FLAVA: A Foundational Language and Vision Alignment Model

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, Douwe Kiela

CVPR'24

Presenter: Xiangchi Yuan; Jingyun Xiao

Georgia Tech

Outline

- Problem Statement
- Related Works
- Approach
- Experiments & Results
- Limitations, Societal Implications
- Summary of Strengths, Weaknesses, Relationship to Other Papers

Problem Statement

Previous works:

- SOTA vision and VLMs rely on large-scale visio-linguistic pretraining.
- Such models are either cross-modal (contrastive) or multi-modal (with earlier fusion) but not both.
- They often only target specific modalities or tasks.

A promising direction:

- 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.

Related Work

Recent work either

(i) Focuses on a single target domain **ViLT, VinVL**;

(ii) Targets a specific unimodal domain along with the joint vision-and-language domain **ALIGN, CLIP**;

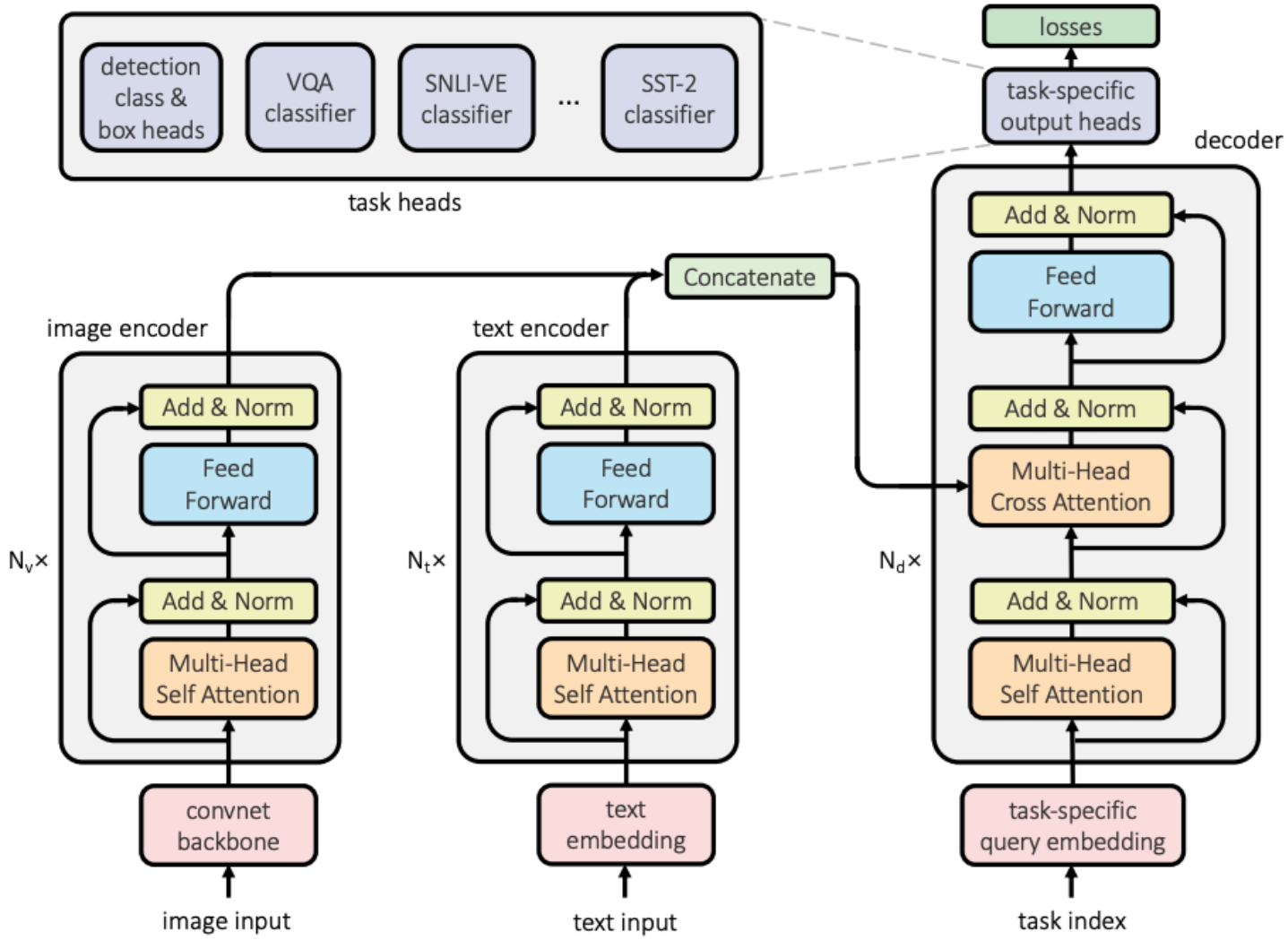
(iii) Targets all domains but only a specific set of tasks in a particular domain

