
Morris Alper*, Michael Fiman*, Hadar Averbuch-Elor
Tel Aviv University

Abstract

Most humans use visual imagination to understand and reason about language, but models such as BERT reason about language using knowledge acquired during text-only pretraining. In this work, we investigate whether vision-and-language pretraining can improve performance on text-only tasks that involve implicit visual reasoning, focusing primarily on zero-shot probing methods. We propose a suite of visual language understanding (VLU) tasks for probing the visual reasoning abilities of text encoder models, as well as various non-visual natural language understanding (NLU) tasks for comparison. We also contribute a novel zero-shot knowledge probing method, Stroop probing, for applying models such as CLIP to text-only tasks without needing a prediction head such as the masked language modelling head of models like BERT. We show that SOTA multimodally trained text encoders outperform unimodally trained text encoders on the VLU tasks while being underperformed by them on the NLU tasks, lending new context to previously mixed results regarding the NLU capabilities of multimodal models. We conclude that exposure to images during pretraining affords inherent visual reasoning knowledge that is reflected in language-only tasks that require implicit visual reasoning. Our findings bear importance in the broader context of multimodal learning, providing principled guidelines for the choice of text encoders used in such contexts.

1. Introduction

Humans are multimodal learners. We communicate with each other about things that we have experienced and knowledge we have gained using our senses—most commonly including sight as well as hearing, touch, smell, and taste. Our communication channel is limited to a single modality—spoken language, signed language, or text—but a reader or listener is expected to use his or her imagination to visualize and reason about the content being described. In general, language is used to describe scenes, events, and images; the words used to describe these are used to conjure up a visual impression in the listener. Therefore, it is natural to consider the types of visual reasoning used in understanding language, and to ask how well we can currently

---

*These authors contributed equally to this work

1Our code will be made available at https://isbertblind.github.io/
Consider, for instance, the questions in Figure 1. Concreteness is typically correlated with how well a concept can be visually imagined. For example, a concrete word such as present often has a unique visual representation. In addition, common associations such as ocean→blue (color) and corn chip→triangle (shape) reflect properties of an imagined visual representation of the item in question. These properties may be difficult to infer from text alone without prior knowledge gained from visual input; for instance, a number of studies have investigated the partial ability of blind English speakers to predict color associations and how it differs from the intuition of sighted speakers2 [40, 50, 51, 54, 65].

There has been a wealth of recent research vision-and-language (V&L) tasks involving both text and image data, and the use of vision-language pretraining (VLP) to create models that are able to reason jointly about both of these modalities together [11, 12, 29, 35]. Notable in this regard is CLIP [46], consisting of paired text and image encoders jointly trained on a contrastive objective, that learns to align text and image embeddings in a shared semantic space. On the other hand, text encoder models such as BERT [14] learn to reason about text in a unimodal vacuum, with knowledge derived from pretraining tasks that only involve textual data.

Prior work has investigated the performance of multimodally trained text encoders on various natural language understanding (NLU) tasks with mixed results, sometimes finding that they are outperformed by unimodally trained models [22] and at other times suggesting improved performance [69]. However, these works fine-tune the models under consideration on NLU tasks before evaluation, making it difficult to disentangle the effects of multimodal pretraining and fine-tuning configuration on the observed performance. Additionally, these works do not address the distinction between NLU tasks requiring implicit visual reasoning and ones that are purely non-visual. We refer to natural language inference involving implicit visual reasoning as visual language understanding (VLU) and propose a suite of VLU tasks that may be used to evaluate visual reasoning capabilities of pretrained text encoders, focusing primarily on zero-shot methods.

We compare multimodally trained text encoders such as that of CLIP to BERT and other unimodally trained text encoders, evaluating their performance on our suite of VLU tasks. We evaluate these models in without modifying their internal weights in order to probe their knowledge obtained during pretraining. A key design aspect of these tests is the probing method used to evaluate knowledge. Previous work has probed the knowledge of BERT and similar models using a masked language modelling (MLM) paradigm [43, 48], but this cannot be directly applied to CLIP since it was not pretrained with MLM. We therefore propose a new zero-shot probing method that we term Stroop probing. This is based on the psychological Stroop effect [39] (described in Section 3.2), which suggests that salient items should have a stronger interference effect on the representation of their context.

