FLAVA: A Foundational Language And Vision Alignment Model

Amanpreet Singh*  Ronghang Hu*  Vedanuj Goswami*
Guillaume Couairon  Wojciech Galuba  Marcus Rohrbach  Douwe Kiela
Facebook AI Research (FAIR)

Abstract

State-of-the-art vision and vision-and-language models rely on large-scale visio-linguistic pretraining for obtaining good performance on a variety of downstream tasks. Generally, such models are often either cross-modal (contrastive) or multi-modal (with earlier fusion) but not both; and they often only target specific modalities or tasks. A promising direction would be to use a single holistic universal model, as a “foundation”, that targets all modalities at once—a true vision and language foundation model should be good at vision tasks, language tasks, and cross- and multi-modal vision and language tasks. We introduce FLAVA as such a model and demonstrate impressive performance on a wide range of 35 tasks spanning these target modalities.

1. Introduction

Large-scale pre-training of vision and language transformers has led to impressive performance gains in a wide variety of downstream tasks. In particular, contrastive methods such as CLIP [82] and ALIGN [50] have shown that natural language supervision can lead to very high quality visual models for transfer learning.

Purely contrastive methods, however, also have important shortcomings. Their cross-modal nature does not make them easily usable on multimodal problems that require dealing with both modalities at the same time. They require large corpora, which for both CLIP and ALIGN have not been made accessible to the research community and the details of which remain shrouded in mystery, notwithstanding well-known issues with the construction of such datasets [9].

In contrast, the recent literature is rich with transformer models that explicitly target the multimodal vision-and-language domain by having earlier fusion and shared self-attention across modalities. For those cases, however, the unimodal vision-only or language-only performance of the model is often either glossed over or ignored completely.

If the future of our field lies in generalized “foundation models” [10] or “universal” transformers [72] with many different capabilities, then the following limitation should be overcome: a true foundation model in the vision and language space should not only be good at vision, or language, or vision-and-language problems—it should be good at all three, at the same time.

Combining information from different modalities into one universal architecture holds promise not only because it is similar to how humans make sense of the world, but also because it may lead to better sample efficiency and much richer representations.

In this work, we introduce FLAVA, a foundational language and vision alignment model that explicitly targets vision, language, and their multimodal combination all at once. FLAVA learns strong representations through joint pretraining on both unimodal and multimodal data while encompassing cross-modal “alignment” objectives and multimodal “fusion” objectives. We validate FLAVA by applying it to 35 tasks across vision, NLP, and multimodal domains and show impressive performance. An important advantage of our approach is that it was trained on a corpus of openly available datasets that is an order of magnitude smaller than datasets used in comparable models. Our models and code are available in https://flava-model.github.io/.
2. Background

The self-supervised pretraining paradigm has significantly advanced the state of the art across various domains, from natural language processing [6, 17–19, 23, 24, 28, 30, 61, 68, 73, 82–84], to computer vision [2, 5, 8, 12, 31, 33, 37, 59, 75, 102, 114], to speech recognition [4, 22, 42, 67, 116] and multimodal domains such as vision and language understanding [12, 16, 34, 43–45, 50, 62–65, 70, 71, 93, 99, 100, 107, 113, 115, 117]. Even though this progress has been based on a shared recipe of self-supervised learning on top of transformers, we are still missing major progress in building foundational models [10] that work well across all of these different domains and modalities at once.

Table 1 shows an extensive comparison of popular and recent models w.r.t. our FLAVA on multiple axes. Recent work either (i) focuses on a single target domain [54, 115]; (ii) targets a specific unimodal domain along with the joint vision-and-language domain [50, 82]; or (iii) targets all domains but only a specific set of tasks in a particular domain.

SimVLM [107], ALIGN [50], and CLIP [82] have demonstrated impressive gains by training transformer-based models on giant private paired image-and-text corpora, as opposed to the previous vision-and-language state-of-the-art such as VinVL [115] and ViLT [54], which were trained on smaller public paired datasets [15, 57, 66, 77, 90].