Related Work

Method	Multimodal Pretraining data			Pretraining Objectives				Target Modalities			
	public	dataset(s)	size	Contr.	ITM	Masking	Unimodal	V	CV&L	MV&L	L
CLIP [83]	✗	WebImageText	400M	✓	–	–	–	✓	✓	–	–
ALIGN [50]	✗	JFT	1.8B	✓	–	–	–	✓	✓	–	–
SimVLM [109]	✗	JFT	1.8B	–	–	PrefixLM	CLM	*	✓	✓	✓
UniT [43]	–	None	–	–	–	–	–	*	–	✓	✓
VinVL [118]	✓	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 [109] and UniT [43] include ImageNet and object detection, respectively).

Unit [1]



[1] Unit: Multimodal multitask learning with a unified transformer, CVPR'21

Georgia Tech

Unit [1]

Key points:

- Unified Transformer arch
- Jointly learned diverse tasks simultaneously and converge
- Multimodal tasks and visual entailment benefit from multi-task training with **uni-modal tasks**

  -> important in FLAVA

Summary:

Unified transformer encoder-decoder architecture. (image encoder, text

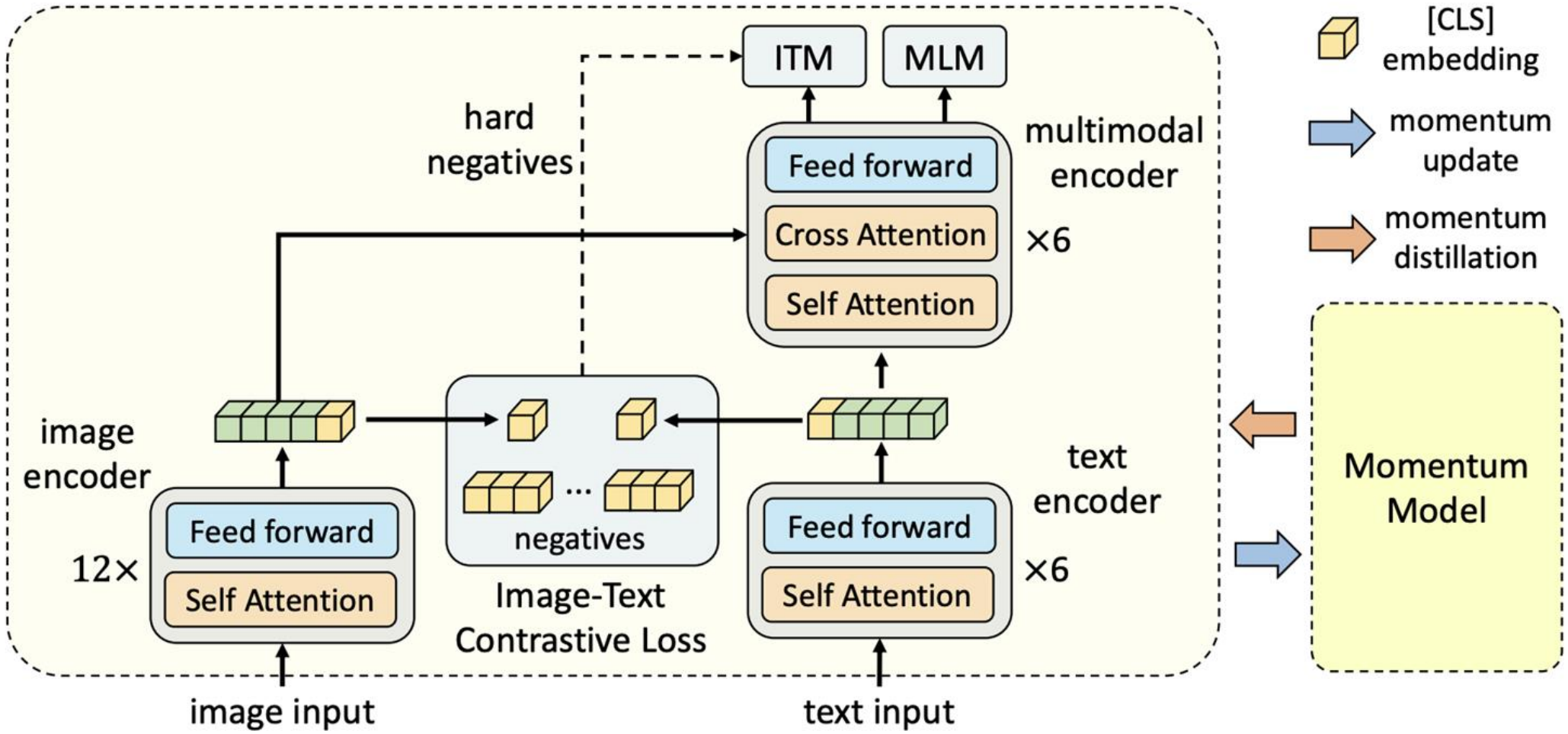
encoder, a joint decoder with per-task query embedding followed by