Strikingly, we find that the multimodally trained text encoders under consideration outperform unimodally trained text encoders on VLU tasks, both when comparing to much larger encoders as well as ones of comparable size. We also compare these models on baseline NLU tasks that do not involve visual reasoning and find that models such as CLIP underperform on these tasks, demonstrating that they do not have a global advantage on NLU tasks. We conclude that exposure to images during pretraining improves performance on text-only tasks that require visual reasoning. Furthermore, our findings isolate the effect of the text component of multimodal models for tasks such as text to image generation, providing principled guidelines for understanding the knowledge that such models inject into downstream vision tasks.

2. Related Work

Building models to create text embeddings for a large range of language tasks has been broadly explored over the past several years. In our work we compare two types of transformer-based models which encode these types of embeddings: those trained on text-only data (unimodally trained), and those exposed to both text and image data during training (multimodally trained).

Since the introduction of the self-attention-based transformer architecture by Vaswani et al. [66], transformers have become the predominant architecture for tasks involving textual data. Devlin et al. [14] introduce BERT, a transformer model encoding contextual information for each token in an input sequence in a bidirectional manner. They suggest a self-supervised pretraining method to compute contextual feature representations of textual data, via masked language modelling and next sentence prediction objectives. This pretrained model can then be applied to other downstream tasks by end-to-end fine-tuning. Subsequently, various other text encoder transformers have been proposed, such as RoBERTa [37], DistilBERT [49], and ERNIE [60, 61]. While these models differ on some architectural details and on precise pretraining objectives, they all share the basic transformer architecture and the use of denoising pretraining objectives. In particular, they are all trained on unimodal data, meaning that they are only exposed to text during training.

In contrast to unimodally trained text encoders, V&L
models have been exposed to both text and image data during training. These models are typically used for tasks that require joint reasoning on text and images, such as visual question answering, grounding referring expressions, and vision-language retrieval [11, 16]. Fusion encoder models such as LXMERT [62], UNITER [12], ViLT [29], and ALBEF [35] output a fused representation of text and image data, while dual encoder models like CLIP [46] and ALIGN [24] consist of dual text and image encoder models that are jointly trained to produce embeddings in a shared semantic space. FLAVA, a vision-and-language model introduced recently by Singh et al. [58], also includes dual text and image encoders, and is trained on both multimodal objectives involving the alignment of images and text, as well as unimodal objectives. In this work we focus on the text encoder component of dual encoder models, since it may be used after V&L pretraining for text-only tasks.

Various works have explored the use of multimodal learning to benefit text understanding. Most related to our work is the very recent study of Zhang et al. [72] which investigates the use of unimodal and multimodal models for understanding visual commonsense in text. Their analysis follows a line of related work investigating the contribution of multimodal learning to visual commonsense knowledge in text, since such knowledge is typically not written explicitly in text but is abundantly present in visual information [32, 36, 67]. Unlike Zhang et al. [72] who only evaluate CLIP with an added set of task-specific learned weights, we are able to probe CLIP and other similar models in the strictly zero-shot setting via our novel Stroop probing method. This allows for directly evaluating properties learned by the models, independent of differences that result, for instance, from specific training configurations. In addition, we also study performance on both visual and non-visual NLU tasks in order to provide a controlled benchmark.

Other works have investigated the use of multimodal learning for NLU in various contexts. Bruni et al. [7] propose an architecture for integrating text and image-based distributional information to improve performance on tasks where meaning is grounded in perception. Kiela and Botou [27] show that integrating features extracted from images using CNN with skip-gram representation vectors improves performance on semantic similarity datasets. Lazari-dou et al. [33] train visual representations extracted using CNNs together with skip-gram embeddings to integrate visual and textual information performing well in both semantic and vision tasks. Kiela et al. [28] train a sentence embedding model using grounded information extracted from image features by attempting to predict the image features. These embeddings improve performance on various NLP tasks in comparison to text only embeddings. They show that using this method for a dataset consisting mainly of abstract words is likely to less benefit from grounding information. Shi et al. [55] show that a syntactic parser can benefit from seeing images during training; however, it was later shown that the model mostly relies on noun concreteness (which we also elaborate on in our work) rather than more complex syntactic reasoning [31]. The use of images for PCFG induction is also investigated by Jin & Schuler [25].