Generally, models in the vision-and-language space can be divided into two categories: (i) dual encoders where the image and text are encoded separately followed by a shallow interaction layer for downstream tasks [50, 82]; and (ii) fusion encoder(s) with self-attention spanning across the modalities [16, 34, 44, 45, 62–65, 70, 71, 99, 100, 107, 115, 117]. The dual encoder approach works well for unimodal [105, 106] and cross-modal retrieval tasks [66, 80] but their lack of fusion usually causes them to underperform on tasks that involve visual reasoning and question answering [39, 53, 91, 94] which is where models based on fusion encoder(s) shine.

Within the fusion encoder category, a further distinction can be made as to whether the model uses a single transformer for early and unconstrained fusion between modalities (e.g., VisualBERT, UNITER, VLBERT, OSCAR [16, 63, 65, 99, 117]) or allows cross-attention only in specific co-attention transformer layers while having some modality-specific layers (e.g., LXMERT, ViLBERT, ERNIE-VIL [70, 71, 100, 113]. Another distinguishing factor between different models lies in the image features that are used, ranging from region features [63, 70, 115], to patch embeddings [54, 62, 107], to convolution or grid features [46, 51].

Dual encoder models use contrastive pretraining to predict the correct N paired combinations among N² possibilities. On the other hand, with fusion encoders, inspired by unimodal pretraining schemes such as masked language modeling [28, 68], masked image modeling [5], and causal language modeling [83], numerous pretraining tasks have been explored: (i) Masked Language Modeling (MLM) for V&L where masked words in the caption are predicted with help of the paired image [63, 70, 100]; (ii) prefixLM, where with the help of an image, the model tries to complete a caption [26, 107]; (iii) image-text matching, where the model predicts whether given pair of image and text match or not; and (iv) masked region modeling, where the model regresses onto the image features or predicts its object class.

Compared to previous work, our model FLAVA works on a wide range of tasks in each of the vision, language, and vision-and-language domains. FLAVA uses a shared trunk which was pretrained on only openly available public paired data. FLAVA combines dual and fusion encoder approaches into one holistic model that can be pretrained with our novel FLAVA pretraining scheme that leverages pretraining objectives from both categories. FLAVA is designed to be able to take advantage of unpaired unimodal data along with multimodal paired data, resulting in a model that can handle unimodal and retrieval tasks as well as cross-modal and multimodal vision-and-language tasks.

3. FLAVA: A Foundational Language And Vision Alignment Model

The goal of this work is to learn a foundational language and vision representation that enables unimodal vision and language understanding as well as multimodal reasoning, all within a single pre-trained model. We show how this can be achieved with a simple and elegant architecture
For while contrastive, masked multimodal modeling (MMM), and image-text matching (ITM) loss are used over paired image-text data. Mask language modeling (MLM) losses are applied onto the image and text encoders over a single image or a text piece, respectively, from prior work, our text encoder has exactly the same architecture as the visual encoder, *i.e.*, we use the same ViT architecture (but with different parameters) for both the visual and textual encoder, *i.e.* ViT-B/16.

**Multimodal encoder.** We use a separate transformer to fuse the image and text hidden states. Specifically, we apply two learned linear projections over each hidden state vector in \{h_I\} and \{h_T\}, and concatenate them into a single list with an additional [CLS_M] token added, as shown in Figure 2. This concatenated list is fed into the multimodal encoder transformer (also based on the ViT architecture), allowing cross-attention between the projected unimodal image and text representations and fusing the two modalities. The output from the multimodal encoder is a list of hidden states \{h_M\}, each corresponding to a unimodal vector from \{h_I\} or \{h_T\} (and a vector \textbf{h}_{CLS,M} for [CLS_M]).

