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

*Equal contribution.

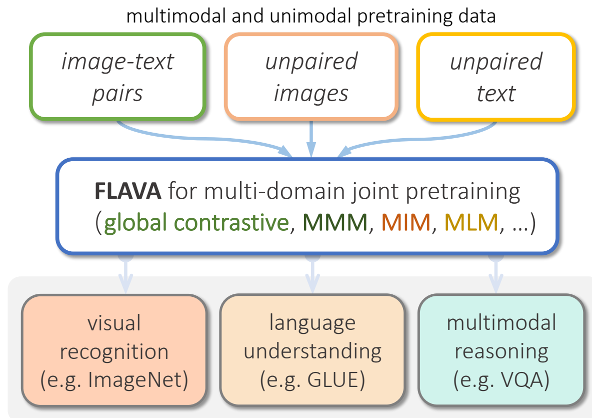

Figure 1. We present FLAVA, a language and vision alignment model that learns strong representations from multimodal (image-text pairs) and unimodal data (unpaired images and text) and can be applied to target a broad scope of tasks from three domains (visual recognition, language understanding, and multimodal reasoning) under a common transformer model architecture.

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/`.

Method	Multimodal Pretraining data			Pretraining Objectives				Target Modalities			
	public	dataset(s)	size	Contr.	ITM	Masking	Unimodal	V	CV&L	MV&L	L
CLIP [82]	✗	WebImageText	400M	✓	–	–	–	✓	✓	–	–
ALIGN [50]	✗	JFT	1.8B	✓	–	–	–	✓	✓	–	–
SimVLM [107]	✗	JFT	1.8B	–	–	PrefixLM	CLM	*	✓	✓	✓
UniT [43]	–	None	–	–	–	–	–	*	–	✓	✓
VinVL [115]	✓	Combination	9M	✓	–	MLM	–	–	✓	✓	–
ViLT [54]	✓	Combination	10M	–	✓	MLM	–	–	✓	✓	–
ALBEF [62]	✓	Combination	5M	✓	✓	MLM	–	–	✓	✓	–
FLAVA (ours)	✓	PMD (Tbl. 2)	70M	✓	✓	MMM	MLM+MIM	✓	✓	✓	✓

Table 1. Comparison of recent models in different modalities. CV&L and MV&L stands for cross-modal and multi-modal vision-and-language. * means the modality is partially targeted (SimVLM [107] and UniT [43] include ImageNet and object detection, respectively).

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 is 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 N^2 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