Along with the rise in visibility of jointly trained V&L transformer models, a number of works have explored the use of these models for text-only tasks, with mixed results. Using terms coined by Sileo et al. [57], these can be broadly split into associative grounding and transfer grounding approaches. Associative grounding uses retrieval methods to associate particular images with related texts; Kiros et al. [30] and Tan & Bansal [63] show that associative grounding methods may improve performance on various text-only NLU benchmark tasks. Transfer grounding applies V&L models directly to text-only input, disregarding the vision component of the model during inference. Wang et al. [69] apply this to weakly-supervised V&L models to outperform BERT on various text-only tasks from the GLUE benchmark. On the other hand, Iki & Aizawa [22] find that V&L-pretrained text encoders have similar or inferior results on NLU tasks including tasks from GLUE. Likewise, Cao et al. [9] find that although visually-aligned text encoders perform well on semantic linguistic probing tasks, BERT still outperforms them.

As discussed above, some prior works suggest that multimodal pretraining aids text understanding while other works show that it can lead to degradation. Our work provides new context for these seemingly contradictory results, allowing them to be reassessed in the new context of visual vs. non-visual natural language understanding.

3. Experimental Setup
3.1. Models Used
For evaluating unimodally trained text encoders, we use BERT [14], RoBERTa [37], DistilBERT and Distil-RoBERTa [49], which are all trained with text-only MLM objectives. We also include results for Sentence-BERT (SBERT) [47], since its output embeddings are trained to have meaningful cosine similarity scores and thus bear more similarity to other models evaluated with Stroop proving. Results on multimodally trained text encoders are reported for CLIP [46] and FLAVA [58]; for these models we use only the text encoder with pretrained weights and discard the other subcomponents. Our tests include checkpoints from both OpenAI and the OpenCLIP open-source implementation of CLIP [13, 23]. Details of the checkpoints used for each model are listed in the supplementary material.

The text encoders of the multimodally trained models range in size from 63M (CLIP) to 109M (FLAVA) param-
eters. We compare to both comparably small unimodally trained text encoders such as DistilBERT (66M parameters) as well as much larger text encoders such as BERT-large (340M). See the supplementary material for an exhaustive list of sizes of the models under consideration.

We use each model with frozen pretrained weights. Our subsequent tests probe the contents of the feature vectors extracted by these models. For MLM probing, we also use the model’s MLM head for prediction. In cases where MLM can be used we have found it to outperform Stroop probing; in such cases we report results for MLM probing here and for Stroop probing in the supplementary material.

3.2. Probing Methods

In order to probe the inherent knowledge of our models, we use the knowledge probing methods described below. The probing methods that follow are strictly zero-shot; in the supplementary material we analyze the use of linear classifiers trained on our models’ frozen embeddings (“linear probing”).

Masked language modelling (MLM). BERT and our other unimodally trained models were all pretrained with MLM objectives and hence can be used for zero-shot prediction of tokens in a masked context. Given a text including a [MASK] token and a set of k possible completions \( C = \{c_1, c_2, \ldots, c_k\} \), a MLM assigns probabilities \( p_1, \ldots, p_k \) to each corresponding token. We use \( \arg \max_i p_i \) as the model’s prediction. Previous works have found that BERT and other MLM can be probed for innate knowledge with this method [43, 48].

Stroop probing (SP). We propose another zero-shot probing method to extract knowledge from models based on the pooled embeddings that they extract. Consider a masked text \( t_m \) and possible completions \( c \in C \), and let \( t_c \) be the text with \( c \) inserted in the mask location. Given a text encoder \( M \), we calculate pooled embeddings \( v_m = M(t_m) \) and \( v_c = M(t_c) \) and unit-normalize them to \( \hat{v}_m = v_m / \|v_m\| \) and \( \hat{v}_c = v_c / \|v_c\| \). Stroop probing considers the cosine similarity scores \( s_c := \hat{v}_m \cdot \hat{v}_c \). These can be used either directly for regression (as in the concreteness task below), or for categorical prediction by selecting \( c^* = \arg \max_c s_c \).

