CLIPPO: Image-and-Language Understanding from Pixels Only

Michael Tschannen, Basil Mustafa, Neil Houlsby
Google Research, Brain Team, Zürich

Abstract

Multimodal models are becoming increasingly effective, in part due to unified components, such as the Transformer architecture. However, multimodal models still often consist of many task- and modality-specific pieces and training procedures. For example, CLIP (Radford et al., 2021) trains independent text and image towers via a contrastive loss. We explore an additional unification: the use of a pure pixel-based model to perform image, text, and multimodal tasks. Our model is trained with contrastive loss alone, so we call it CLIP-Pixels Only (CLIPPO). CLIPPO uses a single encoder that processes both regular images and text rendered as images. CLIPPO performs image-based tasks such as retrieval and zero-shot image classification almost as well as CLIP-style models, with half the number of parameters and no text-specific tower or embedding. When trained jointly via image-text contrastive learning and next-sentence contrastive learning, CLIPPO can perform well on natural language understanding tasks, without any word-level loss (language modelling or masked language modelling), outperforming pixel-based prior work. Surprisingly, CLIPPO can obtain good accuracy in visual question answering, simply by rendering the question and image together. Finally, we exploit the fact that CLIPPO does not require a tokenizer to show that it can achieve strong performance on multilingual multimodal retrieval without modifications.

1. Introduction

In recent years, large-scale multimodal training of Transformer-based models has led to improvements in the state-of-the-art in different domains including vision [2, 10, 74–76], language [6, 11], and audio [5]. In particular, in computer vision and image-language understanding, a single large pretrained model can outperform task-specific expert models [10, 74, 75]. However, large multimodal models often use modality or dataset-specific encoders and decoders, and accordingly lead to involved protocols. For example, such models frequently involve training different parts of the model in separate phases on their respective datasets, with dataset-specific preprocessing, or transferring different parts in a task-specific manner [75]. Such modality and task-specific components can lead to additional engineering complexity, and poses challenges when introducing new pretraining losses or downstream tasks. Developing a single end-to-end model that can process any modality, or combination of modalities, would be a valuable step for multimodal learning. Here, we focus on images and text.

A number of key unifications have accelerated the progress of multimodal learning. First, the Transformer architecture has been shown to work as a universal backbone, performing well on text [6, 15], vision [16], audio [5, 24, 54], and other domains [7, 34]. Second, many papers have explored mapping different modalities into a single shared embedding space to simplify the input/output interface [21, 22, 46, 69], or develop a single interface to many tasks [31, 37]. Third, alternative representations of modalities allow harnessing in one domain neural architectures or training procedures designed for another domain [28, 49, 54, 60]. For example, [60] and [28, 54] represent text and audio, respectively, by rendering these modalities as images (via a spectogram in the case of audio).

In this paper, we explore the use of a pure pixel-based
model for multimodal learning of text and images. Our model is a single Vision Transformer [16] that processes visual input, or text, or both together, all rendered as RGB images. The same model parameters are used for all modalities, including low-level feature processing; that is, there are no modality-specific initial convolutions, tokenization algorithms, or input embedding tables. We train our model using only a single task: contrastive learning, as popularized by CLIP [56] and ALIGN [32]. We therefore call our model CLIP-Pixels Only (CLIPPO).

We find that CLIPPO performs similarly to CLIP-style models (within 1-2%) on the main tasks CLIP was designed for—image classification and text/image retrieval—despite not having modality-specific towers. Surprisingly, CLIPPO can perform complex language understanding tasks to a decent level without any left-to-right language modelling, masked language modelling, or explicit word-level losses. In particular, on the GLUE benchmark [73] CLIPPO outperforms classic NLP baselines, such as ELMo+BiLSTM+attention, outperforms prior pixel-based masked language models [60], and approaches the score of BERT [15]. Interestingly, CLIPPO obtains good performance on VQA when simply rendering the image and text together, despite never having been pretrained on such data.

Pixel-based models have an immediate advantage over regular language models because they do not require pre-determining the vocabulary/tokenizer and navigating the corresponding intricate trade-offs; consequently, we observe improved performance on multilingual retrieval compared to an equivalent model that uses a classical tokenizer.