**Applying to downstream tasks.** The FLAVA model can be applied to both unimodal and multimodal tasks in a straightforward manner. For visual recognition tasks (*e.g.* ImageNet classification), we apply a classifier head (*e.g.* a linear layer or a multi-layer perceptron) on top of the unimodal \textbf{h}_{CLS,I} from the image encoder. Similarly, for language understanding and multimodal reasoning tasks, we apply a classifier head on top of \textbf{h}_{CLS,T} from the text encoder or \textbf{h}_{CLS,M} from the multimodal encoder, respectively. We pretrain the FLAVA model once, and evaluate it separately on each downstream task. More details about finetuning, linear, and zero-shot evaluation on specific tasks can be found in the supplemental.

**3.2. Multimodal pretraining objectives**

We aim to obtain strong representations through pretraining on both multimodal data (paired image and text)
as well as unimodal data (unpaired images or text). FLAVA pretraining involves the following multimodal objectives.

**Global contrastive (GC) loss.** Our image-text contrastive loss resembles that of CLIP [82]. Given a batch of images and text, we maximize the cosine similarities between matched image and text pairs and minimize those for the unmatched pairs. This is accomplished by linearly projecting each $h_{CLS,I}$ and $h_{CLS,T}$ into an embedding space, followed by L2-normalization, dot-product, and a softmax loss scaled by temperature.

Large models are often trained using multiple GPUs data parallelism, where the samples in a batch are split across GPUs. When gathering embeddings for the image and text contrastive objective, the open-source CLIP implementation [48] only back-propagates the gradients of the contrastive loss to the embeddings from the local GPU where the dot-product is performed. In contrast, through experiments that can be found in the supplemental, we observe a noticeable performance gain by performing full back-propagation across GPUs compared to only doing back-propagation locally. We call our loss “global contrastive” $L_{GC}$ to distinguish it from “local contrastive” approaches.

**Masked multimodal modeling (MMM).** While a number of previous vision-and-language pretraining approaches (e.g., [63]) have focused on masked modeling of the text modality by reconstructing masked tokens from the multimodal input, most of them do not involve masked learning on image modality directly at the image pixel level in an end-to-end manner. Here, we introduce a novel masked multimodal modeling (MMM) pretraining objective $L_{MMM}$ that masks both the image patches and the text tokens and jointly works on both modalities.

Specifically, given an image and text input, we first tokenize the input image patches using a pretrained dVAE tokenizer [88], which maps each image patch into an index in a visual codebook similar to a word dictionary (we use the same dVAE tokenizer as in [5]). Then, we replace a subset of image patches based on rectangular block image regions following BEiT [5] and 15% of text tokens following BERT [28] with a special [MASK] token. Then, from the multimodal encoder’s output $\{h_M\}$, we apply a multilayer perceptron to predict the visual codebook index of the masked image patches, or the word vocabulary index of the masked text tokens.

This objective can be seen as an extension of the multimodal masked language modeling such that it incorporates masking on the image side. In our experiments, we find that our MMM pretraining leads to improvements over and in addition to the contrastive loss pretraining, especially for multimodal downstream tasks such as VQA. Note that we apply global contrastive loss on image patches and text tokens without any masking, which are forwarded through the image and text encoders separately from the MMM loss.

**Image-text matching (ITM).** Finally, we add an image-text matching loss $L_{ITM}$ following prior vision-and-language pretraining literature [16, 70, 100]. During pretraining, we feed a batch of samples including both matched and unmatched image-text pairs. Then, on top of $L_{CLS,M}$ from the multimodal encoder, we apply a classifier to decide if an input image and text match each other.

### 3.3. Unimodal pretraining objectives

While the objectives in Sec. 3.2 allow pretraining the FLAVA model on paired image-and-text data, the vast majority of datasets (such as ImageNet for images and CCNews for text) are unimodal without paired data from the other modality. To efficiently learn a representation for a wide range of downstream tasks, we would also like to leverage these datasets and incorporate unimodal and unaligned information into our representations.

