Pixology: Probing the Linguistic and Visual Knowledge of Pixel-based Language Models

Anonymous EMNLP submission

Abstract

Pixel-based language models (LMs) have emerged as a compelling alternative to subword-based LMs, particularly because they can represent virtually any script. PIXEL, a canonical example of such a model, is a vision transformer that has been pre-trained on rendered text. While PIXEL has shown promising cross-script transfer abilities and robustness to orthographic perturbations, it falls short of outperforming monolingual subword counterparts like BERT in most other contexts. This discrepancy raises questions about the amount of linguistic knowledge learnt by these models and whether their performance in language tasks stems more from their visual capabilities than their linguistic ones. To explore this, we probe PIXEL using a variety of linguistic and visual tasks to assess its position on the vision-to-language spectrum. Our findings reveal a substantial gap between the model's visual and linguistic understanding. The lower layers of PIXEL predominantly capture superficial visual features, whereas the higher layers gradually learn more syntactic and semantic abstractions. Additionally, we examine variants of PIXEL trained with different text rendering strategies, discovering that introducing certain orthographic constraints at the input level can facilitate earlier learning of surface-level features. With this study, we hope to provide insights that aid the further development of pixel-based language models.

1 Introduction

Subwords are currently the standard units of processing in language modelling (Sennrich et al., 2016). While they have been shown to work well in monolingual models (Devlin et al., 2019; Liu et al., 2019), in a multilingual context they can lead to an inevitable vocabulary bottleneck with each language competing for space in a finite vocabulary (Rust et al., 2023; Liang et al., 2023). Characters and byte-based models have been proposed as alternatives to subwords, but they lead to longer input sequences (Raffel et al., 2020; Xue et al., 2022; Tay et al., 2022; Clark et al., 2022). Another proposed solution is pixel-based models where patches of pixels are the main unit of representation. A canonical example of this is the PIXEL (Pixel-based Encoder of Language) model (Rust et al., 2023), where text is rendered as a sequence of fixed-sized patches and passed as input to a vision transformer (ViT) (Dosovitskiy et al., 2021). This approach allows the model to represent virtually any script.

Although current versions of the pixel-based language models do not outperform their monolingual subword-based counterparts on most downstream tasks (Rust et al., 2023; Lotz et al., 2023), they are a promising approach to multilingual modelling and offer a unique opportunity to explore modelling language through images. PIXEL is a juxtaposition of a vision and language model: even though it receives image patches as input, the content of those patches is rendered text, making it a visual model of language. With this study, we aim to understand where PIXEL stands on the vision-to-language spectrum. To this end, we probe PIXEL on various visual and language tasks and compare performance with BERT (Devlin et al., 2019) – the language model it is most comparable to – and VIT-MAE (He et al., 2022) – the vision model it is most comparable to. We conduct a comprehensive analysis of the linguistic and visual capabilities of PIXEL that can be used to aid further development of pixel-based language models. Concretely:

RQ1: How much linguistic knowledge is encoded in PIXEL?

RQ2: How much visual knowledge does PIXEL have?

We find that PIXEL learns surface-level linguistic information in the lower layers, resulting in higher-

level syntactic and semantic abstractions appearing in higher layers than BERT (§5.1). When comparing to ViT-MAE, PIXEL underperforms on image tasks, with visual probing accuracy decreasing in the higher layers (§5.2). Thus, the surface-level information is diluted as it acquires linguistic knowledge in the higher layers.

Lotz et al. (2023) trained newer pixel-based language models that add some orthographic constraints to the input that can potentially augment linguistic learning in the lower layers. In this context, we ask the following question:

RQ3: Does adding orthographic constraints to the input enhance the linguistic knowledge in PIXEL?

We find that a rendering strategy that makes word boundaries more explicit in the input enables PIXEL to learn surface-level linguistic features earlier in the model, thereby aiding semantic understanding (§5.3).

Overall, we take inspiration from BERTology, the study of the linguistic capabilities in BERT (Rogers et al., 2020), and aim for this work to foster future explorations and advancements for PIXEL.

2 Background

2.1 PIXEL

The pixel-based language models examined in this study are ViTs (Dosovitskiy et al., 2021) that lie at the confluence of NLP and computer vision. A ViT is an application of the transformer architecture (Vaswani et al., 2017; Devlin et al., 2019) to process images. An image is split into patches that are each flattened into a vector and then projected into a lower-dimensional space through a linear transformation. Positional embeddings are added to retain spatial information before feeding these patch vectors into a transformer encoder.

Inspired by the self-supervised masked language modelling paradigm, a variant of ViT is the masked auto-encoder (He et al., 2022), or ViT-MAE, that learns image representations by masking random image patches. A decoder reconstructs the image from the latent representation of the mask tokens. The PIXEL model by Rust et al. (2023) is trained on the ViT-MAE architecture. It takes a rendered image of text sized 16×8464 as input, which is split into patches of 16×16 pixels. Instead of randomly masking *individual* patches, PIXEL randomly masks *spans* of patches to force the model to learn higher levels of language abstraction. PIXEL is pre-trained on a rendered version of the English Wikipedia and the BookCorpus (Zhu et al., 2015). Thus, it is comparable to BERT in terms of pre-training data and ViT-MAE in terms of architecture and parameters.