2. Related work

Multimodal and contrastive pretraining Most closely related to CLIPPO are CLIP [57] and ALIGN [32] which developed the paradigm of large-scale contrastive training on noisy data from the web. Follow-ups [55, 85] have scaled further and employed state-of-the-art image representation learning to boost performance.

A number of works have explored model unification via weight-sharing. In the contrastive context, LIMoE [53] and MS-CLIP [80] explore a one-tower model similar to ours, studying the use of mixture of experts and selective sharing of modules, respectively. Outside contrastive training, co-training distinct tasks [1, 46] is a popular strategy, with some approaches [44] involving knowledge distillation and gradient masking. Other works use self-supervised learning algorithms to unify task training [21]. These broadly use discriminative tasks to learn representations for various downstream modalities; generative approaches to multimodal modelling have been scaled to billions of parameters, generating text [2, 10, 74, 82], images [58, 62, 83], videos [27, 72] or audio [5] from various modalities.

Another related domain is document and user interface (UI) understanding. Corresponding models are trained on diverse multimodal data sets and can usually solve a range of document/UI understanding tasks. Many models rely on text extracted using an off-the-shelf OCR pipeline in combination with document images [3, 29], but image-only models are getting more popular [35, 41]. While these models can understand visual cues and text from the input image, they still rely on a tokenized text for training and inference.

Contrastive training in NLP There is a sizable body of work on contrastive pretraining on sentence pairs (see [59] for a recent survey), which we explore as an auxiliary objective for CLIPPO. Popular augmentations to generate text pairs involve word deletion, span deletion, reordering, synonym substitution, and next-sentence-prediction [20, 47, 77]. Other methods use different realizations of dropout masks in the model to emulate sentence pairs, or supervised labels to obtain positive and negative pairs [19].

Visual text and tokenization in NLP The most closely related method to CLIPPO from the NLP domain is PIXEL [60], which is a masked autoencoder (MAE) [26] trained on rendered text. It obtains strong performance on multilingual syntactic (part-of-speech tagging, dependency parsing) and semantic language understanding (named entity recognition, sentence understanding) tasks, while being more robust to noise in the text than BERT. Other applications for which visual text has been explored include sentiment analysis [68] and machine translation [49, 63].

Visual text side-steps the design and construction of an appropriate tokenizer, which is a large area of research of its own, and can hence simplify text processing in certain—in particular multilingual—scenarios. We refer to [52] for a survey on tokenizers. Popular models include WordPiece [15], Byte-Pair Encoding [65], and SentencePiece [39].

Subword-based vocabularies are popular in monolingual setups and usually lead to a good performance trade-off compared to word and character based vocabularies for certain languages including English. In multilingual contexts, appropriately representing the vocabulary of all languages becomes challenging as the number of languages increases [13, 61], which in turn can lead to poor performance in tasks involving underrepresented languages. A variety of mitigation strategies has been developed; we refer to [60, Sec. 5.1] for a more detailed discussion of these strategies.

3. Contrastive language-image pretraining with pixels

Contrastive language-image pretraining has emerged as a powerful, scalable paradigm to train versatile vision models on web-scale data sets [56]. Concretely, this approach relies on image/alt-text pairs which can be automatically collected at large scale from the web. Thereby, the textual
descriptions are usually noisy, and can e.g. consist of single keywords, sets of keywords, or potentially lengthy descriptions with many attributes describing the image content. Using this data, two encoders are jointly trained, namely a text encoder embedding the alt-texts and an image encoder embedding the corresponding images into a shared latent space. These two encoders are trained with a contrastive loss, encouraging the embeddings of matching images and alt-text to be similar, and at the same time to be dissimilar from all other image and alt-text embeddings.

Once trained, such an encoder pair can be used in many ways: It can be specialized to classifying a fixed set of visual concepts via their textual descriptions (zero-shot classification); the embeddings can be used to retrieve images given a textual description and vice-versa; or the visual encoder can be transferred in supervised fashion to a downstream task by fine-tuning on a labeled data set or by training a head on top of the frozen image encoder representation. In principle, the text encoder can be used as a standalone text embedding, but this application—to our knowledge—has not been explored in-depth, with some authors citing the low quality of the alt-texts leading to weak language modeling performance of the text encoder [67].