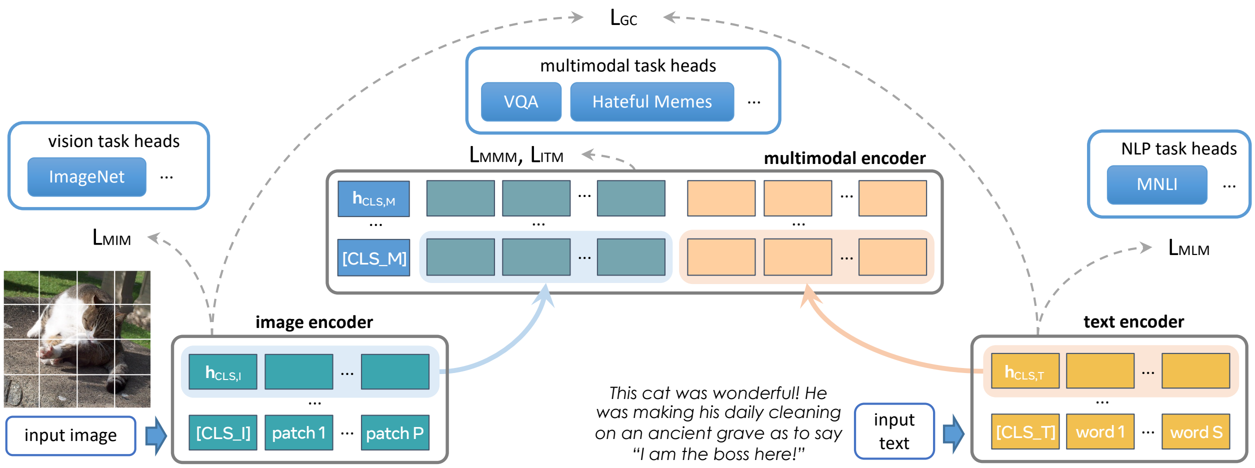

Figure 2. **An overview of our FLAVA model**, with an image encoder transformer to capture unimodal image representations, a text encoder transformer to process unimodal text information, and a multimodal encoder transformer that takes as input the encoded unimodal image and text and integrates their representations for multimodal reasoning. **During pretraining**, masked image modeling (MIM) and mask language modeling (MLM) losses are applied onto the image and text encoders over a single image or a text piece, respectively, while contrastive, masked multimodal modeling (MMM), and image-text matching (ITM) loss are used over paired image-text data. **For downstream tasks**, classification heads are applied on the outputs from the image, text, and multimodal encoders respectively for visual recognition, language understanding, and multimodal reasoning tasks.

based on transformers [103] (Sec. 3.1), which incorporates multimodal pretraining losses on image-text data (Sec. 3.2) as well as unimodal pretraining losses on unimodal data (Sec. 3.3). We discuss additional critical modeling insights in Sec. 3.4. Finally, we demonstrate that our pretrained models can be successfully applied to a wide range of image, text, and multimodal tasks through both zero-shot and fine-tuning evaluations.

3.1. The model architecture

The FLAVA model architecture is shown in Figure 2. The model involves an *image encoder* to extract unimodal image representations, a *text encoder* to obtain unimodal text representations, and a *multimodal encoder* to fuse and align the image and text representations for multimodal reasoning, all of which are based on transformers.

Image encoder. We adopt the ViT architecture [31] for the image encoder. Given an input image, we resize it to a fixed image size and split the image into patches, which are then linearly embedded and fed into a transformer model (along with positional embeddings and an extra image classification token `[CLS_I]`). The image encoder output is a list of image hidden state vectors $\{\mathbf{h}_I\}$, each corresponding to an image patch, plus an additional $\mathbf{h}_{\text{CLS},I}$ for `[CLS_I]`. We use the ViT-B/16 architecture for our image encoder.

Text encoder. Given an input piece of text (*e.g.*, a sentence or a pair of sentences), we first tokenize and embed it into a list of word vectors following [28]. Then, we apply a transformer model over the word vectors to encode them into a list of hidden state vectors $\{\mathbf{h}_T\}$, including $\mathbf{h}_{\text{CLS},T}$ for the text classification `[CLS_T]` token. Importantly, different 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 $\{\mathbf{h}_I\}$ and $\{\mathbf{h}_T\}$, 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 $\{\mathbf{h}_M\}$, each corresponding to a unimodal vector from $\{\mathbf{h}_I\}$ or $\{\mathbf{h}_T\}$ (and a vector $\mathbf{h}_{\text{CLS},M}$ 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 $\mathbf{h}_{\text{CLS},I}$ from the image encoder. Similarly, for language understanding and multimodal reasoning tasks, we apply a classifier head on top of $\mathbf{h}_{\text{CLS},T}$ from the text encoder or $\mathbf{h}_{\text{CLS},M}$ 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 $\mathbf{h}_{\text{CLS},I}$ and $\mathbf{h}_{\text{CLS},T}$ 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 $\{\mathbf{h}_M\}$, 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 $\mathbf{h}_{\text{CLS},M}$ 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 CC-News 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 $\{\mathbf{h}_I\}$ 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 $\{\mathbf{h}_T\}$.

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

Figure 3. Representative examples from various subsets of our pretraining dataset (details in Sec. 3.5).

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, MMM, 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 MMF [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.

	#Image-Text Pairs	Avg. text length
COCO [66]	0.9M	12.4
SBU Captions [77]	1.0M	12.1
Localized Narratives [81]	1.9M	13.8
Conceptual Captions [90]	3.1M	10.3
Visual Genome [57]	5.4M	5.1
Wikipedia Image Text [97]	4.8M	12.8
Conceptual Captions 12M [14]	11.0M	17.3
Red Caps [27]	11.6M	9.5
YFCC100M [101], filtered	30.3M	12.7
Total	70M	12.1

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.

