FLAVA

A Foundational Language And Vision Alignment Model

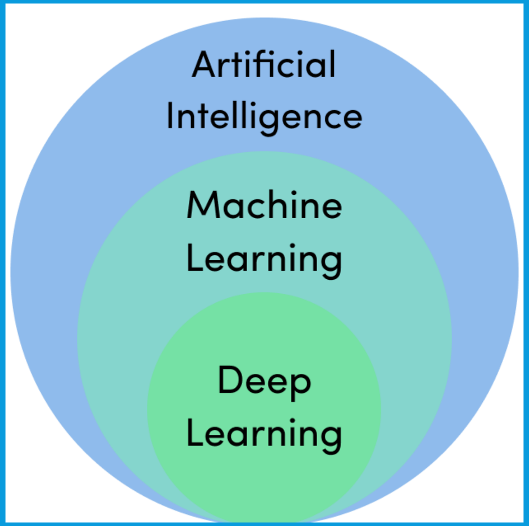
OUTLINE

1. Introduction
2. Flava Model Architecture
3. Other Vision and Language models
4. Pre-Training tasks
5. Results
6. Conclusion
7. References

INTRODUCTION

Meta AI

- Foundational Language And Vision Alignment Model (FLAVA)
- Released in 2021 by Facebook AI Research Team
- Language vision alignment model that learns strong representations from multimodal data (image-text pairs) and unimodal data (unpaired images and text



INTRODUCTION - TASKS

There are several categories of tasks which span across modalities:

Unimodal Vision Tasks (Image)

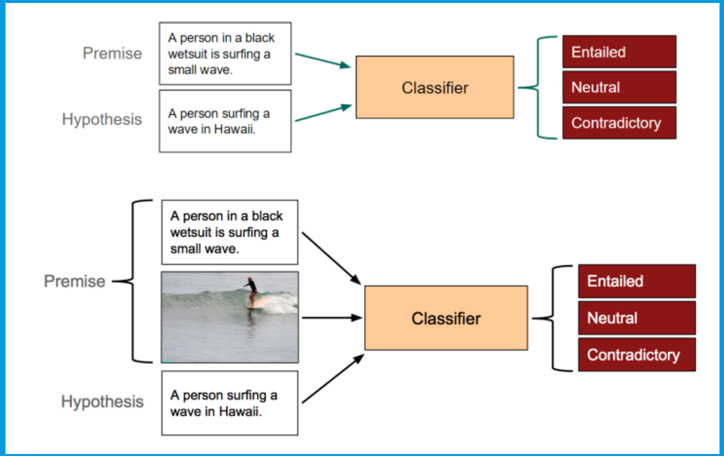
1. Object Identification/Facial Recognition
2. Image Segmentation
3. Image Classification

Unimodal Language Tasks (Tex)

1. Natural Language Inference
2. Sentiment Analysis
3. Named Entity Recognition

Multimodal Tasks (Image + Text)

1. Image-Text retrieval
2. Visual Question Answering (VQA)
3. Multimodal Natural Language Inference (MNLI)



Premise

A person in a black wetsuit is surfing a small wave.

Hypothesis

A person surfing a wave in Hawaii.

Classifier

Entailed

Neutral

Contradictory

Premise

A person in a black wetsuit is surfing a small wave.

Hypothesis

A person surfing a wave in Hawaii.

