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.

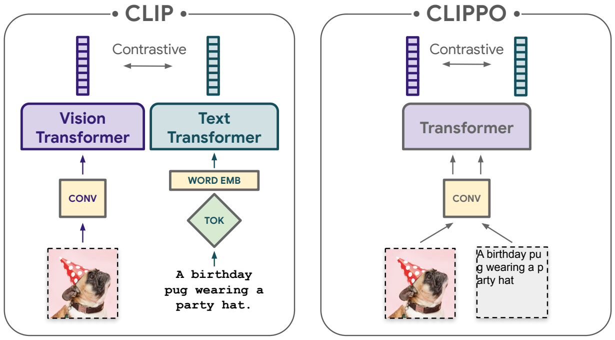

Figure 1. CLIP [56] trains separate image and text encoders, each with a modality-specific preprocessing and embedding, on image/alt-text pairs with a contrastive objective. CLIPPO trains a pure pixel-based model with equivalent capabilities by rendering the alt-text as an image, encoding the resulting image pair using a shared vision encoder (in two separate forward passes), and applying same training objective as CLIP.

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

Code and pretrained models are available as part of big_vision [4] https://github.com/google-research/big_vision.

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 predetermining 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 [67] 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 vision 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 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^{-4}.

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" CLIPPO 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

	#param.	training dataset	I1k 10s.	I1k 0s.	C I→T	C T→I	F I→T	F T→I
CLIP*	203M	WebLI	55.8	65.1	48.5	31.3	79.2	59.4
1T-CLIP	118M	WebLI	53.9	62.3	48.0	30.3	77.5	58.2
CLIPPO	93M	WebLI	53.0	61.4	47.3	30.1	76.4	57.3
CLIPPO	93M	WebLI + 25%C4	52.1	57.4	40.7	26.7	68.9	51.8
CLIPPO	93M	WebLI + 50%C4	48.0	53.1	35.2	23.4	64.8	47.2
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[3]. 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 [64] 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 sentence 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].

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. 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 correspoding 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

[3] https://cloud.google.com/translate

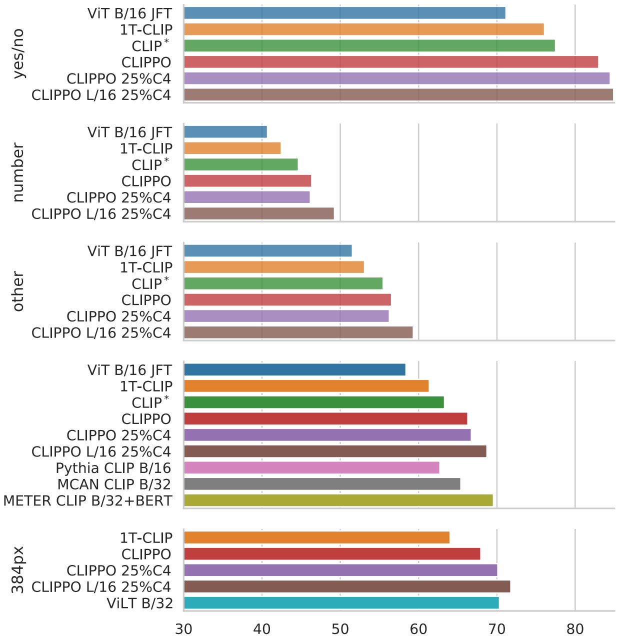

Figure 2. Results on the VQAv2 benchmark (test-dev set). In addition to CLIPPO and baselines produced in this work, we also compare to Pythia and MCAN models with ViT encoders from [67], and with comparably sized METER [17] and ViLT [36] models. CLIPPO outperforms CLIP* and 1T-CLIP clearly on "yes/no" questions and gets similar performance as task-specific models.

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 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 manly 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.

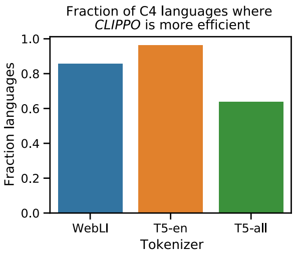

Figure 3. Tokenization efficiency analyzed in terms of the sequence length produced by a given method. CLIPPO produces smaller sequences for the majority of languages compared to 1T-CLIP with alternative tokenizers.