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 VQAv2 [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_C trained with only image-text contrastive loss, FLAVA_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.

Method	Vision Avg.	NLP Avg.	Multi-modal Avg.	Macro Avg.
1 MIM	57.46	–	–	19.15
2 MLM	–	71.55	–	23.85
3 FLAVA_{C}	64.80	79.14	66.25	70.06
4 FLAVA_{MM}	74.22	79.35	69.11	74.23
5 FLAVA w/o unimodal init	75.55	78.29	67.32	73.72
6 FLAVA	**78.19**	**79.44**	**69.92**	**75.85**

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.

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_{MMM} 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).

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

		MIM 1	MLM 2	FLAVA_{C} 3	FLAVA_{MM} 4	FLAVA w/o init 5	FLAVA 6	CLIP 7	CLIP 8
Datasets	Eval method	PMD	PMD	PMD	PMD	(PMD+IN-1k+CCNews+BC)		PMD	400M [82]
MNLI [109]	fine-tuning	–	73.23	70.99	76.82	78.06	**80.33**	32.85	33.52
CoLA [108]	fine-tuning	–	39.55	17.58	38.97	44.22	**50.65**	11.02	25.37
MRPC [29]	fine-tuning	–	73.24	76.31	79.14	78.91	**84.16**	68.74	69.91
QQP [49]	fine-tuning	–	86.68	85.94	88.49	98.61	**88.74**	59.17	65.33
SST-2 [95]	fine-tuning	–	87.96	86.47	89.33	90.14	**90.94**	83.49	88.19
QNLI [87]	fine-tuning	–	82.32	71.85	84.77	86.40	**87.31**	49.46	50.54
RTE [7, 25, 36, 40]	fine-tuning	–	50.54	51.99	51.99	54.87	**57.76**	53.07	55.23
STS-B [1]	fine-tuning	–	78.89	57.28	84.29	83.21	**85.67**	13.70	15.98
NLP Avg.		–	71.55	64.80	74.22	75.55	**78.19**	46.44	50.50
ImageNet [89]	linear eval	41.79	–	74.09	74.34	73.49	**75.54**	72.95	80.20
Food101 [11]	linear eval	53.30	–	87.77	87.53	87.39	**88.51**	85.49	91.56
CIFAR10 [58]	linear eval	76.20	–	**93.44**	92.37	92.63	92.87	91.25	94.93
CIFAR100 [58]	linear eval	55.57	–	**78.37**	78.01	76.49	77.68	74.40	81.10
Cars [56]	linear eval	14.71	–	**72.12**	72.07	66.81	70.87	62.84	85.92
Aircraft [74]	linear eval	13.83	–	**49.74**	48.90	44.73	47.31	40.02	51.40
DTD [20]	linear eval	55.53	–	76.86	76.91	75.80	**77.29**	73.40	78.46
Pets [79]	linear eval	34.48	–	**84.98**	84.93	82.77	84.82	79.61	91.66
Caltech101 [32]	linear eval	67.36	–	94.91	95.32	94.95	**95.74**	93.76	95.51
Flowers102 [76]	linear eval	67.23	–	96.36	**96.39**	95.58	96.37	94.94	97.12
MNIST [60]	linear eval	96.40	–	98.39	98.58	**98.70**	98.42	97.38	99.01
STL10 [21]	linear eval	80.12	–	98.06	98.31	98.32	**98.89**	97.29	99.09
EuroSAT [41]	linear eval	95.48	–	97.00	96.98	97.04	**97.26**	95.70	95.38
GTSRB [98]	linear eval	63.14	–	78.92	77.93	77.71	**79.46**	76.34	88.61
KITTI [35]	linear eval	86.03	–	87.83	88.84	88.70	**89.04**	84.89	86.56
PCAM [104]	linear eval	85.10	–	85.02	85.51	**85.72**	85.31	83.99	83.72
UCF101 [96]	linear eval	46.34	–	82.69	82.90	81.42	**83.32**	77.85	85.17
CLEVR [52]	linear eval	61.51	–	79.35	**81.66**	80.62	79.66	73.64	75.89
FER 2013 [38]	linear eval	50.98	–	59.96	60.87	58.99	**61.12**	57.04	68.36
SUN397 [110]	linear eval	52.45	–	81.27	81.41	81.05	**82.17**	79.96	82.05
SST [82]	linear eval	57.77	–	56.67	**59.25**	56.40	57.11	56.84	74.68
Country211 [82]	linear eval	8.87	–	27.27	26.75	27.01	**28.92**	25.12	30.10
Vision Avg.		57.46	–	79.14	79.35	78.29	**79.44**	76.12	82.57
VQAv2 [39]	fine-tuning	–	–	67.13	71.69	71.29	**72.49**	59.81	54.83
SNLI-VE [111]	fine-tuning	–	–	73.27	78.36	78.14	**78.89**	73.53	74.27
Hateful Memes [53]	fine-tuning	–	–	55.58	70.72	**77.45**	76.09	56.59	63.93
Flickr30K [80] TR R@1	zero-shot	–	–	68.30	**69.30**	64.50	67.70	60.90	82.20
Flickr30K [80] TR R@5	zero-shot	–	–	93.50	92.90	90.30	**94.00**	88.90	96.60
Flickr30K [80] IR R@1	zero-shot	–	–	60.56	63.16	60.04	**65.22**	56.48	62.08
Flickr30K [80] IR R@5	zero-shot	–	–	86.68	87.70	86.46	**89.38**	83.60	85.68
COCO [66] TR R@1	zero-shot	–	–	43.08	**43.48**	39.88	42.74	37.12	52.48
COCO [66] TR R@5	zero-shot	–	–	75.82	**76.76**	72.84	**76.76**	69.48	76.68
COCO [66] IR R@1	zero-shot	–	–	37.59	**38.46**	34.95	38.38	33.29	33.07
COCO [66] IR R@5	zero-shot	–	–	67.28	**67.68**	64.63	67.47	62.47	58.37
Multimodal Avg.		–	–	66.25	69.11	67.32	**69.92**	62.02	67.29
Macro Avg.		19.15	23.85	70.06	74.23	73.72	**75.85**	61.52	66.78