The intuition behind Stroop probing is that items which are more surprising, incongruous, or salient in the given context may have a stronger interference effect on the encoding of the surrounding text. This is analogous to the Stroop effect in human psychology. When presented with congruent and incongruent stimuli such as color words printed in the same or differing colors (e.g. “red” printed in blue), readers take significantly longer on average to read the incongruent stimuli, a phenomenon known as the Stroop effect [39]. We use Stroop probing for multiple tasks, including predicting color associations, as described below.

3.3. Prompts Used

For each task, we test the probing methods above on a wide variety of prompts in order to show the robustness of the described phenomena. In our results below we report the maximum metric value for each model over all of the prompts, since this represents a rough bound on our ability to extract intrinsic knowledge from the models under consideration. A full list of prompts used for each task and an analysis of model performance across prompts are provided in the supplementary material.

In some cases our prompt contains an empty slot, which we indicate below as \([\ast]\). Some models under consideration have a dedicated mask token, but for those such as CLIP that do not, we insert a fixed token in this slot, detailed further in the supplementary material.

4. VLU Tasks

We present three VLU tasks to probe the ability of our models to understand language with implied visual context: concreteness prediction (Section 4.1), color association prediction (Section 4.2) and shape association prediction (Section 4.3). Note that each of these tasks is performed on text alone, but requires visual reasoning to complete.

4.1. Concreteness Prediction

**Task description.** Words and phrases can be roughly classified as either concrete or abstract. A concrete concept is something that can be pointed to or directly sensed, while abstract concepts refer to things that cannot be easily visualized [53]. This can be conceptualized on a scale, ranging from the most abstract to the most concrete. Psychological research suggests that concrete words are easier for humans to understand and remember than abstract words [53]. Similarly, it has been shown that concreteness correlates with the learnability of visual concepts for machine learning models [18], and that MLM pretraining of V&L models may be improved by preferentially masking concrete words [4].

Because concreteness is a property of text that is tightly coupled with the visual domain, we consider concreteness prediction to be a VLU task, requiring some knowledge of the visual content of language to complete. We note that this task has been addressed in various previous works [10, 18, 20, 21, 45]. In contrast to these approaches, our unsupervised concreteness estimation procedure evaluates the concreteness of a word or phrase in a given textual context, rather than being limited to a fixed set of lexical items or discrete categories in a dataset.

---

3For example, try saying these colors out loud (not the printed words): **Green, Red, Blue, Purple, Red, Purple.**
Experimental details. We probe our models for the concreteness of words in context by using a cloze task paradigm with Stroop probing. For example, using the prompt $t_m = “I see the \{ \ast \}’’$ and testing word $\langle w \rangle$, we insert the word into the prompt to obtain $t_w = “I see the \langle w \rangle’’$, and use cosine similarity score $s_w$ between embeddings of $t_m$ and $t_w$ as the regression output. All prompts used are listed in the supplementary material.

We test our approach on the dataset introduced by Brysbaert et al. [8]. This dataset contains 39,954 English unigrams and bigrams along with human-labelled concreteness scores on a scale from 1 (abstract) to 5 (concrete), averaged over annotators. We only use the unigram nouns from this list, totaling 14,562 items. Note that unlike prior concreteness prediction techniques that train supervised models on this dataset [10], we perform zero-shot prediction on this task with no supervised training, using the dataset for testing only.

Also note that we do not report results for DistilBERT or DistilRoBERTa since the checkpoints used do not contain a trained pooling layer, which is required for Stroop probing.

Evaluation metrics. We report absolute values of Pearson, Spearman, and Kendall correlations between the predicted concreteness and ground truth scores ($|\rho|$, $|r_s|$, and $|\tau|$ respectively).

4.2. Color Association Prediction

Task description. Some concepts are highly associated with particular colors—for example, the word banana is highly associated with the color yellow, while a word like child does not have a strong color association. These color associations have been widely studied in experimental psychology and neuroscience [3,5]. We propose a task of color association prediction—given a noun (or noun phrase) $\langle w \rangle$, identify the color with which $\langle w \rangle$ is normally associated.

Experimental details. To probe our models for color associations, we use the MLM and SP methods described above.