Classifier

Entailed

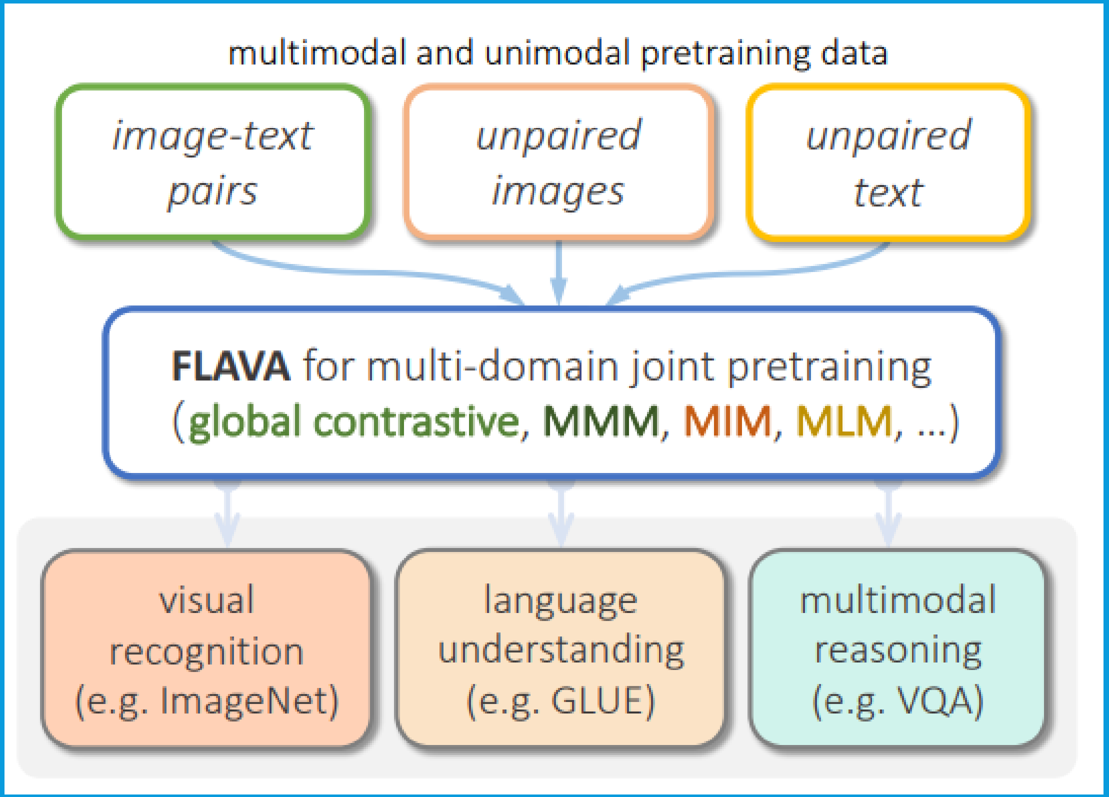
Neutral

Contradictory

Caption	A person in a black wetsuit is surfing a small wave.
Entailment	A person is surfing a wave.
Contradiction	A woman is trying to sleep on her bed.
Neutral	A person surfing a wave in Hawaii.

Figure: An example of the data given in the SNLI dataset, with the associated image from the Flickr30k dataset

INTRODUCTION - FLAVA


multimodal and unimodal pretraining data
image-text pairs
unpaired images
unpaired text
FLAVA for multi-domain joint pretraining
(global contrastive, MMM, MIM, MLM, ...)
visual recognition (e.g. ImageNet)
language understanding (e.g. GLUE)
multimodal reasoning (e.g. VQA)

- State-of-the-art 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.
- FLAVA targets **all modalities at once** – a true vision and language foundation model good at vision tasks, language tasks, and cross- and multi-modal vision and language tasks.
- FLAVA demonstrates impressive performance on a wide range of 35 tasks spanning these target modalities (visual recognition, language understanding, and multimodal reasoning).
- FLAVA uses a Transformer Architecture

FLAVA MODEL

The FLAVA model core is composed by:

- an image encoder transformer to capture unimodal image representations
- a text encoder transformer to process unimodal text information
- a multimodal encoder transformer that takes as input the encoded unimodal image and text and integrates their representations for multimodal reasoning

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.

