not* lead to any change in model be-

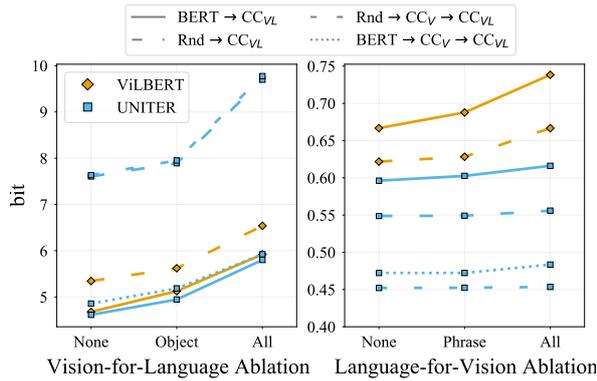


Figure 4: MLM and MRC-KL for different pretraining initialisations. Pretraining affects overall performance, but not ablation behaviour (lines are parallel).

haviour: Ablating the reference phrase leads to an increase in XE of 1%, compared to the standard model (**None** ablation); ablating **All** text increases it by 4%, in line with the KL results in Figure 2.

6.3 Initialisation and Pretraining Order

All tested models share the same pretraining sequence: they are first initialised with BERT weights and then pretrained on CC to model vision and language. We refer to this configuration as **BERT**→**CC**_{V_L}. Here, we ask whether the BERT initialisation leads to the asymmetric lack of language-for-vision behaviour: does the strong language modelling capability embedded in the Bert weights constrain the model from adapting towards the vision prediction task? In addition, we explore whether different pretraining regimes can also lead to different behaviours in V&L BERTs. We consider the following setups:

Rnd→**CC**_{V_L} Instead of initialising with BERT, the models are randomly initialised and trained on CC using all three standard losses.

Rnd→**CC**_V→**CC**_{V_L} Here, randomly initialised models are first pretrained on CC using only the visual MRC-KL loss, and then pretrained on CC as usual with all three losses. This is an attempt at ‘vision-first’ models.

BERT→**CC**_V→**CC**_{V_L} Same as above, but starting from a BERT initialisation.

We pretrain both ViLBERT and UNITER architectures in the **Rnd**→**CC**_{V_L} setting, and UNITER for the others. The results for all these setups are displayed in Figure 4, and show that (i) pretraining from scratch increases MLM loss, compared to models initialised with BERT, in both ViLBERT

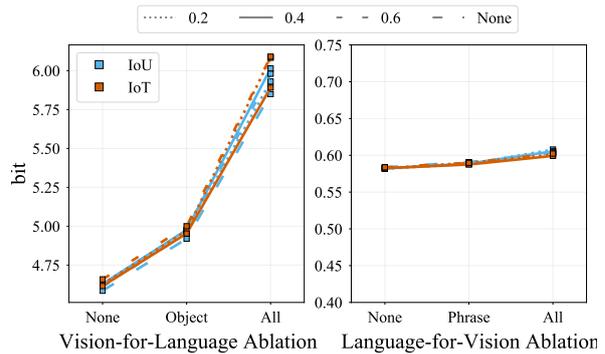


Figure 5: MLM and MRC-KL for different pretraining thresholds for the UNITER model. Neither co-masking threshold nor co-masking function (IoU or IoT) affect model performance nor behaviour in our diagnostics.

and UNITER; and (ii) pretraining with vision-first leads to lower MRC-KL loss for UNITER. Nevertheless, our empirical findings indicate that initialisation is *not* a strong factor for the observed cross-modal asymmetric behaviour.

6.4 Leaking Visual Features

The visual contextual features may (accidentally) provide sufficient information for correct predictions. In particular, given that many automatically proposed regions in an image overlap, if these are not correctly co-masked with the object region, they could leak information about the object.

We perform an experiment where we pretrain UNITER with different overlapping thresholds $\tau \in \{0.2, 0.4, 0.6\}$ as well as without any co-masking (**None**), testing both IoU and IoT masking functions. Figure 5 shows that varying the amount of co-masking, or the masking function, does *not* affect the recruitment of language-for-vision. Our result counters the hypothesis of Lu et al. (2020) that co-masking “*forces the model to rely more heavily on language to predict image content.*”

6.5 Silver Object Annotations

Finally, we examine the data used for representing the visual modality. The models are trained and evaluated on object predictions, ‘silver distributions,’ from Faster R-CNN, which contain noise because they are automatically predicted categories. During evaluation, noise in the target distributions may lead to discounting of the role of language if the evaluation set contains many examples in which the target object class predictions conflict with the aligned textual features.