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.

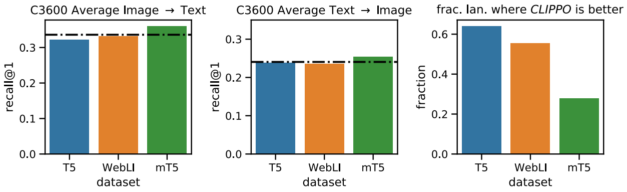

Figure 4. Zero-shot image/text retrieval performance on Cross-Modal3600 [70]. Although specialized (mc4) tokenizers can be leveraged to improve multilingual performance CLIPPO (dashed black line) broadly matches or exceeds comparable 1T-CLIP models trained with vocabulary size 32,000 (the word embeddings result in a 27% increase in parameter count compared to CLIPPO).

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.

	training dataset	MNLI-M/MM	QQP	QNLI	SST-2	COLA	STS-B	MRPC	RTE	avg
BERT-Base	Wiki + BC	84.0 / 84.2	87.6	91.0	92.6	60.3	88.8	90.2	69.5	83.1
PIXEL	Wiki + BC	78.1 / 78.9	84.5	87.8	89.6	38.4	81.1	88.2	60.5	76.3
BiLSTM		66.7 / 66.7	82.0	77.0	87.5	17.6	72.0	85.1	58.5	68.1
BiLSTM+Attn, ELMo		72.4 / 72.4	83.6	75.2	91.5	44.1	56.1	82.1	52.7	70.0
CLIP* img enc.	WebLI	66.4 / 67.5	78.6	69.4	78.6	0.0	5.2	81.2	52.7	55.5
CLIP* text enc.	WebLI	71.8 / 72.5	82.7	73.0	86.2	6.6	65.0	81.4	53.8	65.9
1T-CLIP text enc.	WebLI	72.6 / 73.0	83.8	80.7	84.9	0.0	79.6	83.3	57.0	68.3
CLIPPO	WebLI	73.0 / 72.6	84.3	81.2	86.8	1.8	80.5	84.1	53.4	68.6
CLIPPO	WebLI + 25%C4	77.7 / 77.2	85.3	83.1	90.9	28.2	83.4	84.5	59.2	74.4
CLIPPO	WebLI + 50%C4	79.2 / 79.2	86.4	84.2	92.9	38.9	83.4	84.8	59.9	76.6
CLIPPO	C4	79.9 / 80.2	86.7	85.2	93.3	50.9	84.7	86.3	58.5	78.4
CLIPPO L/16	WebLI + 25%C4	76.6 / 75.5	87.1	79.9	93.2	48.2	84.1	84.6	56.0	76.1
CLIPPO L/16	WebLI + 50%C4	82.3 / 82.4	87.9	86.7	94.2	55.3	85.8	85.9	59.2	80.0

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 F_1 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-

	WMT19	WMT19 BT	C4 NSP
GLUE score	61.2	66.6	77.6

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.

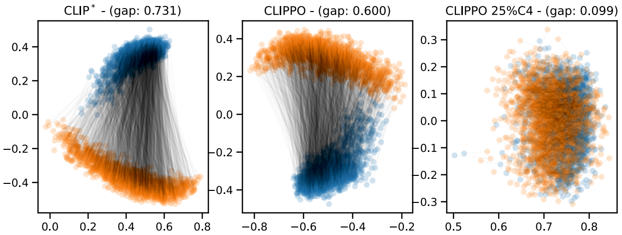

Figure 5. Visualization of the modality gap for CLIP* and CLIPPO optionally trained with 25% C4 data. The visualization follows the analysis from [45] and shows embedded images (blue dots) and corresponding alt-text (orange dots) from the WebLI validation set, projected to the first two principal components of the validation data matrix. CLIPPO has a slightly smaller modality gap than CLIP*; co-training with C4 data strongly reduces the gap.

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 parameter 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.