PIXEL follows the idea of visual text representations by Salesky et al. (2021), who embed rendered text using 2D convolutions for continuous open-vocabulary machine translation. They demonstrate that visual text representations are more robust to noise and provide the benefits of a tokenization-free text processing pipeline.

Lotz et al. (2023) further improved PIXEL by experimenting with different text rendering strategies. Their work provides insights into the semantic modelling capabilities of PIXEL models and correlates that to frequency bias. We include some of these models in our study.

2.2 Model Interpretability

The survey by Zhao et al. (2024) categorises model interpretability into local explanations of predictions and global explanations of model behaviour. Global explanations aim to understand the general concepts encoded in the individual components of a language model. The most prominent method for global explanations of linguistic understanding in language models is *probing*, specifically classifier-based probing (Belinkov, 2022).

In this approach, model weights are frozen and for each of its layers, a small classifier is trained to solve a task given a pooled representation of the intermediate embeddings at that layer. The task is designed to isolate an aspect of linguistic understanding that may or may not be present in the embedding (Adi et al., 2016; Hewitt and Manning, 2019; Şahin et al., 2020; Zhu et al., 2022). The same idea has been used for investigating computer vision models (Alain and Bengio, 2018; Basaj et al., 2021) and, more recently, multi-modal models (Dahlgren Lindström et al., 2020).

A standard framework for linguistic probing is SentEval (Conneau and Kiela, 2018), which includes various probing tasks that uncover different levels of linguistic information in sentence embeddings. SentEval has been extensively employed to analyse models for sentence-level semantics (Ma et al., 2019; Krasnowska-Kieraś and Wróblewska, 2019; Ravichander et al., 2021), and it is the dataset we adopt in this study.

Linguistic probing has been used prominently in BERTology (Rogers et al., 2020) to understand

Type	Name	Predict for a given sentence...	Labels
Linguistic Probing			
Surface	Sentence Length (**SentLen**)	the length.	6 bins
	Word Content (**WC**)	which one of 1000 possible words is in it.	1000
Syntactic	Bigram Shift (**BShift**)	if the order of two random words was inverted.	2
	Tree Depth (**TreeDepth**)	the depth of the syntactic tree.	5-12
	Top Constituents (**TopConst**)	the sequence of top constituents directly below the sentence (S) node.	20
Surface Semantic	Tense (**Tense**)	the tense of the main verb.	3
	Subject Number (**SubjNum**)	the number of the subject.	2
	Object Number (**ObjNum**)	the number of the object.	2
Complex Semantic	Semantic Odd Man Out (**SOMO**)	if a noun or a verb has been switched out for another.	2
	Coordination Inversion (**CoordInv**)	if the two coordinate clauses have been inverted.	2
Visual Probing			
Visual	Max Count (**MaxCount**)	the frequency of the character with the max count.	4 bins
	Argmax Count (**ArgmaxCount**)	the character that has the max count.	5 bins

Table 1: Description of probing tasks used in this study.

the levels of linguistic information stored in BERT embeddings (Tenney et al., 2019b; Jawahar et al., 2019; Mehrafarin et al., 2022). It has been established that BERT tends to encapsulate more syntactic knowledge in its middle layers, while semantic comprehension is more pronounced in the higher layers (Tenney et al., 2019a). In this context, we aim to gain analogous insights about pixel-based language models.

3 Probing Tasks

We now introduce the probing tasks used in our experiments. We probe PIXEL on two levels: linguistic and visual. For linguistic probing we rely on the SentEval framework mentioned above. ViTs have more direct access to surface-level information than subword-based models, since their input is segmented into units of fixed visual size (as opposed to variable-sized tokens) and shown to the model after a continuous linear projection (as opposed to a lookup). Thus, we also employ tasks that are designed to verify whether orthographic information is more easily identifiable throughout PIXEL.

3.1 Linguistic Probing Tasks

The SentEval framework contains probes that quantify three levels of linguistic knowledge present in sentence embeddings: *surface*, *syntactic*, and *semantic* (Conneau et al., 2018). Table 1 presents all the tasks, with their type and description. We evaluate the performance of the models at each layer on these tasks to explain the hierarchy of linguistic understanding contained within the model.

We note, however, that all tasks falling under the *semantic* category do not all probe for the same kind of information. `Tense`, `SubjNum` and `ObjNum` can be solved by trivial surface cues like the presence of certain morphemes like the suffixes *-ed* and *-es*. However, unlike surface tasks, performance on these tasks does not drastically degrade in the upper layers as the model gains semantic understanding (Jawahar et al., 2019), and they can be predictors of downstream semantic performance (Zhu et al., 2022). Thus, we dub these tasks *surface semantic*.