Previous works [46, 53] have shown that the image and text encoder can be realized with a single shared transformer model (henceforth referred to as single tower model, or 1T-CLIP), where the images are embedded using a patch embedding, and the tokenized text is embedded using a separate word embedding. Apart from the modality-specific embeddings, all model parameters are shared for the two modalities. While this type of sharing usually leads to a minor performance drop on image/image-language tasks it also halves the number of model parameters.

CLIPPO takes this idea one step further: text inputs are rendered on blank images, and are subsequently dealt with entirely as images, including the initial patch embedding (see Fig. 1 for an illustration). By training this single vision transformer contrastively as prior works, we obtain a single vision transformer model that can understand both images and text through the single interface of vision and provides a single representation which can be used to solve image, image-language, and pure language understanding tasks.

Alongside multimodal versatility, CLIPPO alleviates common hurdles with text processing, namely the development of an appropriate tokenizer and vocabulary. This is particularly interesting in a massively multilingual setup, where the text encoder has to handle dozens of languages.

We find that CLIPPO trained on image/alt-text pairs performs comparably with its 1T-CLIP counterpart on common image and image-language benchmarks, and is competitive with strong baseline language models on the GLUE benchmark [73]. However, due to the low quality of the alt-texts which are often not grammatical sentences, learning language understanding exclusively from alt-texts is fundamentally limited. Therefore, we augment image/alt-text contrastive pretraining with language-based contrastive training. Specifically, we use positive pairs of consecutive sentences sampled from a text corpus which is seamlessly integrated into the contrastive training by supplementing batches of image/alt-texts with (rendered) text/text pairs.

4. Experiments
4.1. Training details and models

We rely on a single training setup for all our baselines and visual text models. This setup was tuned to produce good results for standard image/alt-text contrastive training as in [56] (using exactly the same loss function as [56], following the pseudocode in [56, Fig. 3]) and we found that it readily transfers to 1T-CLIP and CLIPPO (including variants with text/text co-training).

Our default architecture is a ViT-B/16 [16] and we perform a subset of experiments with a ViT-L/16 architecture to study the effect of scale (we equip both models a MAP head [40] to pool embeddings). In all cases, the representation dimension used for the contrastive loss is 768. We set the batch size to 10,240 and train the main models for 250k steps, using a minimum 100k training steps for ablations. For models co-trained with a certain percentage of text/text data, we scale the number of iterations such that the number of image/alt-text pairs seen matches the number of iterations of the corresponding model without text/text data (e.g. when 50% of the data is text/text pairs we increase the number of iterations from 250k to 500k). The contrastive loss is computed across the full batch. We use the Adafactor optimizer [66] with a learning rate of $10^{-3}$ and decoupled weight decay with weight $10^{-3}$.

Baseline CLIP-style models are trained using the T5-en SentencePiece tokenizer [57]; we use the abbreviation CLIP* for the two tower model from [56] trained from scratch using the setup described above, to avoid confusion with the model released by [56]. A sequence length of 196 is used, as this matches the number of visual text "tokens" CLIPO can process with patch size 16 has at 224px resolution (which we use throughout unless noted otherwise).

**Visual text** For visual text rendering [60, 63] relied on the Google Noto font family\(^1\) which supports the majority of Unicode code points. Here, we use the GNU Unifont bitmap font\(^2\), which has a similar coverage but allows for efficient, lookup-based on-the-fly rendering in our preprocessing pipeline. We emphasize that this rendering strategy does not slow down training compared to tokenizer-based models. In preliminary explorations, we found this to be performance-neutral when compared to the Noto font.

\(^1\)https://fonts.google.com/noto
\(^2\)http://unifoundry.com/unifont
### 4.2. Evaluations and metrics

To evaluate the vision and vision/language understanding capabilities of our models we use standard metrics from the literature [53, 56, 85]: “zero-shot” transfer, which uses (embedded) textual description of the classes to be classified/retrieved and compares these with image embeddings.