FLAVA MODEL

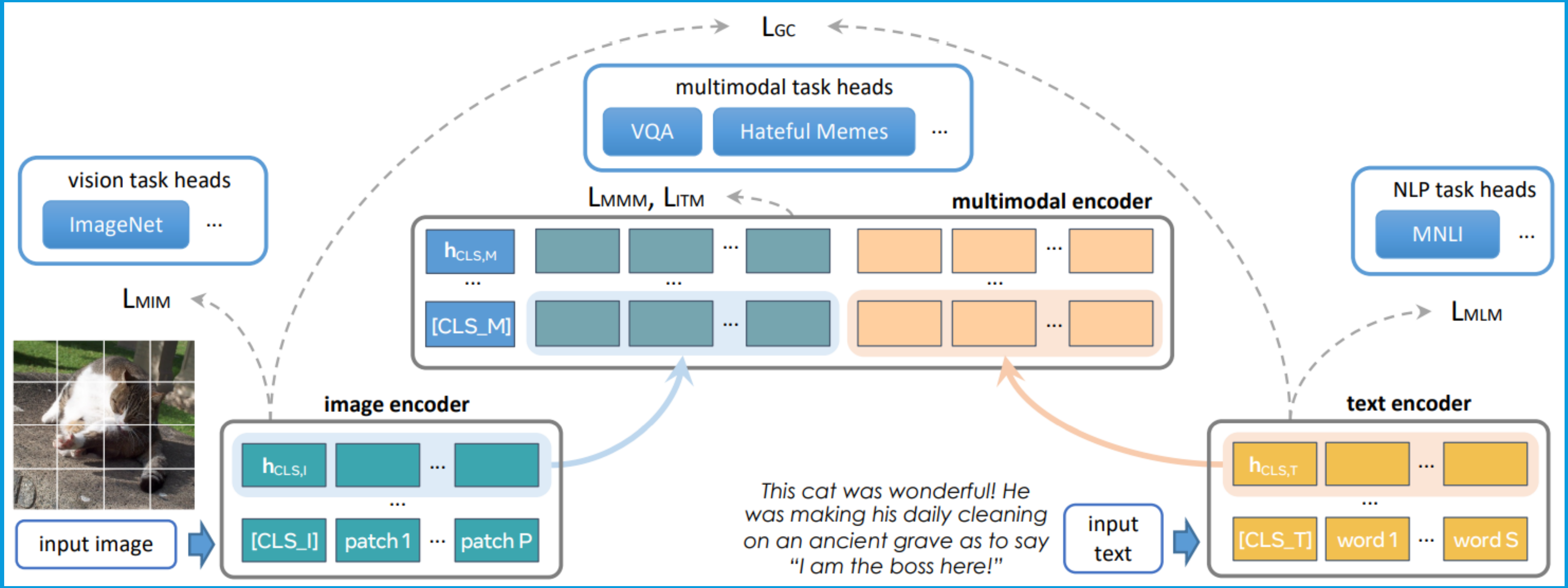


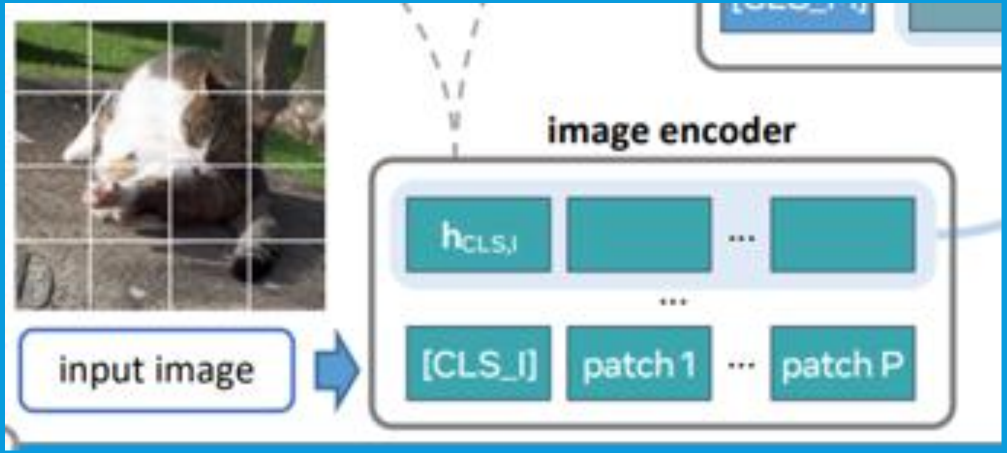
Figure: Overview of the FLAVA model

FLAVA MODEL - IMAGE ENCODER

Thee ViT-B/16 architecture is adopted for the image encoder.