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 PCC 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 Flickr30K and COCO based on their matching scores from the contrastive loss and report top-1 and top-5 recall. For 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).

in Table 5 row 2. Compared to CLIP, we train FLAVA on just 70M data which is $\sim$**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 8) and the CLIP trained on PMD (column 7).

FLAVA also has comparable performance to SimVLM [107] (Table 5 row 3) on language tasks while underper-

	public data		Multimodal Tasks VQAv2	SNLI-VE	HM	Language Tasks CoLA	SST-2	RTE	MRPC	QQP	MNLI	QNLI	STS-B	ImageNet linear eval
1	✓	BERT$_{base}$ [28]	–	–	–	54.6	92.5	62.5	81.9/87.6	90.6/87.4	84.4	91.0	88.1	–
2	✗	CLIP-ViT-B/16 [82]	55.3	74.0	63.4	25.4	88.2	55.2	74.9/65.0	76.8/53.9	33.5	50.5	16.0	80.2
3	✗	SimVLM$_{base}$ [107]	77.9	84.2	–	46.7	90.9	63.9	75.2/84.4	90.4/87.2	83.4	88.6	–	80.6
4	✓	VisualBERT [63]	70.8	77.3†	74.1‡	38.6	89.4	56.6	71.9/82.1	89.4/86.0	**81.6**	87.0	81.8	–
5	✓	UNITER$_{base}$ [16]	72.7	78.3	–	37.4	89.7	55.6	69.3/80.3	89.2/85.7	80.9	86.0	75.3	–
6	✓	VL-BERT$_{base}$ [99]	71.2	–	–	38.7	89.8	55.7	70.6/81.8	89.0/85.4	81.2	86.3	82.9	–
7	✓	ViLBERT [70]	70.6	75.7†	74.1‡	36.1	90.4	53.7	69.0/79.4	88.6/85.0	79.9	83.8	77.9	–
8	✓	LXMERT [100]	72.4	–	–	39.0	90.2	57.2	69.7/80.4	75.3/75.3	80.4	84.2	75.3	–
9	✓	UniT [43]	67.0	73.1	–	–	89.3	–	–	90.6/ –	81.5	**88.0**	–	–
10	✓	CLIP-ViT-B/16 (PMD)	59.8	73.5	56.6	11.0	83.5	53.1	63.5/68.7	75.4/43.0	32.9	49.5	13.7	73.0
11	✓	FLAVA (ours)	**72.8**	**79.0**	**76.7**	**50.7**	**90.9**	**57.8**	**81.4/86.9**	**90.4/87.2**	80.3	87.3	**85.7**	**75.5**