<table>
<thead>
<tr>
<th>#param.</th>
<th>training dataset</th>
<th>I1k 10s.</th>
<th>I1k 0s.</th>
<th>C I→T</th>
<th>C T→I</th>
<th>F I→T</th>
<th>F T→I</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP*</td>
<td>203M</td>
<td>WebLI</td>
<td>55.8</td>
<td>65.1</td>
<td>48.5</td>
<td>31.3</td>
<td>79.2</td>
</tr>
<tr>
<td>1T-CLIP</td>
<td>118M</td>
<td>WebLI</td>
<td>53.9</td>
<td>62.3</td>
<td>48.0</td>
<td>30.3</td>
<td>77.5</td>
</tr>
<tr>
<td>CLIPPO</td>
<td>93M</td>
<td>WebLI</td>
<td>53.0</td>
<td>61.4</td>
<td>47.3</td>
<td>30.1</td>
<td>76.4</td>
</tr>
<tr>
<td>CLIPPO</td>
<td>93M</td>
<td>WebLI + 25%C4</td>
<td>52.1</td>
<td>57.4</td>
<td>40.7</td>
<td>26.7</td>
<td>68.9</td>
</tr>
<tr>
<td>CLIPPO</td>
<td>93M</td>
<td>WebLI + 50%C4</td>
<td>48.0</td>
<td>53.1</td>
<td>35.2</td>
<td>23.4</td>
<td>64.8</td>
</tr>
</tbody>
</table>

| 1T-CLIP L/16 | 349M | WebLI | 60.8   | 67.8  | 50.7  | 32.5  | 81.0  | 61.0  |
| CLIPPO L/16  | 316M | WebLI | 60.3   | 67.4  | 50.6  | 33.4  | 79.2  | 62.6  |
| CLIPPO L/16  | 316M | WebLI + 25%C4 | 60.5   | 66.0  | 44.5  | 29.8  | 72.9  | 57.3  |
| CLIPPO L/16  | 316M | WebLI + 50%C4 | 56.8   | 61.7  | 39.7  | 27.3  | 70.1  | 54.7  |

Table 1. Vision and vision-language cross-modal results. We report ImageNet-1k 10-shot linear transfer validation accuracy (I1k 10s.), ImageNet-1k zero-shot transfer validation accuracy (I1k 0s.), image-to-text and text-to-image retrieval recall@1 on MS-COCO (C I→T and C T→I) and on Flickr30k (F T→I and F I→T). CLIPPO and 1T-CLIP incur a minor drop in these evaluations compared to CLIP*, while only using about half of the model parameters. Co-training with text pairs from C4 (models with + xx%C4) degrades performance on some cross-modal tasks (but leads to improved language understanding capabilities, see Table 2).

Image/alt-text data  We use the WebLI data set introduced in [10] which comprises 10 billion images with 12 billion corresponding alt-texts. Importantly, WebLI comprises alt-texts in 109 languages (unlike previous data sets such as LAION-400M [64] which only contain English alt-texts) and it is therefore a great foundation to study multilingual language-image pretraining and its applications. Please refer to [10, Fig. 3] for details on the alt-text language distribution. For English-only models we obtain English versions of non-English alt-texts via GCP Translation API. In addition to alt-text, WebLI also provides OCR annotations, which we do not use in this paper. Finally, WebLI was processed with a de-duplication step removing all images from various splits of the image evaluation sets used in this paper. Please refer to [10, Sec. 3.2] for more details on the WebLI data set and to [10, Appendix B] for a datasheet.

We also present a subset of results based on LAION-400M [63] and YFCC-100M [71] as an additional comparison points, see Appendix C.1 and C.2, respectively.

Text/text data For co-training with text/text pairs we primarily rely on the publicly available Colossal Clean Crawled Corpus (C4; default/English split) [57]. We randomly sample pairs of consecutive sentences and contrastively train on these pairs, i.e., the model is trained for embedding-based next embedding prediction (NSP) [47]. We also experiment with pairs of parallel sentences in different languages from the WMT19 data set [18] as well as back-translated English sentences derived from C4 following the strategy described in [12].

We report the classification accuracy on ImageNet-1k [14] as well as the recall@1 for cross-modal retrieval on MS-COCO [9] and Flickr30k [81]. Furthermore, we test the low-data transfer performance of the models by means of the linear adaptation protocol from [16], reporting the 10-shot accuracy on ImageNet-1k.