Given an input image:

1. the image is resized to a fixed size
2. the image is split into patches
3. The patches are linearly embedded and fed into a transformer model (along with positional embeddings and an extra image classification token$[CLS_I]$).
4. The image encoder output is a list of image hidden state vectors $\{h_I\}$, each corresponding to an image patch, plus an additional $h_{CLS,I}$ for $[CLS\_I]$.

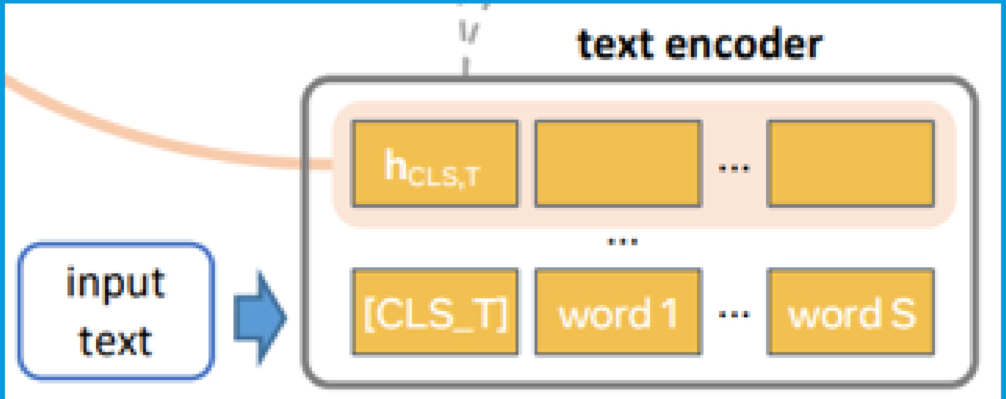


FLAVA MODEL - TEXT ENCODER

Given an input piece of text (e.g., a sentence or a pair of sentences):

1. The sentence is first tokenized
2. The sentence is embedded it into a list of word vectors.
3. Then, a transformer model is applied over the word vectors to encode them into a list of hidden state vectors $\{h_t\}$, including $h_{CLS,T}$ for the text classification $[CLS\_T]$ token.

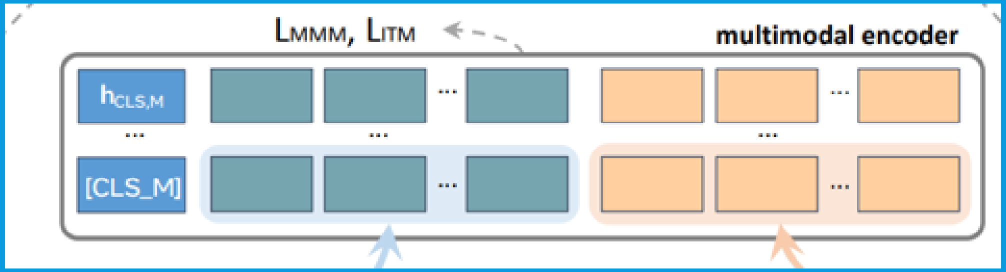
Importantly, different from prior work, the text encoder has exactly the same architecture as the visual encoder). The same ViT architecture is used (but with different parameters) for both the visual and textual encoder i.e. ViT-B/16.