Table 5. **Comparing FLAVA (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 FLAVA (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.

forming it on multimodal tasks and ImageNet linear evaluation. FLAVA is pretrained using a much smaller dataset compared to 1.8B image-text pairs in [107], and we anticipate that FLAVA's performance will further heavily improve as the pretraining dataset size increases.

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. FLAVA 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 Piktus 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.

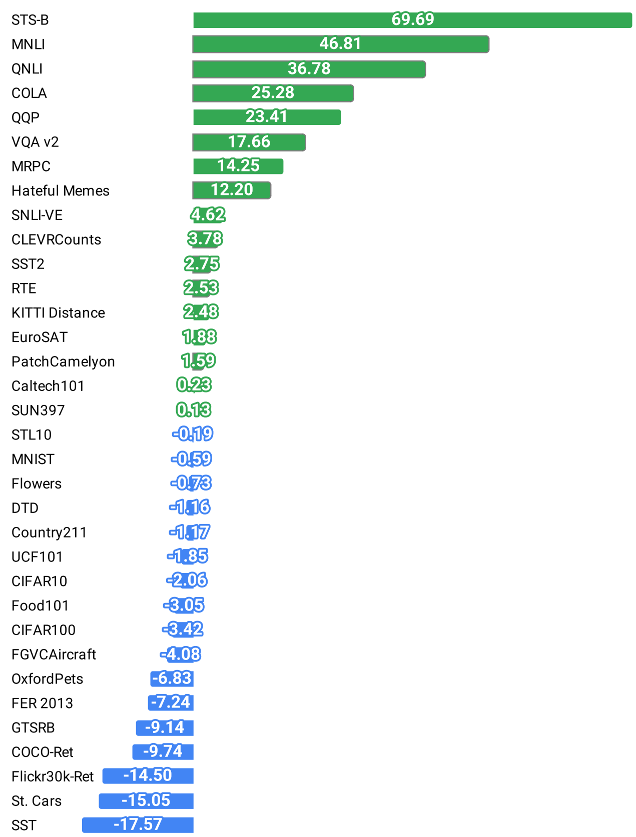

Figure 4. The performance difference (relative, in %) between FLAVA and the released CLIP-ViT-B/16 (400M) [82] on vision, language and multimodal tasks (positive means FLAVA is better).

References

[1] Eneko Agirre, Llu'is M'arquez, and Richard Wicentowski, editors. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics, Prague, Czech Republic, June 2007. 7

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021. 2

[3] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016. 5

[4] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of NeurIPS*, 2020. 2

[5] Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT pre-training of image transformers. *CoRR*, abs/2106.08254, 2021. 2, 4

[6] Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, and Hsiao-Wuen Hon. UniLMv2: Pseudo-masked language models for unified language model pre-training. In *Proceedings of ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 642–652. PMLR, 2020. 2

[7] Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. The fifth PASCAL recognizing textual entailment challenge. *TAC*, 2009. 7

[8] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021. 2

[9] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021. 1

[10] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1, 2

[11] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *Proceedings of ECCV*, 2014. 7