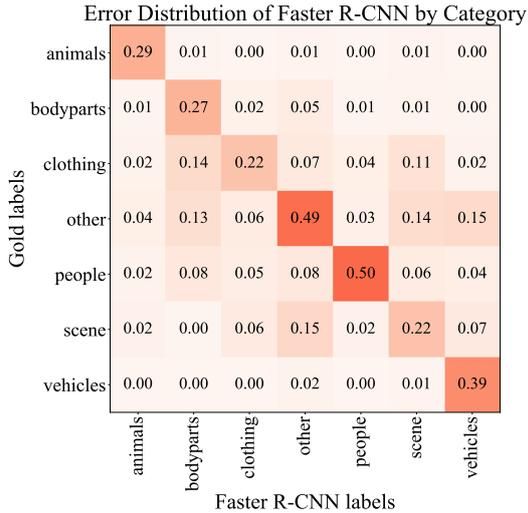


Figure 6: Confusion matrix showing the proportion of incorrect class predictions (out of all predictions), grouped into categories. 60% of Faster R-CNN “people” category label predictions are incorrect (sum of column); 50% of all predictions are incorrect within the set of “people” classes.

To examine this, we construct a subset of the Flickr30k Entities validation set, called LabelMatch, in which the nouns of an aligned phrase – extracted using Stanza (Qi et al., 2020) – perfectly match with one of the Faster R-CNN’s object labels. (Note that this does not mean that Faster R-CNN predicted that category for that example, only that it is possible for the model to have done so.) This results in 10,159 data points out of the original 14k. We consider these labels, extracted from Flickr30k Entities, as human-generated “gold” object labels.

The silver distributions are noisy On the LabelMatch subset, the highest-probability class from Faster R-CNN agrees with the gold label only on 38% of examples. The gold label is in the top-3 55% of the time, in top-5 64%, and in top-10 75% of the time. Figure 6 shows the distribution of errors, grouped into the higher-level categories defined in the Flickr30k Entities dataset. Faster R-CNN is mostly making mistakes within categories, especially within the “people” category, where more than half of all predictions are wrong according to the gold label. There is also a long tail of wrong predictions not in the Flickr30k Entities object classes within the “other” predictions.

Evaluating on gold labels Using the LabelMatch as gold labels for evaluation does not lead to any clear difference in ablated MRC performance, as seen in Figure 7 (right). (Figure 7 (left) confirms

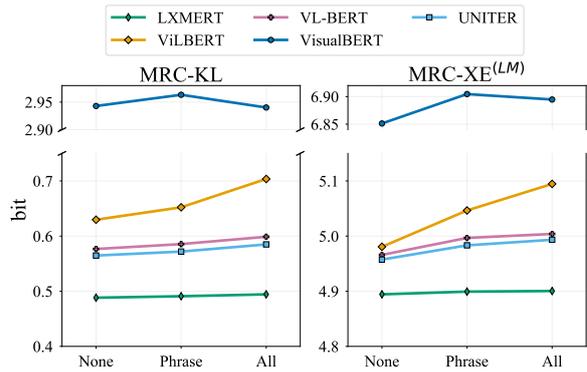


Figure 7: LabelMatch subset results: MRC-KL against silver distributions and MRC-XE against gold labels.

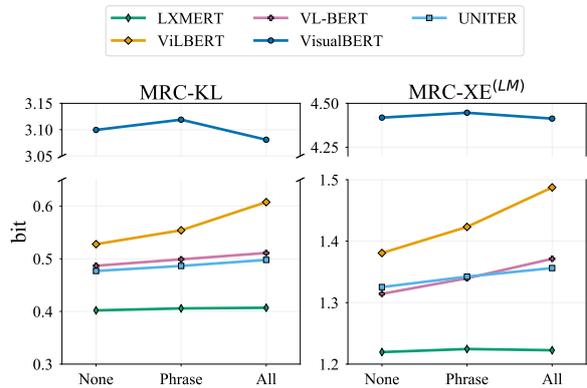


Figure 8: MRC-KL and MRC-XE in the subset of LabelMatch where Faster R-CNN’s prediction match with gold labels. We observe similar patterns as in LabelMatch showing that models do not recruit text for vision even when the predicted object classes are correct.

that this subset does not behave differently from the full validation set on MRC-KL against the silver distribution labels.) Models still do *not* rely on text when predicting the (gold) region class, even though — in the un-ablated setting — the region label is given in the text input in this dataset.

Finally, we also measure MRC-KL on the subset of LabelMatch in which the silver distribution mode and the gold labels agree. This tests whether models use language for vision when the most likely class of silver distribution is correct: perhaps these instances are recognisably cleaner to the model. However, we again find no difference in ablated performance (see Figure 8).