FLAVA MODEL - MULTIMODAL ENCODER

A separate transformer is used to fuse the image and text hidden states:

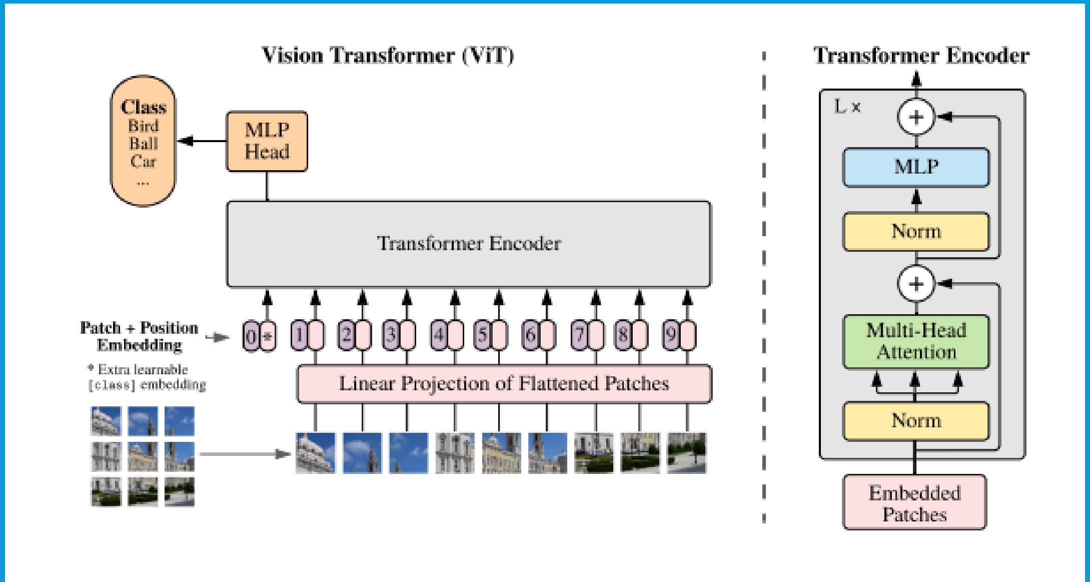
1. Two learned linear projections are applied over each hidden state vector in $\{h_I\}$ and $\{h_T\}$.
2. These projections are concatenated into a single list with an additional $[CLS\_M]$ token added.
3. This concatenated list is fed into the multimodal encoder transformer (also based on the ViT architecture), allowing cross-attention between the projected unimodal image and text representations and fusing the two modalities.
4. 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 $h_{CLS,M}$ for $[CLS\_M]$).



VISION TRANSFORMER

Model overview:

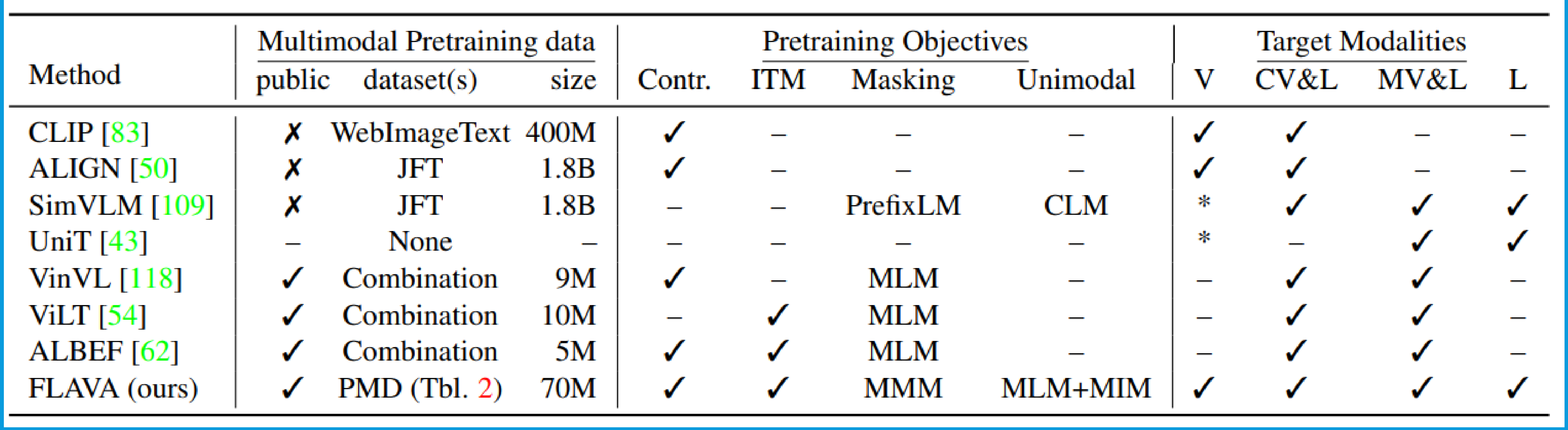
1. Split an image into fixed-size patches, linearly embed each of them and add position embeddings
2. feed the resulting sequence of vectors to a standard Transformer encoder
3. In order to perform classification, the standard approach of adding an extra learnable “classification token” to the sequence is used