[12] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of ECCV*, 2020. 2

[13] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 4

[14] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of CVPR*, pages 3558–3568, 2021. 5

[15] 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. *arXiv preprint arXiv:1504.00325*, 2015. 2

[16] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: universal image-text representation learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Proceedings of ECCV*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer, 2020. 2, 4, 8

[17] Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xianling Mao, and Heyan Huang. Cross-lingual natural language generation via pre-training. *CoRR*, abs/1909.10481, 2019. 2

[18] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of NAACL-HLT*, pages 3576–3588, Online, June 2021. Association for Computational Linguistics. 2

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

[20] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *Proceedings of CVPR*, 2014. 7

[21] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, 2011. 7

[22] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020. 2

[23] 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 *Proceedings of ACL*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. 2

[24] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In *Proceedings of NeurIPS*, volume 32. Curran Associates, Inc., 2019. 2

[25] Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer, 2006. 7

[26] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. *arXiv preprint arXiv:2006.06666*, 2020. 2

[27] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. RedCaps: Web-curated image-text data created by the people, for the people. In *NeurIPS Datasets and Benchmarks*, 2021. 5

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019. 2, 3, 4, 5, 8

[29] William B Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing*, 2005. 7

[30] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *Proceedings of NeurIPS*, pages 13042–13054, 2019. 2

[31] 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 16x16 words: Transformers for image recognition at scale. In *Proceedings of ICLR*, 2021. 2, 3

[32] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *CVPR Workshop*, 2004. 7

[33] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 214–229. Springer, 2020. 2

[34] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Proceedings of NeurIPS*, 2020. 2

[35] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 7

[36] Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9. Association for Computational Linguistics, 2007. 7

[37] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of CVPR*, pages 244–253, 2019. 2

[38] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 7

[39] 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 *Proceedings of CVPR*, pages 6904–6913, 2017. 2, 5, 7

[40] R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, 2006. 7

[41] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 7

[42] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *arXiv preprint arXiv:2106.07447*, 2021. 2

[43] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of ICCV*, 2021. 2, 8

[44] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *Proceedings of CVPR*, pages 9992–10002, 2020. 2

[45] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of CVPR*, pages 12976–12985. Computer Vision Foundation / IEEE, 2021. 2

[46] 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. 2

[47] Taichi Iki and Akiko Aizawa. Effect of vision-and-language extensions on natural language understanding in vision-and-language models. *arXiv preprint arXiv:2104.08066*, 2021. 8

[48] Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. 4, 6

[49] Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. First Quora dataset release: Question pairs. `https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs`. Jan 2017. 7

[50] 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 Marina Meila and Tong Zhang, editors, *Proceedings of ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 2021. 1, 2

[51] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of CVPR*, pages 10267–10276, 2020. 2

[52] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of CVPR*, pages 2901–2910, 2017. 7

[53] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Proceedings of NeurIPS*, 33, 2020. 2, 5, 7, 8

[54] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region su-

pervision. In Marina Meila and Tong Zhang, editors, *Proceedings of ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR, 2021. 2

[55] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Proceedings of ICLR*, 2015. 5

[56] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 7

[57] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*, 2016. 2, 5

[58] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Citeseer*, 2009. 7

[59] Hung Le, Doyen Sahoo, Nancy F Chen, and Steven CH Hoi. Multimodal transformer networks for end-to-end video-grounded dialogue systems. *arXiv preprint arXiv:1907.01166*, 2019. 2

[60] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010. 7

[61] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 2

[62] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *CoRR*, abs/2107.07651, 2021. 2

[63] 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. 2, 4, 8

[64] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. UNIMO: towards unified-modal understanding and generation via cross-modal contrastive learning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the ACL/IJCNLP*, pages 2592–2607. Association for Computational Linguistics, 2021. 2

[65] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Proceedings of ECCV*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer, 2020. 2

[66] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of ECCV*, pages 740–755. Springer, 2014. 2, 5, 7

[67] Andy T Liu, Shang-Wen Li, and Hung-yi Lee. Tera: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2351–2366, 2021. 2