`SOMO` and `CoordInv`, on the other hand, need more complex semantic learning to be solved. We therefore term these tasks as *complex semantic*. The distinction between *surface semantic* and *complex semantic* can also be justified by the differences in accuracies between human evaluation and model performance for tasks in both these categories as reported by Conneau et al. (2018). Most neural models are able to either match or surpass human evaluation for the *surface semantic* tasks, but not for the *complex semantic* tasks. This recategorization also helps to identify consistencies in linguistic understanding, particularly when explaining trends with BERT.

3.2 Visual Probing Tasks

We introduce two new tasks to probe for purely visual information – `MaxCount` and `ArgmaxCount` (see Table 1). Every word in every sentence of the `SentLen` task is replaced by a random English

word generated with the `wonderwords`[1] library to create synthetic datasets. By using random words instead of a sentence, we ensure that the task is purely visual, but does not disadvantage the BERT tokenizer (as opposed to using random characters which could result in single-character tokens). This also distinguishes them from the *surface* tasks in SentEval since there is no underlying linguistic pattern to this data. The labels are binned to ensure a uniform distribution and we down-sample the labels that occur with a very high frequency (for example, 'e' is the most frequent letter in 50% of the dataset). More task details are in Appendix A.

MNIST As a final task to probe for purely visual information, we rely on MNIST (Deng, 2012), which consists of white-on-black images of handwritten digits (0 to 9). It is an image classification benchmark dataset and its resemblance to rendered text as well as the simplicity of the task make it suitable for probing.[2] We do not evaluate BERT on this task since it cannot represent images.

4 Experimental Setup

4.1 Models

Our analysis will primarily focus on the PIXEL-base model trained by Rust et al. (2023), further termed PIXEL. We also make a comparison with its variants introduced by Lotz et al. (2023) for **RQ3**. Specifically, we look at PIXEL-bigrams, pre-trained using the `bigrams` rendering strategy which ensures that every patch contains at most 2 characters, and that no patch overlaps a word boundary, adding extra space where needed. We also look at PIXEL-small-words, trained on the `words` rendering strategy that merely enforces the second constraint. Since it has no base version released, we additionally probe PIXEL-small-bigrams and PIXEL-small for a fair comparison.[3] All these are compared against BERT and ViT-MAE. An overview of the model parameters is in Appendix B.

4.2 Probing

We follow the same probing setup as defined by Conneau and Kiela (2018). Sentence representations for each example in the datasets are obtained by mean-pooling the token or patch embeddings generated at every hidden layer for each model. These embeddings are passed to a classifier that learns to predict the corresponding class label using a cross-entropy loss. For our experiments, we use the implementation and default hyper-parameters proposed by Araujo et al. (2022) for both linguistic and visual tasks.

4.3 Fine-tuning

For a better understanding of the general linguistic abilities of vision models (**RQ1**), we fine-tune ViT-MAE on universal dependencies (UD) (Nivre et al., 2016) POS-tagging, dependency parsing and GLUE (Wang et al., 2018) using the same hyperparameters as Rust et al. (2023). We re-use PIXEL's text rendering configuration, and render text into a square image of 224×224 to match the input size of ViT-MAE. To gauge the general visual abilities of PIXEL (**RQ2**), we fine-tune PIXEL and ViT-MAE on the CIFAR100 (Krizhevsky and Hinton, 2009) image classification dataset.

5 Results and Analysis

5.1 RQ1: How much linguistic knowledge is encoded in PIXEL?

To investigate this question, we first compare PIXEL and ViT-MAE fine-tuned on language tasks. This is to assess the extent to which PIXEL's pre-training regime makes it better at language tasks than a regular vision transformer. Results are in Table 2.

Task	PIXEL	ViT-MAE	BERT
PoS	96.7	93.1	97.2
DP	88.7	78.2	90.6
GLUE avg.	74.1	58.1	80.0

Table 2: Language fine-tuning results for PIXEL, BERT (taken from Rust et al. (2023)) and ViT-MAE

It is clear that PIXEL has an advantage over ViT-MAE. Since PIXEL performs substantially better than ViT-MAE on GLUE, it can be argued that PIXEL learns some semantics. This can be explained either by the domain similarity between PIXEL pre-training and the downstream task input or because its pre-training on language actually enables the model to learn linguistic abstractions.

To investigate which of the two factors explains the advantage, we run the linguistic probing tasks on PIXEL, BERT and ViT-MAE, illustrated in Figure 1. Each plot also includes the majority baseline[4] for that task as a lower bound for each model.

[1] `github.com/mrmaxguns/wonderwordsmodule`

[2] Each image is 28×28 pixels that we resize to 16×16, the image patch size for one patch in PIXEL.

[3] `huggingface.co/Team-PIXEL`

[4] The accuracy if always predicting the most frequent label.