TRANSFORMER VS CNN

Inductive bias

OTHER MODELS



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	✓	✓	✓	✓

Figure: Comparing FLAVA with previous models on multimodal tasks, language tasks, and ImageNet linear evaluation

META-ANALYSIS AND A UNIFIED FRAMEWORK OF VISION-AND-LANGUAGE BERTS

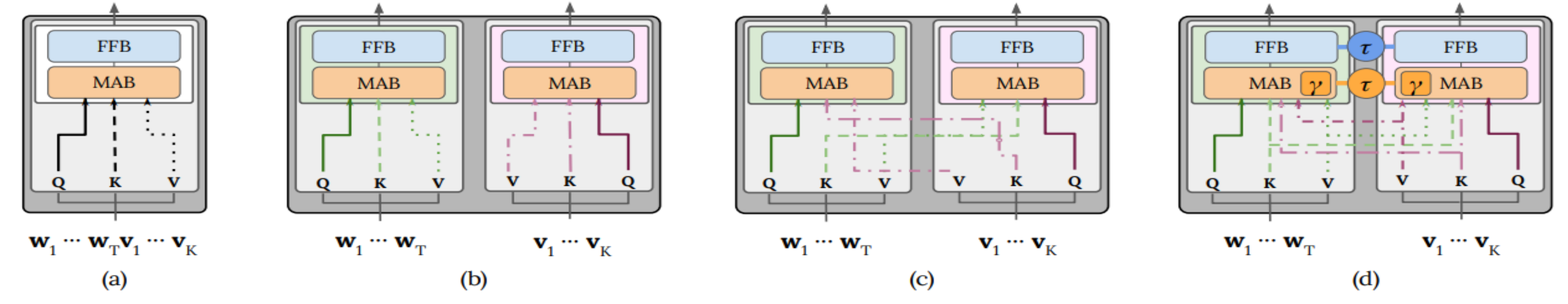

FFB
MAB
Q
K
V
$\mathbf{w}_1 \cdots \mathbf{w}_T \mathbf{v}_1 \cdots \mathbf{v}_K$
(a)
$\mathbf{w}_1 \cdots \mathbf{w}_T$
$\mathbf{v}_1 \cdots \mathbf{v}_K$
(b)
(c)
τ
γ
(d)

Figure: Visualisation of the (a) single-stream, (b) dual-stream intra-modal and (c) dual-stream intermodal Transformer layers. (d) shows our gated bimodal layer. The inter-modal layer attends across modalities, while the intra-model layer attends within each modality. Gated bimodal can attend to either or both.