[68] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2, 4

[69] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 5

[70] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Proceedings of NeurIPS*, pages 13–23, 2019. 2, 4, 8

[71] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of CVPR*, pages 10437–10446, 2020. 2

[72] Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. Pretrained transformers as universal computation engines. *arXiv preprint arXiv:2103.05247*, 2021. 1

[73] Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *CoRR*, abs/2106.13736, 2021. 2

[74] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 7

[75] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. *arXiv preprint arXiv:2102.00719*, 2021. 2

[76] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008. 7

[77] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Proceedings of NeurIPS*, pages 1143–1151, 2011. 2, 5

[78] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019. 5

[79] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of CVPR*, 2012. 7

[80] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of ICCV*, pages 2641–2649. IEEE Computer Society, 2015. 2, 5, 7

[81] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *Proceedings of ECCV*, pages 647–664. Springer, 2020. 5

[82] 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 Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 1, 2, 4, 6, 7, 8

[83] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI Blog. https://openai.com/blog/language-unsupervised/*, 2018. 2

[84] 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. 2

[85] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. *preprint arXiv:1910.02054*, 2019. 5

[86] Samyam Rajbhandari, Olatunji Ruwase, Jeff Rasley, Shaden Smith, and Yuxiong He. Zero-infinity: Breaking the gpu memory wall for extreme scale deep learning. *arXiv preprint arXiv:2104.07857*, 2021. 5

[87] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP/IJCNLP*, pages 2383–2392. Association for Computational Linguistics, 2016. 7

[88] A. Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092, 2021. 4

[89] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 2015. 4, 6, 7

[90] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, pages 2556–2565, 2018. 2, 5

[91] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Proceedings of ECCV*, pages 742–758. Springer, 2020. 2

[92] Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Mmf: A multimodal framework for vision and language research. https://github.com/facebookresearch/mmf, 2020. 5

[93] Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. Are we pretraining it right? digging deeper into visio-linguistic pretraining. *arXiv preprint arXiv:2004.08744*, 2020. 2, 8

[94] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of CVPR*, pages 8317–8326, 2019. 2

[95] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*, pages 1631–1642, 2013. 7

[96] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 7

[97] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. *arXiv preprint arXiv:2103.01913*, 2021. 5

[98] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE, 2011. 7

[99] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: pre-training of generic visual-linguistic representations. In *Proceedings of ICLR*. OpenReview.net, 2020. 2, 8

[100] Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of EMNLP-IJCNLP*, pages 5099–5110. Association for Computational Linguistics, 2019. 2, 4, 8

[101] 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. *Communications of the ACM*, 59(2):64–73, 2016. 5

[102] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020. 2

[103] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of NeurIPS*, 2017. 3

[104] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pages 210–218. Springer, 2018. 7

[105] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*, 2019. 2

[106] 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 *Proceedings of ICLR*, 2019. 2, 5, 8

[107] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *CoRR*, abs/2108.10904, 2021. 2, 7, 8

[108] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *TACL*, 7:625–641, 2019. 7

[109] Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT*, 2018. 7

[110] Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *Proceedings of IJCV*, 119(1):3–22, 2016. 7

[111] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019. 5, 7

[112] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *Proceedings of ICML*, pages 10524–10533. PMLR, 2020. 5

[113] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 1:12, 2020. 2

[114] Zhenxun Yuan, Xiao Song, Lei Bai, Zhe Wang, and Wanli Ouyang. Temporal-channel transformer for 3d lidar-based video object detection for autonomous driving. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 2

[115] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of CVPR*, pages 5579–5588. Computer Vision Foundation / IEEE, 2021. 2

[116] Yu Zhang, James Qin, Daniel S Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V Le, and Yonghui Wu. Pushing the limits of semi-supervised learning for automatic speech recognition. *preprint arXiv:2010.10504*, 2020. 2

[117] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, pages 13041–13049, 2020. 2

[118] Yukun Zhu, Ryan Kiros, Rich 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 *Proceedings of CVPR*, pages 19–27, 2015. 4