References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. VATT: transformers for multimodal self-supervised learning from raw video, audio and text. In *NeurIPS*, 2021. 2

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: A visual language model for few-shot learning. In *NeurIPS*, 2022. 1, 2

[3] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. DocFormer: End-to-end transformer for document understanding. In *ICCV*, pages 973–983, 2021. 2

[4] Lucas Beyer, Xiaohua Zhai, and Alexander Kolesnikov. Big Vision. `https://github.com/google-research/big_vision`, 2022. 1

[5] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matthew Sharifi, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. AudioLM: a language modeling approach to audio generation. *CoRR*, abs/2209.03143, 2022. 1, 2

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 1

[7] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *NeurIPS*, pages 15084–15097, 2021. 1

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 8

[9] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. 4

[10] Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A jointly-scaled multilingual language-image model. In *ICLR*, 2023. 1, 2, 4, 5

[11] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling language modeling with pathways. *CoRR*, abs/2204.02311, 2022. 1

[12] Geoffrey Cideron, Sertan Girgin, Anton Raichuk, Olivier Pietquin, Olivier Bachem, and Léonard Hussenot. vec2text with round-trip translations. *CoRR*, abs/2209.06792, 2022. 4

[13] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *ACL*, pages 8440–8451, 2020. 2

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 4

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 1, 2, 4, 5

[16] 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. An image is worth 16×16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 2, 3, 4, 5, 14, 16, 17, 24

[17] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. In *CVPR*, pages 18145–18155, 2022. 5, 14, 17

[18] Wikimedia Foundation. ACL 2019 Fourth Conference on Machine Translation (WMT19), Shared Task: Machine Translation of News. `http://www.statmt.org/wmt19/translation-task.html`. 4

[19] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP*, pages 6894–6910, 2021. 2

[20] John M. Giorgi, Osvald Nitski, Bo Wang, and Gary D. Bader. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *ACL/IJCNLP*, pages 879–895, 2021. 2

[21] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. OmniMAE: Single model masked pretraining on images and videos. *CoRR*, abs/2206.08356, 2022. 1, 2

[22] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *CVPR*, pages 16081–16091, 2022. 1

[23] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. 7, 21

[24] Yuan Gong, Yu-An Chung, and James R. Glass. AST: audio spectrogram transformer. In *Interspeech*, pages 571–575, 2021. 1

[25] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 4, 13

[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. 2

[27] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *CoRR*, abs/2210.02303, 2022. 2

[28] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. *CoRR*, abs/2207.06405, 2022. 1

[29] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document AI with unified text and image masking. In *ACMMM*, pages 4083–4091, 2022. 2

[30] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers. *CoRR*, abs/2004.00849, 2020. 6

[31] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J. Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *ICLR*, 2022. 1

[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 *ICML*, 2021. 2

[33] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0.1: The winning entry to the VQA challenge 2018. *CoRR*, abs/1807.09956, 2018. 5

[34] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. 1

[35] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. OCR-Free document understanding transformer. In *ECCV*, pages 498–517, 2022. 2

[36] Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-language transformer without convolution or region supervision. In *ICML*, pages 5583–5594. PMLR, 2021. 5, 17

[37] Alexander Kolesnikov, André Susano Pinto, Lucas Beyer, Xiaohua Zhai, Jeremiah Harmsen, and Neil Houlsby. UViM: A unified modeling approach for vision with learned guiding codes. In *NeurIPS*, 2022. 1

[38] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *ICML*, pages 3519–3529, 2019. 22

[39] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP*, 2018. 2, 6

[40] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *ICML*, pages 3744–3753, 2019. 3, 14

[41] Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. *CoRR*, abs/2210.03347, 2022. 2

[42] Yoann Lemesle, Masataka Sawayama, Guillermo Valle-Perez, Maxime Adolphe, Hélène Sauzéon, and Pierre-Yves Oudeyer. Language-biased image classification: evaluation based on semantic representations. In *ICLR*, 2022. 7, 21

[43] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557, 2019. 6

[44] Qing Li, Boqing Gong, Yin Cui, Dan Kondratyuk, Xianzhi Du, Ming-Hsuan Yang, and Matthew Brown. Towards a unified foundation model: Jointly pre-training transformers on unpaired images and text. *CoRR*, abs/2112.07074, 2021. 2

[45] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 2022. 7, 8, 22, 23

[46] Valerii Likhosherstov, Anurag Arnab, Krzysztof Choromanski, Mario Lucic, Yi Tay, Adrian Weller, and Mostafa Dehghani. Polyvit: Co-training vision transformers on images, videos and audio. *Trans. Machine Learning Research*, 2023. 1, 2, 3