In particular, we conceive of this task as categorical prediction over a set of basic color words $C$. For example, using the prompt “A picture of a $\{ \ast \} \langle w \rangle’’ where $\langle w \rangle$ is the item being tested, our probing methods search for the most suitable color to place in the $\{ \ast \}$ slot. All prompts tested are listed in the supplementary material. For MLM probing, we predict the color $c \in C$ with the highest predicted probability in the $\{ \ast \}$ slot of the prompt. For SP, we predict $c^* = \arg \max s_c$, using similarity scores as defined above.

To test this method on our chosen text encoders, we use two datasets. The Color Terms Dataset (CTD) [6] provides a list of 52 concrete words and their color. The Natural-Color Dataset (NCD) of fruit images [2] is a colorization task containing images of 20 types of fruits and vegetables paired with colors. We use the provided list of fruits and colors as a fixed set of words with strong color associations, discarding the image data. For the latter, we filter objects with the color label purple as this label contains multiple WordPiece tokens and thus is not directly comparable with MLM probing for models such as BERT. This results in 15 unique fruits and vegetables. For each model, we calculate color predictions using the probing methods described above out of the set: \{red, orange, yellow, green, blue, black, white, grey, brown\}.

Evaluation metrics. We report categorical accuracy of predictions on the CTD and NCD datasets ($\text{acc}_{\text{CTD}}$ and $\text{acc}_{\text{NCD}}$) relative to the ground truth color labels.

4.3. Shape Association Prediction

Task description. Another salient visual feature of language is the association between concrete nouns or noun phrases and particular shapes. For example, the nouns wheel and compass have a circular association, while pyramid and corn chip have a triangular association. Shape associations have been studied in the psychological literature in contexts such as child language acquisition [70] and semantic memory representation [68]. Building on this line of research, we propose the task of shape association prediction—given a noun (or noun phrase) $\langle w \rangle$, identify the basic shape that is most associated with $\langle w \rangle$. Because the space of possible shapes is complex and difficult to categorize unambiguously, we restrict $\langle w \rangle$ under consideration to nouns associated with a few basic shapes, as described below.

Experimental details. We construct the ShapeIt benchmark for shape associations\(^4\). This contains 109 items total, each consisting of a noun or noun phrase along with the basic shape most associated with it from the set \{rectangle, circle, triangle\}. The benchmark was constructed by performing a user study requiring users to choose a shape associated with a given word, and selecting for only those words which were consistently classified by the users. Data collection methods used in constructing this benchmark are detailed in the supplementary material, along with further analysis of its contents. Probing methods used for this task are equivalent to the color association prediction task. Prompts used for probing include “A $\{ \ast \}$ shaped $\langle w \rangle’’ where $\langle w \rangle$ is the shape associated word; the full list of prompts used is detailed in the supplementary material. We use both shape nouns (e.g. circle) and associated adjectives (e.g. circular) and report the highest accuracy achieved between these two settings.

Evaluation metric. We report categorical accuracy of predictions (acc) relative to the ground truth shape labels.

\(^4\)available in our code repository
5. Non-visual NLU Tasks

We also present three non-visual NLU tasks to serve as a baseline comparison for our models:

5.1. Factual Knowledge Probing

Task description. It has been observed that language models have an emergent knowledge base property, in which they may be probed for factual knowledge about the world [43, 48]. Various works on probing BERT and other language models for commonsense world knowledge have found that they show an impressive ability to memorize knowledge, although they may be deficient in applying this knowledge to reasoning about the physical world [15]. In this task, we probe our models for fine-grained factual knowledge via a cloze test, where an empty slot must be filled in with a word. We test on factual knowledge about geographical locations since this requires factual knowledge that does not explicitly rely on visual reasoning.

Experimental details. For this task, we use the Comparative Question Completion dataset introduced by [71]. This consists of questions in which one of a pair of coordinated elements is masked; the target is the masked phrase. Specifically, we use the cities dataset which masks the names of geopolitical entities such as cities and countries. Example sentences from the dataset include: which country has more part time jobs new zealand or [•]? (the correct answer being australia) and which is older saudi arabia or [•]? (the correct answer being persia). The original dataset has 1,187 questions with 447 unique locations as answers. In order to fit the general method of masking tasks, we filter masked phrases with more than one token (e.g. the west coast) similar to the protocol presented in the original paper. As this results in an extremely limited set of candidates for MLM models such as RoBERTa that use Byte Pair Encoding tokenization, we restrict the MLM models under comparison to BERT and DistilBERT. The filtered dataset contains 825 questions with 216 unique locations.