In this work, we introduce knowledge and information from these unimodal datasets through 1) pretraining the image encoder and text encoder on unimodal datasets; 2) pretraining the entire FLAVA model jointly on both unimodal and multimodal datasets; or 3) a combination of both by starting from pretrained encoders and then jointly training. When applied to stand-alone image or text data, we adopt masked image modeling (MIM) and masked language modeling (MLM) losses over the image and text encoders respectively, as described in what follows.

**Masked image modeling (MIM).** On unimodal image datasets, we mask a set of image patches following the rectangular block-wise masking in BEiT [5] and reconstruct them from other image patches. The input image is first tokenized using a pretrained dVAE tokenizer [88] (same as the one used in the MMM objective in Sec. 3.2), and then a classifier is applied on the image encoder outputs $\{h_I\}$ to predict the dVAE tokens of the masked patches.

**Masked language modeling (MLM).** We apply a masked language modeling loss [28] on top of the text encoder to pretrain on stand-alone text datasets. A fraction (15%) of the text tokens are masked in the input, and reconstructed from the other tokens using a classifier over the unimodal text hidden states output $\{h_T\}$.

**Encoder initialization from unimodal pretraining.** We use three sources of data for pretraining: unimodal image data (ImageNet-1K [89]), unimodal text data (CCNews [68] and BookCorpus [118]), and multimodal image-text paired data (Sec. 3.5). We first pretrain the text encoder with the MLM objective on the unimodal text dataset. We experiment with different ways for pretraining the image encoder: we pretrain the image encoder on unpaired image datasets with either MIM or the DINO objective [13], before joint training on both unimodal and multimodal datasets. We empirically found the latter to work quite well, despite the
switch to an MIM objective on images post-initialization (more details in supplemental). Then, we initialize the whole FLAVA model with the two respective unimodally-pretrained encoders, or when we train from scratch, we initialize randomly. We always initialize the multimodal encoder randomly for pretraining.

**Joint unimodal and multimodal training.** After unimodal pretraining of the image and text encoders, we continue training the entire FLAVA model jointly on the three types of datasets with round-robin sampling. In each training iteration, we choose one of the datasets according to a sampling ratio that we determine empirically (see supplemental) and obtain a batch of samples. Then, depending on the dataset type, we apply unimodal MIM on image data, unimodal MLM on text data, or the multimodal losses (contrastive, MIM, and ITM) in Sec. 3.2 on image-text pairs.

### 3.4. Implementation details

We find that the optimizer hyperparameters play a critical role in effective pretraining. A large batch size, a large weight decay, and a long warm-up are all important for preventing divergence with a large learning rate (we use 8,192 batch size, 1e-3 learning rate, 0.1 weight decay, and 10,000 iteration warm-up in our pretraining tasks together with the AdamW optimizer [55, 69]). In addition, the ViT transformer architecture (which applies layer norm [3] before the multi-head attention rather than after [112]) provides more robust learning for the text encoder under large learning rate than the BERT [28] transformer architecture. FLAVA is implemented using the open-source MIMF [92] and fairseq [78] libraries. We use Fully-Sharded Data Parallel (FSDP) [85, 86] and train in full FP16 precision except the layer norm [3] to reduce GPU memory consumption.

### 3.5. Data: Public Multimodal Datasets (PMD)

For multimodal pretraining, we constructed a corpus out of publicly available sources of image-text data, which are presented in Table 2 with examples in Fig. 3. The total count of text-image pairs is 70M, including 68M unique images, and the average caption length is 12.1 words. For the YFCC100M dataset [101], we filter the image-text data by discarding non-English captions and only keeping captions that contain more than two words. We first consider the description field of each image, if this does not pass our filters we consider the title field. Other than that, we did not do any additional filtering. Importantly, this corpus entirely consists of open datasets that are freely accessible by other researchers, facilitating reproducibility and enabling future work by the community.