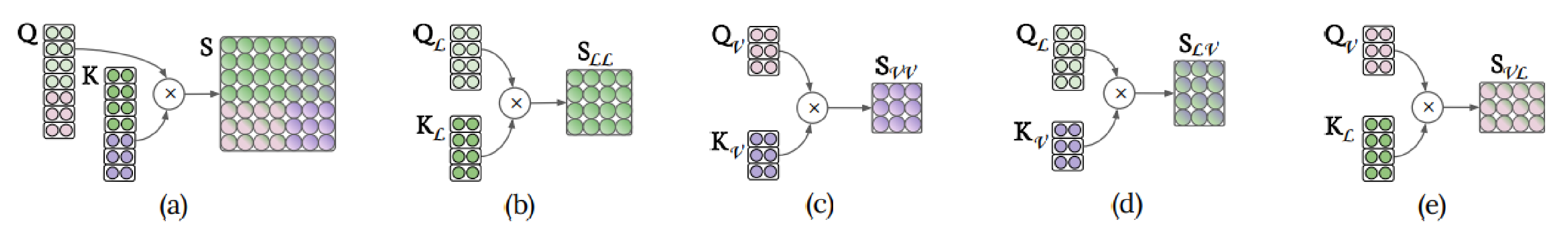

Q
K
×
S
(a)
$\mathbf{Q}_{\mathcal{L}}$
$\mathbf{K}_{\mathcal{L}}$
$\mathbf{S}_{\mathcal{LL}}$
(b)
$\mathbf{Q}_{\mathcal{V}}$
$\mathbf{K}_{\mathcal{V}}$
$\mathbf{S}_{\mathcal{VV}}$
(c)
$\mathbf{Q}_{\mathcal{L}}$
$\mathbf{K}_{\mathcal{V}}$
$\mathbf{S}_{\mathcal{LV}}$
(d)
$\mathbf{Q}_{\mathcal{V}}$
$\mathbf{K}_{\mathcal{L}}$
$\mathbf{S}_{\mathcal{VL}}$
(e)

Figure : Visualisation of the score matrix for (a) single-stream, (b) text–text, (c) vision–vision, (d) text–vision, and (e) vision–text interactions. Shades of green denote the text modality, while purple ones denote the vision modality. Dual-stream scores are sub-matrices of the single-stream scores matrix.

BACKGROUND – PRETRAINING/FINE TUNING

Pretraining/Fine-tuning work as follows:

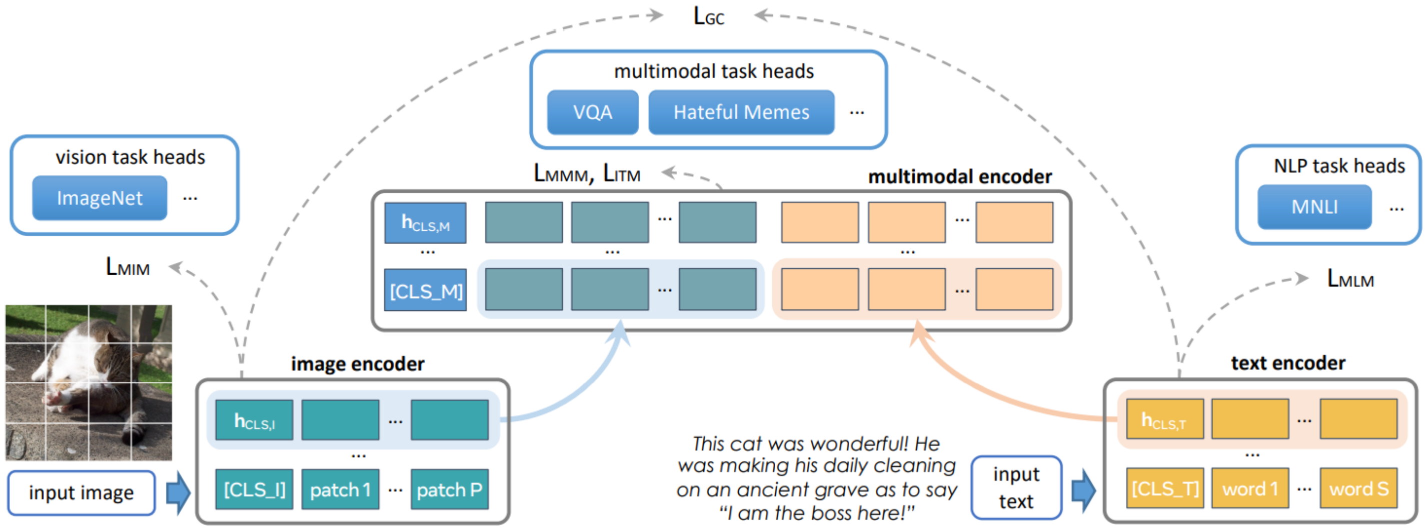
1. You have a machine learning model ***m***
2. **Pre-training:** You have a dataset ***A*** on which you train ***m***
3. You have a dataset ***B***. Before you start training the model, you initialize some of the parameters of ***m*** with the model which is trained on ***A***.
4. **Fine-tuning:** You train ***m*** on ***B***.