We treat this task as a categorical classification task, choosing only from the set of unique locations given in the dataset per sample, and evaluating how often the correct target is chosen. We use MLM probing and Stroop probing for categorical prediction as described above. Similarly to our other tasks, the intuition is that more surprising completions should have a larger interference effect on the text’s encoding, if the relevant information is encoded in the embedding.

Evaluation metrics. We report recall at one and five (R@1, R@5), measuring how often the ground truth answer is found among the model’s top one or five predicted candidates.

5.2. Language Proficiency Probing

Task description. In order to evaluate our models’ intrinsic knowledge of general language usage, we consider the task of reconstructing English text in order to produce natural-sounding language. Multiple-choice cloze tests are commonly used in language assessment tasks for students to evaluate their proficiency [1, 59, 64]. Similarly, a model with a good grasp of English language usage should be able to fill in missing words in cloze contexts to produce fluent English. This requires grammatical and semantic knowledge, but in general, it is not directly related to visual reasoning when applied to arbitrary masked contexts. As noted by Trace [64], cloze tasks may evaluate global reading comprehension or local contextual information in the cloze context; we focus on the latter case and refer to this task applied to our models as language proficiency probing.

Experimental details. To evaluate language proficiency, we use the Children’s Book Test (CBT) cloze dataset provided by Meta research [19], consisting of book passages with accompanying masked sentences and possible mask completions. We discard the book passages and only consider the sentences and completions, to focus on the task of reconstructing well-formed text. Completions are grouped by part of speech (POS); we use the noun (N), verb (V), and preposition (P) groups and discard the named entity groups since the latter require long-distance context to predict while N, V, and P can often be inferred from localsentential context. In total, each of the N, V, and P groups contains 2,500 sentences with 10 possible completions each. We filter out long sentences since our multimodally trained models have shorter expected input lengths. After filtering we are left with 1,588 noun, 1,747 verb, and 2,382 preposition completion sentences. In addition, we only use sentences that have a one-word token answer for all tokenizers. For example, one sentence from the V group is I [•] not a fellow; I am a young lady! and the set of possible completions is \{am, born, find, picking, pricked, said, sat, seems, streamed, thinking\}. We use MLM and Stroop probing to evaluate our models on this data.

Evaluation metrics. We report categorical accuracy per POS group (accV, accN, and accP), measuring how often the ground truth answer is selected in each of these groups.

5.3. Sentiment Analysis

Task description. Sentiment analysis refers to the task of predicting speaker emotion or affect, a well-studied problem in natural language processing [26, 34, 44]. We focus on sentiment analysis in text as a subset of text classification. Since text describing the same visual scene may have a positive or negative sentiment (This cake is delicious vs.
This cake tastes bad), we consider this task to be a non-visual NLU task.

**Experimental details.** For this task, we use the IMDB movie review dataset consisting of 50K movie reviews with binary sentiment labels (positive/negative) [38]. In order to provide a fairer comparison between models rather than biasing towards models trained on longer texts, we use only a single random sentence from each review in the IMDB dataset. In addition, we filter long sentences which are too long for multimodal encoders leaving 42,567 examples. Using only a single sentence makes this task more challenging since the randomly chosen sentence is not guaranteed to contain sufficient context for understanding the review’s sentiment, but we find that significantly better than random performance is achievable, as seen in the results section. We also differ from the more common learned sentiment analysis paradigm by using strictly zero-shot prediction via engineered prompts. For example, one prompt used is “sentiment expressed for the movie is [ ]. (s)”, where ⟨s⟩ indicates the sentence chosen from the initial review, and [ ] may be filled with one of {good, bad}. We apply MLM and Stroop probing for binary prediction, and report categorical accuracy achieved for each model.

**Evaluation metric.** We report categorical accuracy of predictions (acc) relative to the ground truth sentiment labels.

### 6. Results and Discussion

Results for the tasks described above are provided in Table 1. For tasks with multiple prompts the listed metrics are the maximum over prompts, providing a rough upper bound on each model’s ability to perform the task in question. Further analysis of performance by prompt, as well as SP results for models shown here with MLM, are provided in the supplementary materials.