### 4. Experiments

We evaluate FLAVA across vision, language, and multimodal tasks. For vision, we evaluate on 22 common vision tasks. For NLP, we evaluate on 8 tasks from the GLUE [106] benchmark. For multimodal, we evaluate on VQA-v2 [39], SNLI-VE [111], Hateful Memes [53], as well as Flickr30K [80] and COCO [66] image and text retrieval.

We compare our joint pretraining method (FLAVA in Table 3 and 4) with other settings, on this diverse array of 35 tasks. We report the average performance on the NLP, vision, and multimodal tasks, and an additional macro average across all the three modalities in Table 3, and also the detailed the performance on each task in Table 4.

**Full FLAVA pretraining achieves the best results.** Table 3 shows baselines and different ablation settings of FLAVA, including: models trained with unimodal MIM and MLM losses, FLAVA\textsubscript{C} trained with only image-text contrastive loss, FLAVA\textsubscript{MM} trained only on multimodal data, models without unimodal initialization, and the full model (each setting is detailed in the paragraphs below). The full FLAVA model in row 6 outperforms all other settings in average performance over NLP, vision, and multimodal tasks.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Image-Text Pairs</th>
<th>Avg. text length</th>
</tr>
</thead>
<tbody>
<tr>
<td>COCO [66]</td>
<td>0.9M</td>
<td>12.4</td>
</tr>
<tr>
<td>SBU Captions [77]</td>
<td>1.6M</td>
<td>12.1</td>
</tr>
<tr>
<td>Localized Narratives [81]</td>
<td>1.9M</td>
<td>13.8</td>
</tr>
<tr>
<td>Conceptual Captions [90]</td>
<td>3.1M</td>
<td>10.3</td>
</tr>
<tr>
<td>Visual Genome [57]</td>
<td>5.4M</td>
<td>5.1</td>
</tr>
<tr>
<td>Wikipedia Image Text [97]</td>
<td>4.8M</td>
<td>12.8</td>
</tr>
<tr>
<td>Conceptual Captions 12M [14]</td>
<td>11.0M</td>
<td>17.3</td>
</tr>
<tr>
<td>Red Caps [27]</td>
<td>11.6M</td>
<td>9.5</td>
</tr>
<tr>
<td>YFCC100M [101], filtered</td>
<td>30.3M</td>
<td>12.7</td>
</tr>
</tbody>
</table>

Table 2. Public Multimodal Datasets (PMD) corpus used in FLAVA multimodal pretraining, which consists of publicly available datasets with a total size of 70M image and text pairs.

---

**Figure 3.** Representative examples from various subsets of our pretraining dataset (details in Sec. 3.5).
Effective global contrastive loss in FLAVA. We next perform a step-by-step ablation of our model (Table 4). We first train a restricted version of FLAVA using only the global contrastive loss \( L_{GC} \) in Sec. 3.2 on multimodal data, denoted as FLAVA\(_C\) in column 3. This restricted setting is a conceptually similar model to CLIP [82] that also involves a contrastive loss, and we compare against the CLIP model trained on the same PMD data with the same ViT-B/16 image encoder as a baseline (using the open-source implementation in [48]), denoted as CLIP in column 7.\(^1\) Comparing column 3 vs 7, we see that FLAVA\(_C\) outperforms it in all vision, language, and multimodal domains. This can be attributed to mostly two factors: different model details of FLAVA (e.g., 768 text encoder hidden size instead of 512) and performing global back-propagation across all GPU workers as mentioned in Sec. 3.2. In a more detailed analysis, we find that the latter improves our macro average over vision, NLP, and multi-modal tasks by +1.65% with only minor additional computation overhead, indicating the global back-propagation implementation in contrastive loss are critical to effective pretraining.

**MMM and ITM objectives benefit multimodal tasks.** Next, we include the other multimodal objectives from Sec. 3.2 into our pretraining, using \( L_{MIM} \) and \( L_{ITM} \) along with \( L_{GC} \). The results are denoted as FLAVA\(_MM\) in Table 4 column 4. Compared to FLAVA\(_C\) with only the contrastive loss \( L_{GC} \) (column 3 vs 4), this setting improves multimodal average score by +2.86%, NLP average score by +9%, and also vision average score slightly by +0.3%.