[47] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. In *ICLR*, 2018. 2, 4

[48] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pages 13–23, 2019. 6

[49] Elman Mansimov, Mitchell Stern, Mia Xu Chen, Orhan Firat, Jakob Uszkoreit, and Puneet Jain. Towards end-to-end in-image neural machine translation. *CoRR*, abs/2010.10648, 2020. 1, 2, 8

[50] Joanna Materzyńska, Antonio Torralba, and David Bau. Disentangling visual and written concepts in CLIP. In *CVPR*, pages 16410–16419, 2022. 7, 21

[51] Joanna Materzynska, Antonio Torralba, and David Bau. Disentangling visual and written concepts in clip. In *CVPR*, 2022. 21, 22

[52] Sabrina J. Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, and Samson Tan. Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. *CoRR*, abs/2112.10508, 2021. 2

[53] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with LIMoE: the language-image mixture of experts. In *NeurIPS*, 2022. 2, 3, 4

[54] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. Masked spectrogram modeling using masked autoencoders for learning general-purpose audio representation. *CoRR*, abs/2204.12260, 2022. 1

[55] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, Mingxing Tan, and Quoc V. Le. Combined scaling for open-vocabulary image classification. *CoRR*, abs/2111.10050, 2021. 2

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3, 4, 6, 14, 17

[57] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. 3, 4, 6, 20

[58] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831, 2021. 2

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

[60] Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. Language modelling with pixels. In *ICLR*, 2023. 1, 2, 3, 7, 18

[61] Phillip Rust, Jonas Pfeiffer, Ivan Vulic, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In *ACL/IJCNLP*, pages 3118–3135, 2021. 2

[62] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 2

[63] Elizabeth Salesky, David Etter, and Matt Post. Robust open-vocabulary translation from visual text representations. In *EMNLP*, pages 7235–7252, 2021. 2, 3

[64] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. *CoRR*, abs/2111.02114, 2021. 4, 15

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

[66] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *ICML*, 2018. 3, 14

[67] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can CLIP benefit vision-and-language tasks? In *ICLR*, 2022. 2, 3, 5, 17

[68] Baohua Sun, Lin Yang, Catherine Chi, Wenhan Zhang, and Michael Lin. Squared english word: A method of generating glyph to use super characters for sentiment analysis. In *Workshop on Affective Content Analysis*, volume 2328, pages 140–151, 2019. 2

[69] Zineng Tang, Jaemin Cho, Yixin Nie, and Mohit Bansal. TVLT: Textless vision-language transformer. In *NeurIPS*, 2022. 1

[70] Ashish Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset. In *EMNLP*, 2022. 4, 6, 19

[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

[72] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. In *ICLR*, 2023. 2

[73] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*, 2019. 2, 3, 4, 7, 18

[74] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. GIT: A generative image-to-text transformer for vision and language. *Trans. Machine Learning Research*, 2022. 1, 2

[75] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: BEiT pretraining for all vision and vision-language tasks. *CoRR*, abs/2208.10442, 2022. 1

[76] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple visual language model pretraining with weak supervision. In *ICLR*, 2022. 1, 8

[77] Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. CLEAR: contrastive learning for sentence representation. *CoRR*, abs/2012.15466, 2020. 2, 4

[78] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Trans. Assoc. Comput. Linguistics*, 10:291–306, 2022. 6

[79] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *NAACL-HLT*, 2021. 6, 20

[80] Haoxuan You, Luowei Zhou, Bin Xiao, Noel Codella, Yu Cheng, Ruochen Xu, Shih-Fu Chang, and Lu Yuan. Learning visual representation from modality-shared contrastive language-image pre-training. In *ECCV*, pages 69–87, 2022. 2

[81] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78, 2014. 4

[82] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Trans. Machine Learning Research*, 2022. 2

[83] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Trans. Machine Learning Research*, 2022. 2

[84] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, pages 6281–6290, 2019. 5

[85] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-shot transfer with locked-image text tuning. In *CVPR*, pages 18102–18112, 2022. 2, 4, 5, 6, 14

[86] Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*, pages 19–27, 2015. 7, 18