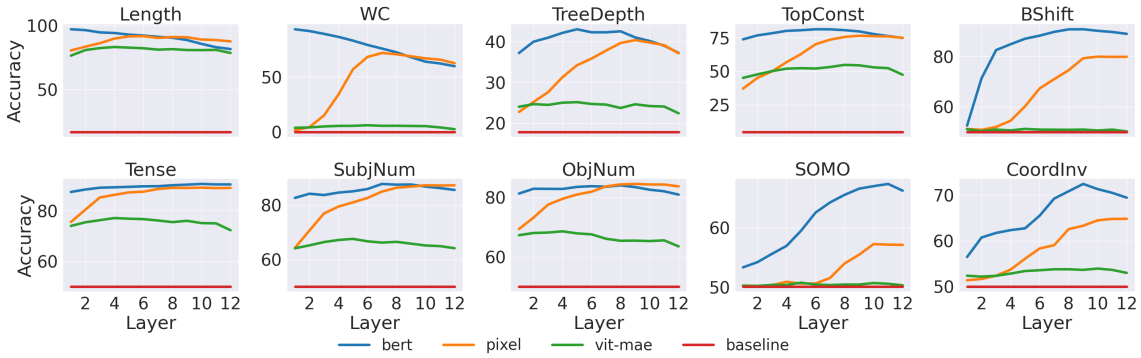

Figure 1: Linguistic probing results on PIXEL, BERT and VIT-MAE, along with the majority baseline.

Figure 2: Example of "cool" being rendered differently in different contexts for PIXEL. The red lines represent patch boundaries.

If the embeddings do not contain any useful information for the task, we would expect the performance to be equivalent to the majority baseline.

The performance of BERT is consistent with what is documented in literature. Surface features are encoded in the lower layers, syntactic features are represented in the middle layers, and semantic features are found in the upper layers (Jawahar et al., 2019). The performance for VIT-MAE for all layers, for most tasks, is very close to the majority baseline. For tasks where some visual information can be useful, for example in `SentLen`, and `Tense` (the visual presence of morpheme *-ed* can be associated with label `PAST`), VIT-MAE performs better than the majority baseline but does not improve or decline through the layers. The performance of PIXEL, when higher than VIT-MAE, can thus be attributed to its linguistic knowledge and not due to having input that is closer to the downstream task.

Across all tasks, PIXEL consistently has an initial monotonic rise in accuracy, starting with a similar performance as VIT-MAE in the lower layers. This indicates that it is using purely visual information in the lower layers, and learns linguistic information in the higher layers. In other words, PIXEL starts as a visual model, and becomes more of a language model through the layers.

However, PIXEL never matches the peak performance of BERT in any layer. This is consistent with the results from Rust et al. (2023), where PIXEL underperforms BERT on the English tasks. We can, therefore, hypothesize that much of PIXEL's capacity is used in recovering from the performance gap between a vision and language model.

Does PIXEL learn syntax and semantics? Unlike BERT, PIXEL does not have a consistent curve across the *surface*, *syntactic* and *semantic* tasks. This is most striking in the *surface* tasks. For BERT, there is an inverse relation between model depth and accuracy. For `SentLen`, the accuracy curve of PIXEL rises until layer 5 and then stagnates. For `WC`, on the other hand, it has a steep rise in the initial layers until layer 7, where it starts to drop. The task requires a good knowledge of word-level features and boundaries - something that is encoded in BERT already at the input level, but PIXEL has to learn during training. We illustrate this further in Figure 2. The patches encoding the word "cool" differ when used in the context of a sentence compared to when it is rendered alone Thus, it may take more layers for PIXEL to reconcile the two different embeddings as the same word. Lotz et al. (2023) have also commented on this phenomenon and linked it to poor downstream semantic performance. They also found that PIXEL-based language models form better contextualised word representations in the upper layers of the model.

We can extrapolate this phenomenon to explain the initial monotonic rise in other tasks. For *syntactic* tasks, PIXEL peaks at layer 9, later than BERT, then stagnates or declines. This delay leads to a delayed learning of higher level abstractions, suggesting that PIXEL needs more layers to match BERT's performance. We leave this question to future work.

The performance across the *surface semantic* tasks for PIXEL shows some consistency. There is a steep rise until layer 3, after which the curve has a more gradual rise, crossing BERT accuracy in the higher layers. For *complex semantic* tasks, both PIXEL and BERT achieve peak performance be-

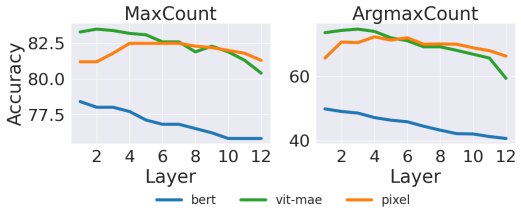

Figure 3: Visual probing results.

tween layers 9 and 12. However, the performance gap between the two is substantial, indicating that PIXEL does not learn semantic abstractions at the same level as BERT. This is also substantiated by the difference in the downstream performance gap between BERT and PIXEL for syntactic and semantic tasks, mentioned in Table 2. PIXEL's performance on dependency parsing and POS-tagging is very close to BERT, while its performance on GLUE, which contains tasks requiring more semantic understanding, is about 6% lower.

The drop in performance for surface tasks in the higher layers also indicates that PIXEL forgets some surface level information as it learns more linguistic abstractions. We substantiate this further with the results on the visual probing tasks below.