We additionally compare FLAVA\(_MM\) with two other baseline settings – the FLAVA model trained with only unimodal MIM or MLM losses in Sec. 3.3, respectively over the images or the text in PMD. These two baselines are shown in Table 4 column 1 and 2, which are largely outperformed by FLAVA\(_MM\). These results indicate that the combined multimodal objectives (contrastive, MMM, ITM) allow FLAVA to learn powerful representations for both unimodal and multimodal downstream tasks.

---

\(^1\)We fine-tune CLIP on multimodal downstream tasks (VQAv2, SNLI-VE, and HM) by applying a classifier on the concatenation of the two output vectors from its image and text encoders (details in supplemental).

### Table 3: Our full FLAVA pretraining (row 6) achieves the best average scores on vision, language, and multimodal tasks compared to ablations. Row 1 to 4 are pretrained on PMD while row 5 and 6 also involve unimodal IN-1k, CCNews, and BookCorpus datasets.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1 MIM</td>
<td>57.46</td>
<td>–</td>
<td>–</td>
<td>19.15</td>
</tr>
<tr>
<td>2 MLM</td>
<td>–</td>
<td>71.55</td>
<td>–</td>
<td>23.85</td>
</tr>
<tr>
<td>3 FLAVA(_C)</td>
<td>64.80</td>
<td>79.14</td>
<td>66.25</td>
<td>70.06</td>
</tr>
<tr>
<td>4 FLAVA(_MM)</td>
<td>74.22</td>
<td>79.35</td>
<td>69.11</td>
<td>74.23</td>
</tr>
<tr>
<td>5 FLAVA w/o unimodal init</td>
<td>75.55</td>
<td>78.29</td>
<td>67.32</td>
<td>73.72</td>
</tr>
<tr>
<td>6 FLAVA</td>
<td>78.19</td>
<td>79.44</td>
<td>69.92</td>
<td>75.85</td>
</tr>
</tbody>
</table>

**Joint unimodal & multimodal pretraining helps NLP.** For the full FLAVA pretraining, we introduce unimodal image data from ImageNet-1k (IN-1k) and text data from CCNews and BookCorpus (BC). In this setting, we apply FLAVA\(_MM\) losses on PMD data batches, MIM loss on IN-1k unimodal image data and MLM loss on CCNews text data, following Sec. 3.3, shown in Table 4 column 5. Comparing it to FLAVA\(_MM\) in column 4 with only multimodal pretraining, this joint unimodal and multimodal pretraining improves the NLP average score from 74.22 to 75.55, which suggests that the additional text data from CCNews and BookCorpus benefits language understanding through the MLM objective.

However, we also observe from column 4 vs 5 that the macro average over all tasks decreases slightly. We suspect that this is because adding different tasks to the mix makes the optimization problem much harder, especially when the whole model is randomly initialized. Also, the round-robin sampling of tasks does not follow any particular curriculum to order the learning sequence of these tasks. Naturally, having some vision and language understanding is important before learning multimodal tasks, which motivates us to explore first leveraging unimodal pretraining before the joint training, as described below.

### Better image and text encoders via unimodal pretraining.

As detailed in Section 3.3, in order to leverage unimodal learning before joint training, we initialize the model from pretrained self-supervised weights for both vision and language encoders. For vision encoder, we initialize from an off-the-shelf DINO model pretrained on ImageNet-1k [89]. For the language encoder, we pretrain a ViT model with MLM loss on CCNews and BookCorpus datasets and use its model weights. Comparing column 5 vs 6, we observe pretrained encoders boost the performance of FLAVA on all tasks. We empirically find that initializing the vision encoder from a DINO self-supervised model gives better performance compared to a BEiT self-supervised model (see supplemental for additional details).