task-specific heads to make the final outputs for each task.

Major Cons: Image and text can't be well aligned

[1] Unit: Multimodal multitask learning with a unified transformer, CVPR'21

ALBEF [2]



[2] Align before Fuse: Vision and Language Representation Learning with Momentum Distillation, NeurIPS'21

ALBEF

Key points:

- “Align before Fuse”, first aligning visual and textual features before fusing them.
- Previous works fuse two modalities early -> suboptimal representations.
- Distillation refine the alignment process by teacher’s stable target representations -> learning from noisy web data

Summary:

Solve the problem:

Existing methods jointly model visual tokens (region-based image features) and word tokens.

Visual tokens and word tokens are unaligned -> hard to learn image-text interactions.

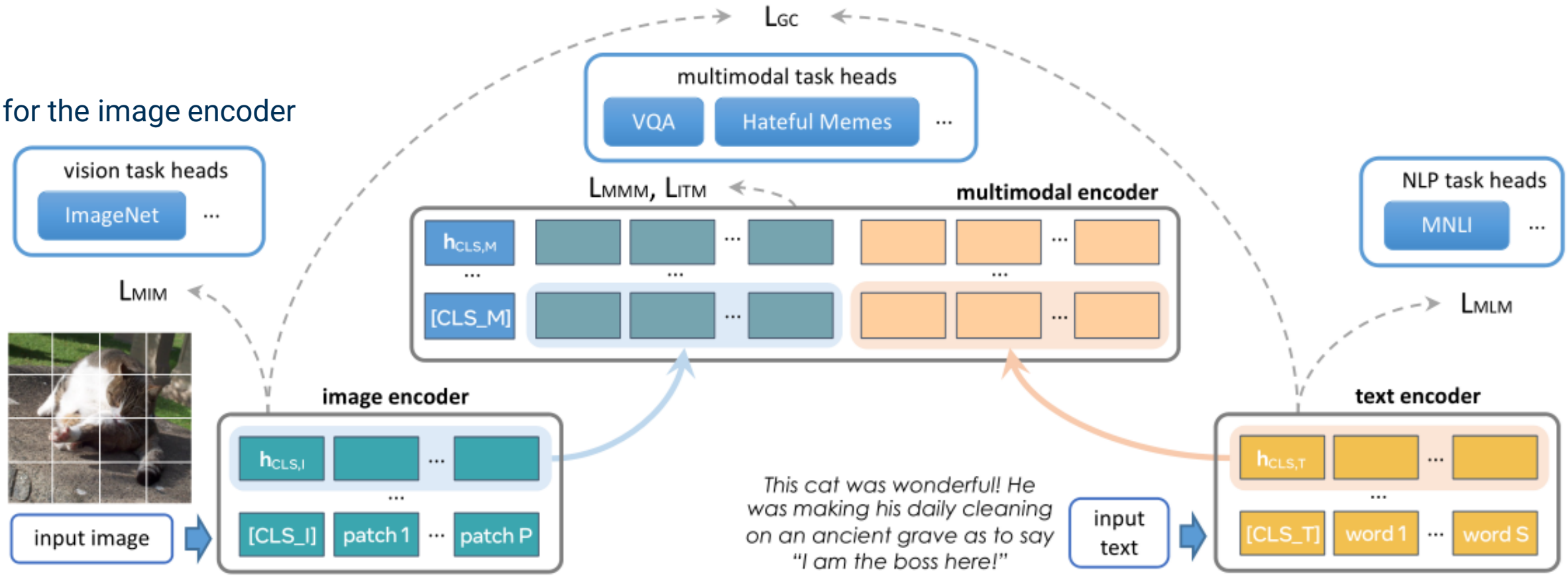
Major Cons:

Can’t do unimodal tasks.