5.2 RQ2: How much visual knowledge does PIXEL have?

We investigate this question by first probing PIXEL on the visual tasks introduced in §3.2 to understand whether it is indeed forgetting the surface level information in the higher layers. Results are shown in Figure 3.

For both `MaxCount` and `ArgmaxCount`, we see that VIT-MAE has the highest performance in the lower layers, followed by PIXEL and then BERT. BERT performance has a steady decline, much like the *surface* tasks in Figure 1. PIXEL's performance is much closer to VIT-MAE, but it does not have much decline through the layers, leading to a higher performance than VIT-MAE in the higher layers. PIXEL has slight increases in performance in the middle layers, analogous to the performance peaks in *surface* tasks. The substantially higher performance than BERT, combined with the similarity to VIT-MAE performance, indicates that PIXEL still retains much surface level information in the higher layers. PIXEL's high performance on *surface semantic* tasks in the higher layers also substantiate this since PIXEL has access to both surface and semantic information.

Model	Accuracy
PIXEL	0.52
VIT-MAE	0.83
Random Model	0.42

Table 3: Results for PIXEL, VIT-MAE and VIT-MAE with randomised weights fine-tuned on CIFAR100 for image classification.

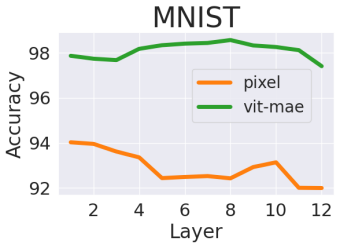

Figure 4: MNIST probing results

Can PIXEL be a vision model? If PIXEL still retains much surface information in the higher layers, is it able to perform well on vision tasks? To investigate this question, we present fine-tuning results for PIXEL and VIT-MAE in Table 3. If PIXEL performs competitively, it implies that PIXEL is fundamentally a vision model that has acquired some language understanding. We also fine-tune a transformer of the same size with randomized weights as a lower bound baseline.

The performance gap between PIXEL and VIT-MAE on image classification is analogous to the performance gap between the two on the GLUE tasks in Table 2. Thus, even though PIXEL is a vision transformer and it retains much surface level information, its pre-training regime on language has lead to a substantially worse performance on image classification, much closer to the random baseline than to VIT-MAE. It can be argued that PIXEL's poorer performance on CIFAR-100 is due to a domain mismatch, stemming from its pre-training on black-and-white text, which offers limited exposure to the color and complexity of the input.

To disentangle this, we probe PIXEL on MNIST at every layer. The results are in Figure 4. The curves for PIXEL are consistent with the curves in Figure 3 and *surface* tasks in Figure 1, in that there is a performance decline through the layers. The difference is that PIXEL performance declines immediately after layer 1, and unlike Figure 1, it is at a lower accuracy than VIT-MAE in the lower layers. Thus, even on input that is similar to the data that PIXEL was pre-trained on, PIXEL does not match VIT-MAE performance.

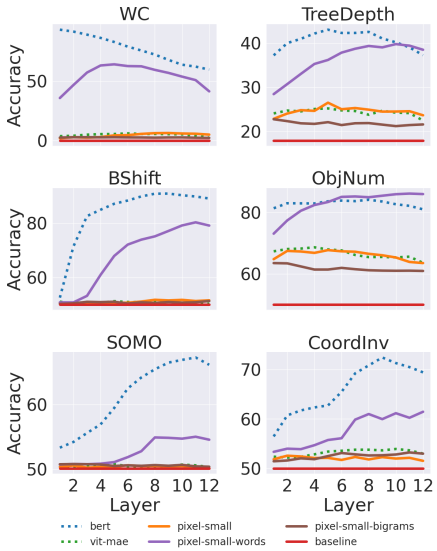

Figure 5: Selected linguistic probing results for small PIXEL variants, base models are indicated with dotted lines.

5.3 RQ3: Does adding orthographic constraints to the input enhance the linguistic capabilities in PIXEL?

Results from §5.1 and §5.2 establish that PIXEL learns surface level information in the lower layers, which leads to delayed learning of higher level semantics. This raises the question of how the gap between visual and linguistic understanding in layers 1 - 6 (the layer with peak performance on surface tasks) can be bridged earlier in the model. Encoding words with differing visual patch representations, as shown in Figure 2, can be made easier by ensuring consistent rendering of words across contexts. The added constraints to the rendering in the PIXEL-variants may lead to a faster learning of surface level information and word boundaries in the lower layers, as discussed in §5.1, thereby making PIXEL behave more like BERT. This idea is further justified by the fact that PIXEL-bigrams and PIXEL-small-words have better downstream performance than PIXEL. Probing results on selected tasks for PIXEL-small, PIXEL-small-words, and PIXEL-small-bigrams are in Figure 5. We also include BERT-base and VIT-MAE-base in the graphs for reference.

At the small scale, PIXEL suffers an almost catastrophic decline in probing performance, showing no more linguistic understanding than VIT-MAE. Similarly, PIXEL-small-bigrams also does not demonstrate any meaningful linguistic understanding. PIXEL-small-words, on the other hand, displays probing performance comparable to PIXEL-base, even at the small scale. It starts with much higher accuracy than VIT-MAE in layer 1 - indicating that there is already linguistic information present in the initial layers due to the imposed structure at the input level. It also achieves peak performance in most tasks earlier than PIXEL-base.