As seen in these results, multimodally trained models consistently outperform unimodally trained models on VLU tasks, including both comparably sized and much larger text encoders, while generally underperforming them on non-visual NLU tasks. This is further illustrated by qualitative analysis of the results in various tasks.

Figure 2 shows the results of concreteness prediction for CLIP and BERT. Nouns predicted as most concrete by CLIP, for example bench and chalk, that can be clearly visualized, while nouns predicted as least concrete (i.e., abstract) such as story and name, do not have a clear visual representation. In comparison, BERT’s predictions are significantly noisier, with nouns such as seed and jelly predicted as abstract.

Figures 3 and 4 shows color and shape association predictions of BERT-base and CLIP on samples from the relevant datasets. Without having access to the associated images, the CLIP text encoder usually predicts the correct matching between the given item and its correct shape or color, while BERT fails in most cases. Our results suggest that these associations are more consistently encoded by multimodally trained encoders. Furthermore, qualitative analysis of the misclassifications of CLIP, OpenCLIP and FLAVA on color association prediction reveals that these are mostly due to ambiguities in the dataset itself; see the supplementary materials for details.

Performance on non-visual NLU tasks, shown on the right side of Table 1, demonstrates that our results are not an artifact of our probing methodology providing a global advantage to multimodally trained models, nor are these models uniformly better at language-related tasks. We also see
Table 1. Results on VLU (left) and non-visual NLU (right) tasks: concreteness prediction, color and shape association prediction, factual knowledge probing, language proficiency probing, and sentiment analysis (Sent.) respectively. For tasks other than concreteness prediction, MLM probing is used for models supporting it (BERT, DistilBERT, RoBERTa, DistilRoBERTa); SP is used elsewhere. The definition of each metric is defined in the relevant task definition in Sections 4-5. DistilBERT and DistilRoBERTa do not have concreteness results due to the pooling layer issue mentioned in Section 4.1, and RoBERTa and DistilRoBERTa do not have results for factual knowledge probing due to the tokenization issue mentioned in Section 5.1. As these results show, V&L models yield superior performance on visual tasks, while underperforming unimodally trained models on non-visual NLU tasks.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Concreteness</th>
<th>Color</th>
<th>Shape</th>
<th>Knowledge</th>
<th>Proficiency</th>
<th>Sent.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$</td>
<td>\rho</td>
<td>$</td>
<td>$</td>
<td>\tau</td>
<td>$</td>
</tr>
<tr>
<td>Unimodal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT-base</td>
<td>0.414</td>
<td>0.416</td>
<td>0.283</td>
<td>0.353</td>
<td>0.400</td>
<td>0.559</td>
</tr>
<tr>
<td>BERT-large</td>
<td>0.348</td>
<td>0.355</td>
<td>0.239</td>
<td>0.490</td>
<td>0.467</td>
<td>0.587</td>
</tr>
<tr>
<td>DistilBERT</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.333</td>
<td>0.400</td>
<td>0.587</td>
</tr>
<tr>
<td>RoBERTa-base</td>
<td>0.433</td>
<td>0.404</td>
<td>0.275</td>
<td>0.431</td>
<td>0.333</td>
<td>0.431</td>
</tr>
<tr>
<td>RoBERTa-large</td>
<td>0.345</td>
<td>0.374</td>
<td>0.253</td>
<td>0.471</td>
<td>0.400</td>
<td>0.431</td>
</tr>
<tr>
<td>DistilRoBERTa</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.411</td>
<td>0.333</td>
<td>0.431</td>
</tr>
<tr>
<td>ERNIE</td>
<td>0.461</td>
<td>0.496</td>
<td>0.338</td>
<td>0.196</td>
<td>0.333</td>
<td>0.449</td>
</tr>
<tr>
<td>ERNIE-large</td>
<td>0.358</td>
<td>0.353</td>
<td>0.233</td>
<td>0.216</td>
<td>0.267</td>
<td>0.458</td>
</tr>
<tr>
<td>SBERT</td>
<td>0.338</td>
<td>0.337</td>
<td>0.228</td>
<td>0.198</td>
<td>0.067</td>
<td>0.513</td>
</tr>
</tbody>
</table>