Multimodal can’t benefit from unimodal training.

Approach

ViT for the image encoder



Georgia Tech

Architecture

- Image encoder

  ViT-B/16 architecture for the image encoder
- Text encoder

  same ViT architecture with visual encoder, i.e. ViT-B/16
- Multimodal encoder

  a separate **ViT** transformer to fuse the image & text hidden states with two learned linear projections
- Applying to downstream tasks

  The FLAVA model can be applied to both unimodal and multimodal tasks directly.

  A question: why they all use ViT as architecture even for text encoder?

CLS tokens: classification tokens

Image classification token [CLS I]

Text classification token [CLS T]

Additional token [CLS M]

These tokens help applying to downstream tasks:

The output (hidden state vectors) can be input of a classification head -> output of classification head solve corresponding tasks.

Training FLAVA once and evaluate separately.

Training objectives

Unimodel objective

- loss on image encoder
- loss on text encoder

Multimodal objective

Training Process

Multimodal pretraining objectives:

- Global contrastive (GC) loss image-text contrastive

  GC loss resembles that of CLIP maximize the cosine similarities and minimize.
- Masked multimodal modeling (MMM)

  Masks both the image patches and the text tokens and jointly works on both modalities.
- Image-text matching (ITM)

  Feed both matched and unmatched image-text pairs, then apply a classifier to decide if match or not.

Global Contrastive loss

- Large models are often trained using multiple GPUs data parallelism, where the samples in a batch are split across GPUs
- CLIP only back-propagates the gradients of the contrastive loss to the embeddings from the local GPU where the dot-product is performed
- Full backpropagation across GPUs compared to only doing backpropagation locally brings noticeable performance.

  “local contrastive” -> “global contrastive”

Training Process

Unimodal pretraining objectives:

- Masked image modeling (MIM)

  mask a set of image patches with the rectangular block-wise masking and reconstruct. MAE?
- Masked language modeling (MLM)

  A fraction of the text tokens are masked in the input, and reconstructed from the other tokens

Encoder initialization from unimodal pretraining:

pretrain the text encoder with the MLM objective on the unimodal text dataset.
pretraining the image encoder on unpaired image datasets (MIM or the DINO objective)
DINO better, than switch to MIM during images post-initialization.

Joint unimodal and multimodal training:

continue training the entire FLAVA model jointly on the three types
of datasets with round-robin sampling

DINO [1] objective

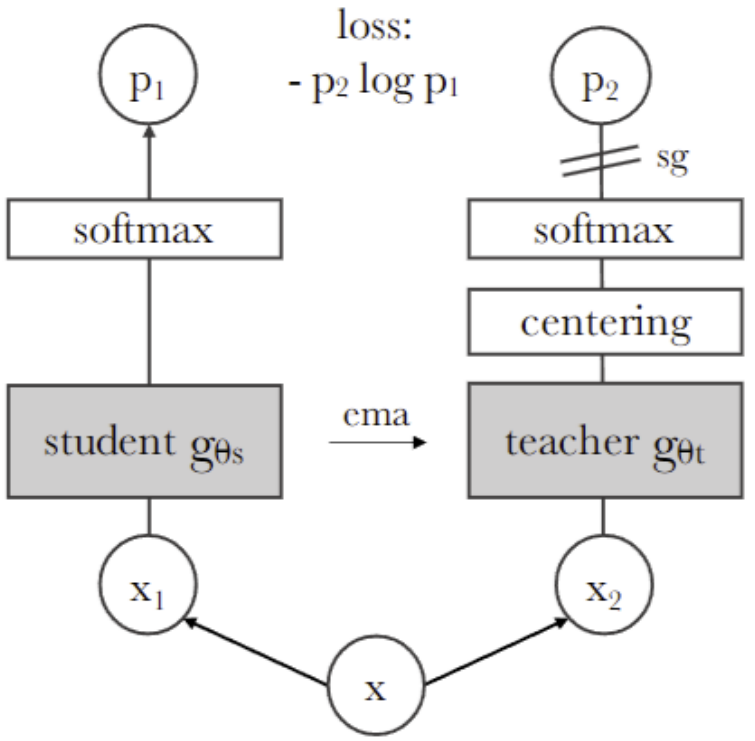
A self-supervised method, a form of self-distillation with no labels. DINO + ViTs = better performance