Specifically, for WC, the accuracy rises only until layer 4 before it declines. The curves for *syntactic* tasks are more similar to BERT, with the lower layers achieving scores higher than PIXEL. A combination of visual and some semantic understanding leads to scores for *surface semantic* tasks being even higher than BERT in the upper layers. For *complex semantic* tasks, however, the curve rises until layers 7-8 and then plateaus, indicating higher semantic abstractions are still not being learnt by the model.

Since PIXEL-small-bigrams and PIXEL-small do not have any meaningful linguistic representations at the small scale, we compare the base versions of the two models on the linguistic probes and find similar trends. PIXEL-bigrams at both the base and small scale performs worse than PIXEL. Specific results and analysis can be found in Appendix C.

Why is fine-tuned PIXEL-bigrams better than fine-tuned PIXEL? The observation above is at odds with the downstream performance of PIXEL-bigrams, which Lotz et al. (2023) found to be better than PIXEL. To understand this discrepancy, we run the linguistic probes on fine-tuned versions of the models. We fine-tuned the PIXEL-base-bigrams model on UD parsing (syntactic) and MNLI (semantic) with the same hyper-parameter setup as PIXEL, and compare them to the fine-tuned PIXEL models made available by Rust et al. (2023) on the same tasks. Results are in Figure 6 and Figure 7.

We see that across all probing tasks, fine-tuned PIXEL-bigrams demonstrates better performance than fine-tuned PIXEL. Merchant et al. (2020) found that finetuning BERT on dependency parsing shows effects throughout the model, but MNLI only affects the top layers. Moreover, fine-tuning can cause the model to potentially forget some linguistic knowledge. Mehrafarin et al. (2022) also echoed that fine-tuning on tasks with larger data sizes (like MNLI) can lead to loss of linguistic information in the pre-trained encodings.

We see this trend in PIXEL, where both UD and MNLI fine-tuning decrease probing performance on BShift and SOMO. There is a slight decline in performance on all other probing tasks with UD

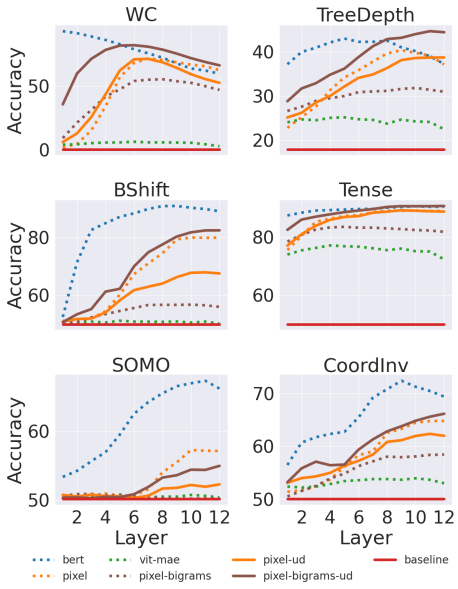

Figure 6: Selected probing results for PIXEL and PIXEL-bigrams finetuned on UD.

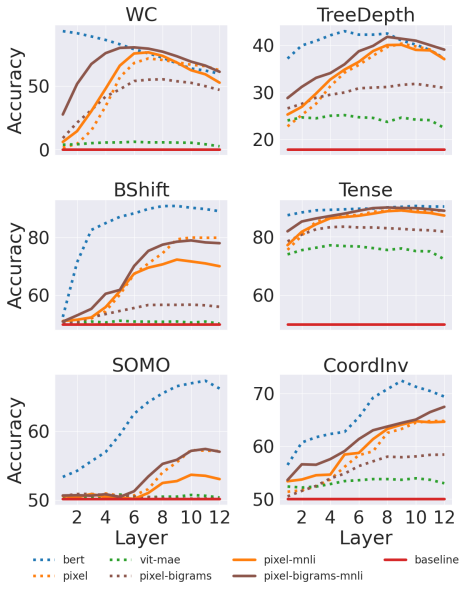

Figure 7: Selected probing results for PIXEL and PIXEL-bigrams finetuned on MNLI.

fine-tuning, but with MNLI fine-tuning, the performance remains similar to pre-trained PIXEL.

We observe the contrary with PIXEL-bigrams. Both UD and MNLI fine-tuning have enhanced the linguistic knowledge encoded in all the layers, with probing performance compared to PIXEL-bigrams pre-trained being much higher. Additionally, UD fine-tuning particularly increases probing performance on *syntactic* tasks in the top layers, and MNLI fine-tuning similarly increases probing performance on the *complex semantic* tasks in the top layers.

Thus, we can speculate that the inductive bias learnt during fine-tuning creates better linguistic representations in PIXEL-bigrams.