V&L

| Metric            | $|\rho|$ | $|\tau|$ | $|\tau|_{acc_{CTD}}$ | $|\tau|_{acc_{NCD}}$ | $\text{R@1}$ | $\text{R@5}$ | $\text{acc}$ | $\text{acc}$ | $\text{acc}$ | $\text{acc}$ | $\text{acc}$ |
|-------------------|-------|-------|----------------------|----------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| CLIP              | 0.603 | 0.624 | 0.437                | 0.843                | 0.800       | 0.798       | 0.009       | 0.118       | 0.134       | 0.133       | 0.126       | 0.560       |
| OpenCLIP          | **0.634** | 0.643 | 0.432                | **0.941**            | 0.800       | **0.853**  | 0.009       | 0.121       | 0.211       | 0.135       | 0.123       | 0.560       |
| FLAVA             | 0.608 | **0.665** | 0.449                | 0.882                | 0.800       | 0.798       | 0.020       | 0.138       | 0.116       | 0.118       | 0.139       | 0.519       |

Table 1. Results on VLU (left) and non-visual NLU (right) tasks: concreteness prediction, color and shape association prediction, factual knowledge probing, language proficiency probing, and sentiment analysis (Sent.) respectively. For tasks other than concreteness prediction, MLM probing is used for models supporting it (BERT, DistilBERT, RoBERTa, DistilRoBERTa); SP is used elsewhere. The definition of each metric is defined in the relevant task definition in Sections 4-5. DistilBERT and DistilRoBERTa do not have concreteness results due to the pooling layer issue mentioned in Section 4.1, and RoBERTa and DistilRoBERTa do not have results for factual knowledge probing due to the tokenization issue mentioned in Section 5.1. As these results show, V&L models yield superior performance on visual tasks, while underperforming unimodally trained models on non-visual NLU tasks.

that the non-visual tasks are highly solvable, with BERT-large and RoBERTa-large achieving high performance on all tasks despite the challenging zero-shot regime and limited information in the task inputs (ambiguity in cloze contexts for factual probing, lack of textual context for proficiency probing and randomly-chosen sentences for sentiment analysis). Despite this, the multimodally trained models show near-random performance.

We note a direct connection to the original Stroop effect in the field of human psychology. Follow-up works to the first Stroop effect demonstration have found it to apply to various types of stimuli, such as color-congruent and incongruent objects (e.g. a yellow banana vs. a purple banana) [41]. Our results, also including color congruence of objects, strengthen the motivation for using Stroop probing applied to tasks involving visual congruence or saliency.

We also note a connection between our results and the reporting bias effect, in which commonsense properties are less likely to be explicitly stated than incongruent properties (e.g. a yellow banana vs. a blue banana). Reporting bias in text has been studied in the context of color associations [42] and in more general contexts [17, 56]. As the multimodally trained models under consideration were trained on paired image-caption data, the distribution of text in image captions differs somewhat from the text used for training models such as BERT. In the supplementary material, we provide an analysis of reporting bias in the LAION dataset [52], the training data for the OpenCLIP model included in our tests. These results provide evidence that the improvement in performance seen from V&L training cannot primarily be attributed to a lack of reporting bias in image caption texts, and emphasizes the significance of the visual modality in these models’ language understanding.

7. Conclusion

We propose a suite of visual language understanding tasks along with non-visual natural language understanding tasks to probe the effect of V&L pretraining on such reasoning capabilities of text encoder models. We introduce Stroop probing as a zero-shot knowledge proving method for evaluating the innate knowledge of text encoders. We also show that exposure to V&L data in pretraining improves the performance of text encoder models on VLU tasks, even though they may underperform unimodally trained text encoders on non-visual NLU tasks. Beyond text-only tasks, these results bear importance in the broader context of multimodal learning, in which the isolated contribution of text encoders has previously been underexplored. Our findings suggest that multimodal pretraining has a significant effect on the knowledge represented by the text encoder component of multimodal models, facilitating in establishing best practices for the design and training of text encoders used in such contexts.

Acknowledgements. We thank Noriyuki Kojima and Gabriel Stanovsky for their helpful feedback.
References


[71] Avishai Zagoury, Einat Minkov, Idan Szpektor, and William W Cohen. What’s the best place for an ai conference, vancouver or \&