### 4.1. Comparison to state-of-the-art models

We compare our full FLAVA model (Table 4 column 6) with several state-of-the-art models on multimodal tasks, language tasks, and ImageNet linear evaluation, in Table 5. FLAVA largely outperforms previous multimodal approaches pretrained on public data (row 4 to 11) on both language and multimodal tasks and approaches the well-established BERT model on several GLUE tasks.

FLAVA combines unimodal and multimodal losses and learns more generic representations which are transferable to vision, language, and multimodal tasks. We evaluate the best released CLIP [82] ViT-B/16 model (pretrained on 400M image-text pairs in [82] with the same image encoder architecture as in FLAVA) on our task benchmark, shown
in Table 5 row 2. Compared to CLIP, we train FLAVA on just 70M data which is ∼6x smaller. In Fig. 4, we observe that FLAVA works significantly better on language and multimodal tasks while slightly worse than CLIP on some vision-only tasks. In addition, we note that FLAVA outperforms the variant of the CLIP model pretrained only on the PMD dataset (Table 5 row 10). Table 4 further shows a breakdown analysis between our model (column 6) and the released CLIP ViT-B/16 (400M) model (column 7) and the CLIP trained on PMD (column 8).

FLAVA also has comparable performance to SimVLM [107] (Table 5 row 3) on language tasks while underperforms the variant of the CLIP model pretrained only on some vision-only tasks. In addition, we note that FLAVA outperforms the variant of the CLIP model pretrained only on the PMD dataset (Table 5 row 10). Table 4 further shows a breakdown analysis between our model (column 6) and the released CLIP ViT-B/16 (400M) model (column 7) and the CLIP trained on PMD (column 8).

Table 4. Comparing our full FLAVA pretraining with other settings, where FLAVA gets the highest macro average score. MNLI numbers are average of MNLI-m and MNLI-mm. MRPC and QQP numbers are average of accuracy and F1. We report F1C for CoLA, MCC for STS-B, and AUROC for Hateful Memes, respectively. We perform zero-shot text retrieval and image retrieval (TR and IR) on all other tasks we report accuracy. Column 8 is the best released model in [82] based on ViT-B/16 pretrained on 400M image-text pairs. The overall best result is underlined while bold signifies the best on public data (PMD and unimodal).
Table 5. Comparing FLA V A (Table 4 column 6) with previous models on multimodal tasks, language tasks, and ImageNet linear evaluation. We report results on development sets of the GLUE benchmark [106]. We report Matthew’s Correlation for CoLA; accuracy/F1 for MRPC and QQP; the Pearson/Spearman correlation for STS-B; average of mismatched and matched accuracy for MNLI; AUROC for Hateful Memes; test-dev VQA score for VQAv2 and accuracy for all other tasks. The results for BERT and other VLP methods on GLUE benchmark are obtained from [47]. The results on V&L tasks are from original papers. For UniT, we use “shared, (COCO init.)” version. Note that SimVLM is pretrained on an order of magnitude more data than FLA V A (1.8B vs 70M). †: taken from [93]; ‡: taken from [53]. The overall best result among the multimodal approaches is underlined while bold signifies the best model trained on public data.