We also evaluate CLIPPO and baselines on the popular VQA benchmark VQAv2 [25]. To construct a VQA model using a single pretrained ViT we render the question at the top end of the corresponding image (using the same Unifont renderer as used for CLIPPO training) and follow the standard prediction setup where the answer is predicted as the most likely answer from the training set, i.e. by classification. Specifically, we replace the last layer of our pretrained CLIPPO and baselines with a randomly initialized one with the appropriate number of outputs and fine-tune on VQAv2. This setup tests the ability of the pretrained ViT to combine image and text in intermediate layers as it has produce a single output from a fused image/text input image, unlike in the other cross-modal tasks (and pretraining), where image and text representations are computed with two separate forward passes. Please refer to Appendix A in the supplementary material for examples images with rendered questions and Appendix B.1 for details on the fine-tuning protocol.

Multilingual capabilities are assessed via zero-shot retrieval on CrossModal3600 [70], which is a geographically diverse set comprising 3600 images each human-annotated with captions in 36 languages. The corresponding recall metric is averaged across all languages and images.

Finally, we evaluate the language understanding capabilities on the General Language Understanding Evaluation (GLUE) benchmark [73] which comprises natural language inference tasks (MNLI, QNLI, RTE), a sentiment analysis task (SST-2), sentence similarity tasks (QQP, STS-B, MRPC), and a linguistic acceptability task (CoLA). Following common practice, we exclude the WNLI task from the benchmark [15, 77]. We transfer our baselines and CLIPPO models by attaching a 2-hidden layer MLP with 768 units to
their representation and following precisely the fine-tuning protocol from BERT [15]. For sentence pair classification tasks we simply render both sentences on the same image, printing [SEP] to mark the start of the second sentence.

4.3. Vision and vision-language understanding

Image classification and retrieval Table 1 shows the performance of CLIPPO along with the baseline models on the benchmarks described in Sec. 4.2. It can be seen that the CLIPPO and 1T-CLIP incur a drop of a 2-3 percentage points absolute compared to CLIP*. This is not surprising and can be attributed to the fact that single tower models only have about half the parameters count of a corresponding two tower model. The difference in performance between the English-only CLIPPO and 1T-CLIP is very small for a B/16 backbone at 100k training steps (see Table 6 in the supplementary material), and vanishes with longer training and/or by increasing the model size, despite the fact that CLIPPO has 25% and 10% fewer parameters than 1T-CLIP for a B/16 and L/16 architecture, respectively (which is due to the absence of the text embedding in CLIPPO).

The multilingual CLIPPO model performs somewhat worse than the corresponding 1T-CLIP, and the gap does not close completely when training longer (see Table 6).

However, when evaluated across a broad set of languages on the CrossModal3600 CLIPPO performs on par with or slightly better than 1T-CLIP (see Sec. 4.4 below).

As we add sentence pairs to the training mix the performance on the cross-modal retrieval metrics decreases. This is not surprising as we keep the total batch size constant so that the effective batch size of image/alt-text contrastive training decreases, which is known to impact performance [85]. Interestingly, the 10-shot transfer performance does not move in tandem, but only decreases significantly when half of the training data is sentence pairs. In exchange, co-training with text data leads to significantly improved language understanding performance (see Sec. 4.5).

VQA In Fig. 2 we report the VQAv2 score of our models and baselines. It can be seen that CLIPPO outperforms CLIP*, 1T-CLIP, as well as a pretrained ViT-B/16 from [16] by a significant margin, achieving a score of 66.3, and co-training with 25% C4 data leads to a slight improvement of the score. The improved score of CLIPPO is mainly due to better performance in “yes/no” questions. Increasing the model size to L/16 adds another 2 points which originate from improvements in the “number” and “other” VQAv2 categories. However, note that for an L/16 architecture 1T-CLIP performs competitively with CLIPPO (see Table 7).

One possible explanation for this could be that 1T-CLIP develops better OCR capabilities thanks to the higher model capacity (alt-texts can correlate with text in images/scene text, see [10, Fig. 3]). Increasing the resolution to 384px adds 2 to 3 points across models.