In conclusion, even when evaluated on gold labels, we still see most models making next to *no* use of textual information for visual predictions. This behaviour is consistent with models that have been pretrained against noisy silver data, where language inputs are not useful for prediction.⁵

⁵We believe it also explains why Chen et al. (2020) found

7 Conclusion

The *cross-modal input ablation diagnostic* introduced in this paper demonstrated an asymmetry in pretrained vision and language models: the prediction of masked text is strongly affected by ablated visual inputs, compared to (almost) no effect of ablating textual inputs when predicting masked image regions. These results offer a useful check of actual model behaviour, and run counter to hypotheses assuming more balanced cross-modal activations.

We conducted several follow-up studies to better understand, and possibly ameliorate, this behaviour. We explored (i) pretraining the model on vision before cross-modal training; (ii) visual leakage from overlapping image regions; and (iii) problems with the silver target distributions produced by Faster R-CNN. There were no discernible effects from (i) or (ii). With regards to (iii), we found the silver labels to be much less reliable than expected. We suspect that these silver target distributions are not appropriate for activating language-for-vision, and we recommend that future V&L models avoid them, if they aim to not only support vision-for-language.

This paper has focused on an intrinsic evaluation of pretrained models, rather than their downstream task performance. We note that the models studied in this paper perform similarly on a number of downstream tasks (Bugliarello et al., 2021), in line with their similar ablation behaviour. Whether cross-modal ablation behaviour can predict downstream performance is a question left for future models with more divergent behaviour.

More broadly, the fact that the tested models showed an effect of vision-for-language but not language-for-vision may be an accumulation of model engineering for multimodal tasks that are more strongly vision-for-language, such as visual question answering or grounded reasoning. Some recent models (Li et al., 2021; Shen et al., 2021; Hendricks and Nematzadeh, 2021) have removed the visual component from the training objective entirely, resulting in vision-for-language architectures. In the future, we advocate for increased work on more language-for-vision tasks, such as text-modulated object detection (Kamath et al., 2021), in order to push pretrained vision-*and*-language models to be truly, bidirectionally, cross-modal.

an advantage of pretraining with a KL-divergence loss instead of a maximum-probability cross-entropy loss: predictions made with KL divergence are more robust to noise.

Acknowledgements

■ We are grateful to the anonymous reviewers and members of the CoAStAL NLP group for their constructive feedback. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801199.

References

- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. [Bottom-up and top-down attention for image captioning and visual question answering](#). In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086. IEEE Computer Society.
- Yonatan Belinkov and James Glass. 2019. [Analysis methods in neural language processing: A survey](#). *Transactions of the Association for Computational Linguistics*, 7:49–72.
- Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. [Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs](#). *Transactions of the Association for Computational Linguistics*.
- 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 *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI*, volume 12351 of *Lecture Notes in Computer Science*, pages 565–580. Springer.
- 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 *European Conference on Computer Vision*, pages 104–120.
- 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.
- Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. 2010. [The Pascal visual object classes \(voc\) challenge](#). *Int. J. Comput. Vision*, 88(2):303–338.

- Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. [Measuring and increasing context usage in context-aware machine translation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, Online. Association for Computational Linguistics.
- 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.
- Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. [Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure](#). *Journal of Artificial Intelligence Research*, 61:907–926.
- Sarthak Jain and Byron C. Wallace. 2019. [Attention is not Explanation](#). 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 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.
- Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. 2021. [MDETR – modulated detection for end-to-end multi-modal understanding](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. [Align before fuse: Vision and language representation learning with momentum distillation](#). *ArXiv preprint*, abs/2107.07651.
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. [Visualbert: A simple and performant baseline for vision and language](#).
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020. [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.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. [Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.
- Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. [12-in-1: Multi-task vision and language representation learning](#). In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10434–10443. IEEE.
- Joe O’Connor and Jacob Andreas. 2021. [What context features can transformer language models use?](#) In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 851–864, Online. Association for Computational Linguistics.
- Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. 2021. [Seeing past words: Testing the cross-modal capabilities of pretrained v&l models on counting tasks](#). In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 32–44, Groningen, Netherlands (Online). Association for Computational Linguistics.
- Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. [Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models](#). In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society.
- Karl Popper. 1935. *Logik der Forschung*. Springer Verlag.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.
- Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. [Faster R-CNN: towards real-time object detection with region proposal networks](#). In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.
- Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. [A primer in BERTology: What we know about how BERT works](#). *Transactions of the Association for Computational Linguistics*, 8:842–866.
- 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 (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

- Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2021. [How much can clip benefit vision-and-language tasks?](#) *ArXiv preprint*.
- 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.
- 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.
- 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.