1. Model passes two different random transformations of an input image to the student and teacher networks

1. Their similarity is then measured with a cross-entropy loss.

1. Apply a stop-gradient (sg) operator on the teacher to propagate gradients only through the student.

1. The teacher parameters are updated with an exponential moving average (ema) of the student parameters



[1] Emerging properties in self-supervised vision transformers, In ICCV'21

dVAE also uses MMM and MIM

MMM and MIM in FLAVA both utilize pretrained **dVAE** tokenizer to tokenize image

dVAE:

- **Target**: Generating images from text descriptions without the need for domain-specific training on datasets like MS-COCO.
- **Input**: Transformers models both text and image tokens as a single sequence.
- **Results**: Can generate images from text input in a “zero-shot” manner, meaning the model can handle unseen categories that were not part of its training data.

Input for inference: Text description

Output: Generated image that matches text description.

Joint uni and multi-modal training

Step 1: unimodal pretraining of the image and text encoders

Step 2: training the entire FLAVA model jointly on the three types of datasets with round-robin sampling.

Round-robin sampling:

In each training iteration, choose one of the datasets according to a sampling ratio empirically obtain a batch of samples. Dataset type determine loss type: unimodal MIM on image data, unimodal MLM on text data, or the multimodal losses (contrastive, MMM, and ITM)

Selected Implementation details

Model:

optimizer hyperparameters preventing divergence with a large learning rate

- large batch size
- large weight decay
- long warm-up

Data:

	#Image-Text Pairs	Avg. text length
COCO [66]	0.9M	12.4
SBU Captions [77]	1.0M	12.1
Localized Narratives [82]	1.9M	13.8
Conceptual Captions [92]	3.1M	10.3
Visual Genome [57]	5.4M	5.1
Wikipedia Image Text [99]	4.8M	12.8
Conceptual Captions 12M [14]	11.0M	17.3
Red Caps [27]	11.6M	9.5
YFCC100M [103], 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.

Experiments and Results

Vision tasks

22 vision tasks

Language tasks

GLUE Benchmark: 8 tasks like sentiment analysis (SST-2), textual entailment (MNLI, RTE), and question answering (QNLI).

Multimodal tasks

VQAv2: interpret visual content and answer related questions

SNLI-VE: infer the correct relationship between the visual content of the image and the given textual hypothesis

Hateful Memes: detect nuanced and context-based hateful content in memes

Flickr30K: zero-shot retrieval tasks

COCO: zero-shot retrieval tasks

Experiments and Results

Ablation studies and insights：

Effective global contrastive loss in FLAVA
(Similar to CLIP)

MMM and ITM objectives benefit multimodal tasks

Joint unimodal & multimodal pretraining helps NLP

Better image and text encoders via unimodal pretraining

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**

FLAVA$_{C}$: only image-text contrastive loss

FLAVA$_{MM}$: only on multimodal data

FLAVA without unimodal initialization, full pretraining

FLAVA full pretraining

Experiments and Results

Comparison with other models :

	public data		Multimodal Tasks VQAv2	SNLI-VE	HM	Language Tasks CoLA	SST-2	RTE	MRPC	QQP	MNLI	QNLI	STS-B	ImageNet linear eval
1	✓	$\text{BERT}_{\text{base}}$ [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 [83]	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	✗	$\text{SimVLM}_{\text{base}}$ [109]	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	✓	$\text{UNITER}_{\text{base}}$ [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	✓	$\text{VL-BERT}_{\text{base}}$ [101]	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 [102]	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**

Limitations & Societal Implications

Limitations:

Bias in Training Data: datasets can contain inherent biases that the model may inadvertently learn. This can lead to biased or unfair outcomes when the model is deployed in real-world applications. Lack of

Complexity of Training Multiple Objectives: multiple objectives (like masked modeling, contrastive learning, and image-text matching), difficult to tune and optimize, hard to evaluate or balance the contributions of different objectives to achieve the best overall performance.

Limitations & Societal Implications

Societal Implications:

Access and Inclusivity: The high computational cost could limit access to research and development, especially in less resourced environments.

Bias and Fairness: As with many AI models, biases in training data can lead to biases in the model, impacting fairness and representation.

Data Privacy: FLAVA uses large amounts of data scraped from the internet, which can include personal information or content where the usage rights are ambiguous.