We also compare CLIPPO with baselines from the literature. Specifically, [17] proposes framework (called METER) for multimodal tasks, where pretrained transformer-based image and text encoders are combined with a transformer-based fusion module. CLIPPO L/16 achieves performance competitive with their model combining a CLIP B/32 vision backbone with a BERT-Base language backbone, which is roughly comparable in size and computational cost with our L/16 models. Another related work is [67], which combines different CLIP vision backbones with two existing VQA systems, Pythia [33] and MCAN [84]. CLIPPO outperforms different CLIP ViT-based Pythia and MCAN models from [67]. Note, however, that ResNet-based CLIP backbones lead to better results when combined with these systems. We further note that both [17] and [67] also investigate training their models on a mix of different image-text data sets with multiple objectives such as grounded masked language modeling and text-image matching, before transferring to the VQA task, which leads to significant improvements. ViLT [36] relies on such a strategy to train a single transformer backbone jointly encoding image and text tokens. At 384px resolution, CLIPPO (with 25% C4 data) obtains a VQA score comparable with that of ViLT (and other models from the
literature such as ViLBERT [48], VisualBERT [43], and PixelBERT [30]), despite only using a contrastive objective for pretraining.

### 4.4. Multilingual vision-language understanding

For typical language models, tokenizer choice can be a challenging process [78]. Commonly used English-language tokenizers generalize poorly to non-latin scripts [85]. This can be alleviated by the use of larger, multilingual vocabularies, at the expense of very large parameter counts. CLIPPO bypasses this issue, removing any language-related bias stemming from unbalanced or restrictive tokenizers. We consider multilingual image/text retrieval on Crossmodal3600 and compare CLIPPO, trained on WebLI with multilingual alt-texts, against 1T-CLIP with a number of SentencePiece tokenizers; one trained from 300M WebLI multilingual alt-texts, English (T5-en) and multilingual (T5-all) tokenizers from T5 [57], and a multilingual tokenizer (mT5) from mT5 [79], all with a vocabulary size of 32,000. The results are shown in Fig. 4. On average, CLIPPO achieves comparable retrieval performance to these baselines. In the case of mT5, the use of extra data to create the specialized vocabulary can boosts performance above that of CLIPPO; the leveraging of such extra parameters and data in the multilingual context will be an interesting future direction for CLIPPO.

#### Tokenization efficiency

If a tokenizer is well suited to a particular dataset, it will tokenize to shorter sequences—this is especially the case when byte fallback [39] is enabled. SentencePiece tokenizers have the advantageous ability to tokenize entire—possibly quite long—words to single tokens. CLIPPO cannot learn any such compression, but benefits from equal treatment of all languages and words: it will by definition generalize equally well to all data, as its tokenization schema has not been trained on a specific dataset. We analyze 20,000 samples for each of the 104 C4 languages. Each CLIPPO token is assumed to be a 16 × 16 patch; though in typical computations all approaches considered here would pad to a fixed length, we compute CLIPPO’s sequence length according to the last patch which contains rendered text. Fig. 3 shows the fraction of C4 languages where CLIPPO processes tokens more efficiently than the vocabularies discussed above. We conservatively define “more efficient” as producing a shorter token sequence for over 75% of examples. Even so, CLIPPO is indeed more efficient across the majority of languages. Per-language breakdowns of multilingual retrieval performance and tokenization efficiency are further discussed in Appendix C.3.

### 4.5. Language understanding

Table 2 shows the GLUE benchmark results of CLIPPO and baselines. One can observe that CLIPPO trained on WebLI performs competitively with the BiLSTM+Attn+ELMo baseline which relies on deep word embeddings trained on a large language corpus. Also, it can be seen that CLIPPO along with 1T-CLIP outperform the language encoder trained using standard contrastive language vision pretraining (CLIP†). This indicates that multimodal training in a single encoder benefits language understanding. Furthermore, CLIPPO achieves a much higher GLUE score than the CLIP† image encoder, which in turn leads to significantly better results than fine-tuning a ViT-B/16 from scratch on GLUE (see Appendix C.2 for additional results). Unsurprisingly, the models pretrained on WebLI cannot do better than random guessing on the CoLA evaluation which requires to assess the grammatical correctness of sentences (recall that alt-texts are rarely grammatical sentences). Also the accuracy of CLIP† and 1T-CLIP vision encoders we observe for SST-2 is in agreement with what was reported in [56, Table 10] for CLIP with a ViT-B/16 image encoder.