<table>
<thead>
<tr>
<th>public data</th>
<th>Multimodal Tasks</th>
<th>Language Tasks</th>
<th>ImageNet linear eval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VQAv2</td>
<td>CoLA</td>
<td>MNLI</td>
</tr>
<tr>
<td></td>
<td>SNLI-VE   HM</td>
<td>SST-2</td>
<td>QQP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RTE</td>
<td>MRPC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>QQP</td>
<td>NLI</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>STS-B</td>
</tr>
<tr>
<td>1 ✓ BERTbase [28]</td>
<td>–</td>
<td>54.6 92.5 62.5</td>
<td>81.9/87.6 90.6/87.4</td>
</tr>
<tr>
<td>2 ✓ CLIP-ViT-B/16 [82]</td>
<td>55.3 74.0 63.4</td>
<td>25.4 88.2 55.2</td>
<td>74.9/65.0 76.8/53.9</td>
</tr>
<tr>
<td>3 ✓ SimVLMbase [107]</td>
<td>77.9 84.2 –</td>
<td>46.7 90.9 63.9</td>
<td>75.2/84.4 90.4/87.2</td>
</tr>
<tr>
<td>4 ✓ VisualBERT [63]</td>
<td>70.8 77.3 74.1 ‡</td>
<td>38.6 89.4 56.6</td>
<td>71.9/82.1 89.4/86.0</td>
</tr>
<tr>
<td>5 ✓ UNITERbase [16]</td>
<td>72.7 78.3 –</td>
<td>37.4 89.7 55.6</td>
<td>69.3/80.3 89.2/85.7</td>
</tr>
<tr>
<td>6 ✓ VL-BERTbase [99]</td>
<td>71.2 –</td>
<td>38.7 89.8 55.7</td>
<td>70.6/81.8 89.0/85.4</td>
</tr>
<tr>
<td>7 ✓ ViLBERT [70]</td>
<td>70.6 75.7 † 74.1 ‡</td>
<td>36.1 90.4 53.7</td>
<td>69.0/79.4 88.6/85.0</td>
</tr>
<tr>
<td>8 ✓ LXMER T [100]</td>
<td>72.4 –</td>
<td>39.0 90.2 57.2</td>
<td>69.7/80.4 75.3/75.3</td>
</tr>
<tr>
<td>9 ✓ UniT [43]</td>
<td>67.0 73.1 –</td>
<td>89.3 –</td>
<td>90.6/ – 81.5 –</td>
</tr>
<tr>
<td>10 ✓ CLIP-ViT-B/16 (PMD)</td>
<td>59.8 73.5 56.6</td>
<td>11.0 83.5 53.1</td>
<td>63.3/86.7 75.4/43.0</td>
</tr>
<tr>
<td>11 ✓ FLA V A (ours)</td>
<td>72.8 79.0 76.7</td>
<td>50.7 90.9 87.8</td>
<td>81.4/86.9 90.4/87.2</td>
</tr>
</tbody>
</table>

5. Conclusion

In this work, we have presented a foundational vision and language alignment model that performs well on all three target modalities: 1) vision, 2) language, and 3) vision & language. We introduced a novel set of objectives to achieve this goal and conducted experiments on a wide variety of 35 tasks to analyze the model’s performance. FLA V A was trained on a corpus of publicly available datasets that is several orders of magnitude smaller than similar recent models, but still obtained better or competitive performance. Our work points the way forward towards generalized but open models that perform well on a wide variety of multimodal tasks.

Broader impacts and limitations. The models in this work are trained on public datasets widely used in the community. This enables reproducibility and we hope that our work will motivate others to compare models across a wide area of tasks and domains with the same data. However, like all natural data, these datasets have biases, potentially affecting our models. We partly mitigate this by combining several public datasets to increase the diversity and evaluating on an even larger set of target datasets. Still, further study is needed to identify and reduce potentially harmful biases.

Acknowledgements. We thank Devi Parikh for her support and advice on this project. We are grateful to Dmytro Okhonko, Hu Xu, Armen Aghajanyan, Po-Yao Huang, Min Xu, and Aleksandra Pikrus for joint explorations of multimodal data. We thank Ning Zhang, Madian Khabsa, Sasha Sheng, and Naman Goyal for useful technical discussions; Karan Desai for providing access to RedCaps; Vaibhav Singh and others on the Google TPU team for TPU support; Shubho Sengupta, Armand Joulin, Brian O’Horo, Arthur Menezes for compute and storage support; and Ryan Jiang, Kushal Tirumala and Russ Howes for help running experiments.
References


[19] Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Saksham Singhal, Payal Bajaj, Xia Song, and Furu Wei. XLM-R: Cross-lingual language model pre-training via ELECTRA. *arXiv*, abs/2106.16138, 2021. 2