5.4 Summary of Findings

On the spectrum of vision and language, it can be concluded from the results of **RQ1** and **RQ2** that PIXEL is more of a language model than a vision model. The difference in downstream performance between PIXEL and VIT-MAE is much larger than PIXEL and BERT. Although with a lower accuracy, PIXEL's behaviour with linguistic probing is more similar to BERT than VIT-MAE.

However, much of PIXEL's linguistic understanding is surface level. The lower layers in PIXEL learn surface level information, as demonstrated by the visual and linguistic probes. The linguistic knowledge acquired in the upper layers demonstrates some syntactic understanding, but does not capture very strong semantic information. This indicates that adding more layers to the model could allow the model to have better semantic representations.

While that is a solution on the architecture side, from **RQ3** results we can conclude that on the input side PIXEL-words comes out to be the current best solution to bridging the gap between visual and language understanding in the lower layers in the model. Nevertheless, it still lacks in semantic understanding, and as Lotz et al. (2023) have noted it is not very efficient to train.

PIXEL-bigrams has worse linguistic probing performance than PIXEL, but the inductive bias it learns during fine-tuning dramatically improves the linguistic knowledge encoded in its layers. PIXEL, on the contrary, forgets some linguistic information during fine-tuning.

6 Conclusion

This study is a first step towards understanding the language modelling capabilities of pixel-based models. Although these models exhibit substantial linguistic understanding, the nature of image-text representations leads to a gap in visual and linguistic understanding. Pixel-based models need to learn the discrete representations that subword-based models already have access to at the input level. Adding orthographic constraints to the input can help bridge this gap, but further architectural modifications could improve these models more, which is a promising direction for future work.

7 Limitations

Our main approach to understanding the linguistic information encoded in pixel-based language models is probing. We acknowledge that although this is our primary method of inquiry, it comes with its flaws. Belinkov and Glass (2019) have noted that even though certain information is detected by a probe as being present in the embeddings, it does not necessarily imply that the information is used by the model. They also remark that using a deeper auxiliary classifier for the probe may lead to better results. There are other criticisms of the approach like Hewitt and Liang (2019) that question whether the probe uncovers information encoded in the embedding, or just learns the linguistic task itself that it is trained on. Pimentel et al. (2020) challenge this and present evidence of the former. Zhu and Rudzicz (2020) recommend using a control mechanism to select probes, based on discussions about the dichotomy raised above. Thus, although this does not dismiss the validity of our findings, we note that our results and conclusions should be read with these caveats in mind.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *CoRR*, abs/1608.04207.

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes. *Preprint*, arXiv:1610.01644.

Vladimir Araujo, Andrés Carvallo, Souvik Kundu, José Cañete, Marcelo Mendoza, Robert E. Mercer, Felipe Bravo-Marquez, Marie-Francine Moens, and Alvaro Soto. 2022. Evaluation benchmarks for Spanish sentence representations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6024–6034, Marseille, France. European Language Resources Association.

Dominika Basaj, Witold Oleszkiewicz, Igor Sieradzki, Michał Górszczak, Barbara Rychalska, Tomasz Trzcinski, and Bartosz Zieliński. 2021. Explaining self-supervised image representations with visual probing. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 592–598. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an Efficient Tokenization-Free Encoder for Language Representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Adam Dahlgren Lindström, Johanna Björklund, Suna Bensch, and Frank Drewes. 2020. Probing multimodal embeddings for linguistic properties: the visual-semantic case. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 730–744, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Li Deng. 2012. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 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)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Katarzyna Krasnowska-Kieraś and Alina Wróblewska. 2019. Empirical Linguistic Study of Sentence Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5729–5739, Florence, Italy. Association for Computational Linguistics.

A. Krizhevsky and G. Hinton. 2009. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. XLM-V: Overcoming the vocabulary bottleneck in multilingual masked language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13142–13152, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Jonas Lotz, Elizabeth Salesky, Phillip Rust, and Desmond Elliott. 2023. Text rendering strategies for pixel language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10155–10172, Singapore. Association for Computational Linguistics.

Xiaofei Ma, Zhiguo Wang, Patrick Ng, Ramesh Nallapati, and Bing Xiang. 2019. Universal Text Representation from BERT: An Empirical Study. *arXiv preprint*. ArXiv:1910.07973 [cs].

Houman Mehrafarin, Sara Rajaee, and Mohammad Taher Pilehvar. 2022. On the importance of data size in probing fine-tuned models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 228–238, Dublin, Ireland. Association for Computational Linguistics.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

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.

Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. 2023. Language modelling with pixels. In *The Eleventh International Conference on Learning Representations*.

Gözde Gül Şahin, Clara Vania, Ilia Kuznetsov, and Iryna Gurevych. 2020. LINSPECTOR: Multilingual probing tasks for word representations. *Computational Linguistics*, 46(2):335–385.

Elizabeth Salesky, David Etter, and Matt Post. 2021. Robust open-vocabulary translation from visual text representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7235–7252, Online and Punta Cana,

Dominican Republic. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Yi Tay, Vinh Q. Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. 2022. Charformer: Fast character transformers via gradient-based subword tokenization. In *International Conference on Learning Representations*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2).