Adding sentence pairs form the C4 corpus gradually improves the GLUE score, and when half of the examples are sentence pairs our model becomes competitive with PIXEL, while still retaining decent image and vision-language understanding capabilities (cf. Table 1). Note, however, that there is a trade-off between language-only tasks and tasks that involve image understanding. Finally, training CLIPPO only on sentence pairs leads to a model which outperforms PIXEL by a significant margin. However, our model has seen more sentence pairs than PIXEL, so PIXEL might improve as well when training longer.
Table 2. Results for the GLUE benchmark (dev set). The metric is accuracy except for the performance on QQP and MRPC, which is measured using the F₁ score, CoLA which uses Matthew’s correlation, and STS-B which evaluated based on Spearman’s correlation coefficient. “avg” corresponds to the average across all metrics. The results for BERT-Base and PIXEL are from [60, Table 3], and BiLSTM and BiLSTM+Attn, ELMo from [73, Table 6]. All encoders considered here have a transformer architecture comparable to BERT-Base (up to the text embedding layer), except for CLIPPO L/16 which uses a ViT L/16, and the two BiLSTM model variants. Wiki and BC stand for (English) Wikipedia and Bookcorpus [86] data, respectively.

4.6. Ablations and analysis

**Impact of weight sharing across modalities** The fact that CLIPPO 1) uses a shared patch embedding for regular images and text images and 2) this embedding has considerably fewer parameters than the text embedding of 1T-CLIP and CLIP* provokes the question of whether CLIPPO could benefit from separate patch embeddings for text images and regular images. Further, CLIPPO relies on a single head to compute the output representation for images and text, and relaxing this constraint by using separate heads for the two modalities could lead to more expressive representations. The results (deferred to Appendix D.1) show that neither of these variants lead to improved image classification or retrieval metrics compared to CLIPPO.

**Impact of the text location** We test whether rendering the question at the top, middle, or bottom of the image impacts the VQA performance of CLIPPO and find that it does not, provided that we increase the learning rate of the positional embedding during fine-tuning (see Appendix D.2).

**Typographic attacks** Since CLIPPO is trained on large amounts of rendered (alt-)text it is important to check whether it becomes more susceptible to typographic attacks—the tendency of CLIP-style models to zero-shot classify an image according to adversarially injected scene text unrelated to the scene [23, 42, 50]. In Appendix D.3 we present results indicating that CLIPPO is no more vulnerable to typographic attacks than 1T-CLIP and CLIP*.

**Modality gap** Liang et al. [45] discovered that text and image embeddings of CLIP-style models form two distinct clusters rather than both filling the embedding space densely and occupying the same spatial region. They attribute this phenomenon to a combination of initialization conditions and properties of the loss function/training dynamics. Since we consider single tower models here, and also co-train some of these models with text-only pairs it is interesting to see how this affects the modality gap. We compute the gap and visualize it following the recipe from [45] in Fig. 5 (see Appendices D.4 and D.5 for additional visualizations). CLIPPO attains a slightly lower modality gap than CLIP*, but clearly features a clustering structure for image and text embeddings. However, when training contrastively with sentence pairs in addition to image/alt-text pairs, the clustering structure disappears, the image and text embeddings overlap, and the modality gap decreases significantly. A possible explanation for this behavior could be that the additional learning pressure induced by the contrastive loss on sentence pairs encourages text embeddings to spread out more and hence the structure of all embeddings changes.

**Text/text co-training objectives** To corroborate that contrastive NSP is a sensible objective to improve language understanding in the context of CLIPPO, we train CLIPPO without any image/alt-text data on pairs of parallel translated sentences (this is straight-forward in our framework since visual text is language-agnostic), as well as English back-translated data, and evaluate the resulting text representations on GLUE. Table 3 shows that NSP on C4 clearly achieves the highest GLUE score.

5. Discussion and limitations

We proposed and evaluated CLIPPO which produces a single ViT that can understand images and language jointly using images as a sole input modality. Perhaps surprisingly, CLIPPO matches the performance of the 1T-CLIP baseline across many of the considered tasks, and only in-
curs a minor drop compared to the CLIP* baseline, despite having less than half the parameters of CLIP*. As we showed, the image-only interface enables a simple, unified data pipeline for training on and transferring to mixed modalities. CLIPPO opens the door for additional modalities (e.g., spectrograms) and, as we hope, might inspire applications of pixel-only models beyond contrastive training. Nevertheless, several limitations remain, as discussed next.