Summary of Strengths, Weaknesses, Relationships

Strengths

- 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. - FLAVA achieved this!

- Mainly benefitted from joint pretraining on both unimodal and multimodal data while encompassing cross-modal “alignment” objectives and multimodal “fusion” objectives. - Training process matters!

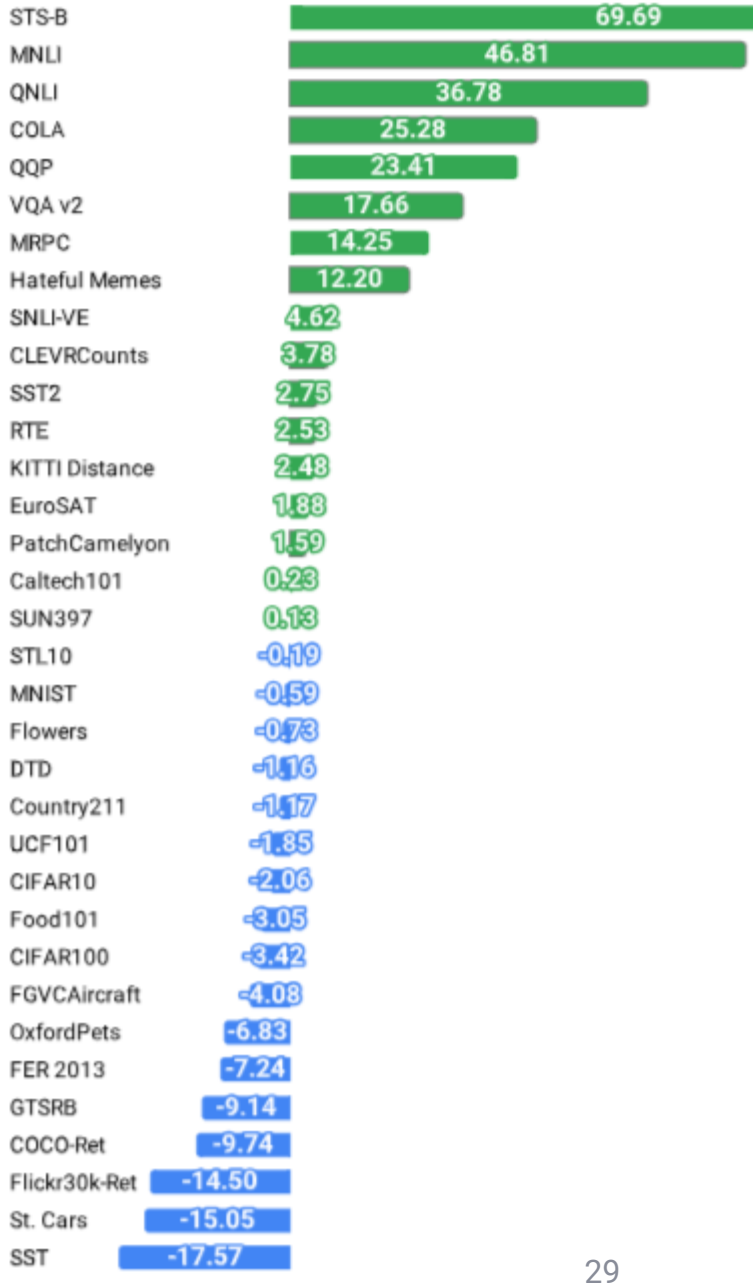
Weakness

Training Complexity: The model requires careful handling of different objectives (contrastive, masked modeling, image-text matching) across unimodal and multimodal data, which adds complexity to the training process.

Large Batch Size Requirement: The model training relies on very large batch sizes and a significant amount of computational resources, which might limit accessibility for some researchers.

Performance on Vision Tasks: While competitive, FLAVA's performance on certain vision-only tasks lags behind specialized models like CLIP, indicating that the model's broader focus on multiple modalities might compromise its ability to compete with models optimized for a single modality.

Potential reasons?



Summary of Strengths, Weaknesses, Relationships

Relationships to Other Works:

Comparison to CLIP and ALIGN: Unlike CLIP and ALIGN, which focus on large-scale image-text pair training, FLAVA incorporates these and extends beyond by integrating unpaired image and text data effectively.

Advancement Over Traditional Models: It improves upon traditional single-modality models by providing a unified approach that leverages cross-modal interactions, enhancing task performance.

Inspirations from Transformers: Incorporates and builds upon the transformer architecture, making significant adaptations for multimodal integration and alignment.