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Zining Zhu and Frank Rudzicz. 2020. An information theoretic view on selecting linguistic probes. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9251–9262, Online. Association for Computational Linguistics.

Zining Zhu, Soroosh Shahtalebi, and Frank Rudzicz. 2022. Predicting fine-tuning performance with probing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11534–11547, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Visual Tasks

Max Count Character (MaxCount) Every letter from the random words in an example is counted. Per example, these raw counts $\hat{f}_\ell$ for each letter ℓ are first normalised to obtain the fractions $f_\ell = \hat{f}_\ell / \sum_{\ell'} \hat{f}_{\ell'}$. We compute $\max_\ell \hat{f}_\ell$ and split the results into 4 uniformly occurring contiguous bins. The task is to predict this bin given the sentence. Examples where multiple letters have the same maximal count are excluded to ensure that the probing task can only be solved by noticing one particular character. We exclude examples with less than 3 unique characters. Details about labels and frequency of each bin are in Table 4

MaxCount		
Bin	**Labels in Bin**	**Bin Size**
1	['3', '4', '5', '6', '7', '8', '9']	21162
2	['10', '11', '12', '13', '14']	21162
3	['15', '16', '17', '18', '19']	26597
4	['20', '21', '22', '23', '24', '25', '26', '27', '28', '29', '30', '31', '32', '33', '34', '35', '36', '37', '38']	25333
	Total	94,254

Table 4: Bin sizes, labels in bin and total data size for `MaxCount`. The labels correspond to the count of the character with the maximum frequency in an example.

Argmax count character (ArgmaxCount) We count the letters in each example again, but now the task is to predict $\ell^* = \arg\max_\ell \hat{f}_\ell$ given the example. The same examples are excluded as above (meaning the argmax is unique), and we skip examples where the argmax is not one of the 26 lowercase Latin letters $\{a, b, \ldots, z\}$. To mitigate against the strong skew towards higher-frequency letters (e, t, a, ...), letters are grouped into an approximation of a uniform distribution of 5 bins (without contiguity constraint) after which the bins are subsampled to have the same amount of sentences as the smallest bin. Details about labels and frequency of each bin are in Table 5.

ArgmaxCount		
Bin	**Labels in bin**	**Bin Size**
1	['b', 'c', 'd', 'f', 'g', 'h', 'k', 'l', 'm', 'p', 'r', 't', 'u', 'w', 'y', 'z']	9413
2	['n', 'o', 's']	9490
3	['i']	9816
4	['a']	11784
5	['e']	9900
	Total	50,403

Table 5: Bin sizes, labels in bin and total data size for `ArgmaxCount`. The labels correspond to the character with the maximum frequency in an example. 'e' has been subsampled from 59,497 to 9900 to ensure a relatively uniform distribution across bins.

B Model Parameters

The models used in this study along with their parameter sizes are in Table 6.

Models	
Name	**Parameters**
BERT	110M
ViT-MAE	112M
PIXEL	86M
PIXEL-bigrams	86M
PIXEL-small	22M
PIXEL-small-bigrams	22M
PIXEL-small-words	22M

Table 6: Size of the probed models

C PIXEL vs PIXEL-bigrams

Select probing results for PIXEL and PIXEL-bigrams base models are in Figure 8. As observed, PIXEL-bigrams performs worse than PIXEL across all probing tasks. We theorize that even though bigrams rendering imposes some structure on the input text, it results in a loss of word boundary information and longer sequences. The rendering strategy adds extra space even within words to ensure that one patch has only two characters, and creates more ambiguity about the structure of the word. This is most prominently seen in the tasks that test for word level information within a sentence - namely, `WC`, `BShift` and `SOMO`. For the later 2, PIXEL-bigrams barely outperforms the majority baseline.

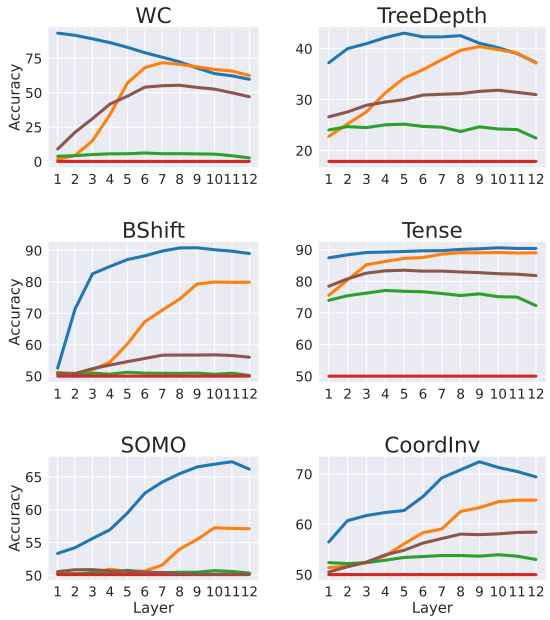

Figure 8: Selected linguistic probing results for PIXEL (orange), PIXEL-bigrams (brown), BERT (blue) and VIT-MAE (green).