**Co-training** First, to achieve language understanding performance competitive with PIXEL and BERT on GLUE, contrastive co-training with text pairs is necessary. While adding 25% C4 data to the batch seems to strike a good balance across all tasks considered, it does induce a non-negligible drop in zero-shot image classification and image/text retrieval. This drop becomes more severe as the fraction of C4 examples increases. We observed an associated change in modality gap, and further investigation of the representation in the context of co-training might help to develop models that achieve better overall performance in the co-training setup.

**Diverse rendered text** CLIPPO currently relies on cleanly rendered text as an input and its capabilities to handle text from documents or web pages without further adaption is limited (besides the basic OCR capabilities that CLIP-style models learn from image/alt-text pairs). We emphasize that sophisticated OCR and document understanding is not a goal of this paper, and training CLIPPO with augmented noisy rendered text that mimics the distribution of documents and websites is likely to lead to worse performance across the considered tasks, since image/alt-text pairs are less correlated and provide a weaker learning signal. However, developing CLIPPO further to handle less clean visual text will open many additional applications.

**Generative modeling** CLIPPO, like CLIP, BERT, PIXEL, and many other models, uses an encoder-only design and hence lacks the ability to generate text outputs. A common approach to equip encoder-only models with generation capabilities (e.g., for image captioning or VQA) is to simply combine them with a (potentially pretrained) language model [8, 76]. This approach naturally also applies to CLIPPO and PIXEL, but defeats the advantages of visual text in certain (e.g. multilingual) scenarios. While visual text outputs have previously been explored in the context of machine translation [49], it remains unclear what a scalable tokenizer-free way to generate text is.

**Multilingual learning** Finally, we showed that CLIPPO obtains strong multilingual image/text retrieval performance without requiring the development of an appropriate tokenizer. For fine-grained adjustment and balancing of the retrieval performance further steps will be necessary, including data balancing and potentially co-training with multi-lingual text data. Furthermore, similar to PIXEL, CLIPPO relies on certain ad-hoc design choices w.r.t. the visual representation, for example the left-to-right rendering of Arabic scripts. This approach leads to decent performance on average, but it is not clear what kind of unwanted effects it introduces and how these could be mitigated.

### 6. Conclusion

We introduced CLIPPO, a model for processing images and text jointly through the lens of vision. This reduces design choices (in particular w.r.t. tokenization) and parameters count, simplifies data processing pipelines and transfer recipes, and increases generality across multiple languages. We also explored methods of enhancing language understanding, where traditional image/alt-text contrastive models trained on web data fall short. We demonstrated this is possible by co-training with text pairs, with CLIPPO models outperforming strong NLP baselines while maintaining solid image understanding capabilities.

Although we presented a unified contrastive training algorithm, CLIPPO suffers somewhat when co-training on multiple tasks, and future work to harmonize the co-training could enhance the models significantly. Deeper understanding of the design choices in rendering text as images, and their impact on performance, is another interesting avenue.

### Acknowledgments
We would like to thank Lucas Beyer, Josip Djolonga, Alexander Kolesnikov, Mario Lucic, Andreas Steiner, Xiao Wang, and Xiaohua Zhai for inspiring discussions and helpful feedback. We also thank Jeffrey Sorensen for help with the text rendering preprocessing.

<table>
<thead>
<tr>
<th></th>
<th>WMT19</th>
<th>WMT19 BT</th>
<th>C4 NSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLUE score</td>
<td>61.2</td>
<td>66.6</td>
<td>77.6</td>
</tr>
</tbody>
</table>

Table 3. Ablation of text pair-based contrastive co-training tasks: Training on parallel translated sentences (WMT19), training on parallel back-translated sentences (WMT19 BT), and NSP for sentences sampled from C4 (C4 NSP). C4 NSP leads to the highest GLUE score by a large margin.
References


[26] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In CVPR, pages 15979–15988, 2022.


[32] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In ICLR, 2021.


[59] Nils Rethmeier and Isabelle Augenstein. A primer on contrastive pretraining in language processing: Methods, lessons learned and perspectives. ACM Computing Surveys, 2021. 2


[65] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In ACL, 2016. 2


[71] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: the new data in multimedia research. Commun. ACM, 59(2):64–73, 2016. 4, 14