This is one form of transfer learning, so you can transfer some of the knowledge obtained on dataset ***A*** to dataset ***B***.

FLAVA MODEL PRETRAINING

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
- contrastive, masked multimodal modeling (MMM), and image-text matching (ITM) loss are used over paired image-text data.



PRETRAINING OBJECTIVES

Pretraining obejectives aim to obtain strong representations through pretraining on both multimodal data (paired image and text) as well as unimodal data (unpaired images or text).

1. Unimodal pretraining objectives:

- Masked Image Modeling (MIM)
- Masked Language Modeling (MLM)
- Encoder initialization from unimodal pretraining

2. Multimodal pretraining objectives:

- Global Contrastive Loss (GC Loss)
- Masked multimodal modeling (MMM)
- Image-texto matching (ITM)

3. Joint unimodal and multimodal training

UNIMODAL OBJECTIVES - MASKED IMAGE MODELING (MIM)

"On unimodal image datasets, we mask a set of image patches following the rectangular block-wise masking in BEiT and reconstruct them from other image patches.

The input image is first tokenized using a pretrained dVAE tokenizer and then a classifier is applied on the image encoder outputs $\{h_I\}$ *to predict the dVAE tokens of the masked patches."*

UNIMODAL OBJECTIVES - MASKED LANGUAGE MODELING (MLM)

"We apply a masked language modeling loss 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 $\{h_T\}$*."*

UNIMODAL OBJECTIVES - ENCODER INITIALIZATION FROM UNIMODAL PRETRAINING

"We use three sources of data for pretraining: unimodal image data (ImageNet-1K), unimodal text data (CCNews and BookCorpus), and multimodal image-text paired data. 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:

1. *We pretrain the image encoder on unpaired image datasets with either MIM or the DINO objective, before joint training on both unimodal and multimodal datasets.*
2. *We empirically found the latter to work quite well, despite the switch to an MIM objective on images post-initialization (more details in supplemental).*
3. *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. "*

MULTIMODAL OBJECTIVES - GLOBAL CONTRASTIVE (GC) LOSS

"Our image-text contrastive loss resembles that of CLIP.

1. *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.*

2. *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 only back-propagates the gradients of the contrastive loss to the embeddings from the local GPU where the dot-product is performed.*

3. *In contrast, through experiments that can be found in the supplemental, we observe a noticeable performance gain by performing full backpropagation across GPUs compared to only doing backpropagation locally. We call our loss "global contrastive" LGC to distinguish it from "local contrastive" approaches."*

MULTIMODAL OBJECTIVES - 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 LMMM that masks both the image patches and the text tokens and jointly works on both modalities.

1. *Specifically, given an image and text input, we first tokenize the input image patches using a pretrained dVAE tokenizer, which maps each image patch into an index in a visual codebook similar to a word dictionary*
2. *Then, we replace a subset of image patches based on rectangular block image regions following BEiT and 15% of text tokens following BERT with a special [MASK] token.*
3. *Then, from the multimodal encoder's output {hM}, 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. "*

MULTIMODAL OBJECTIVES - IMAGE-TEXT MATCHING (ITM)

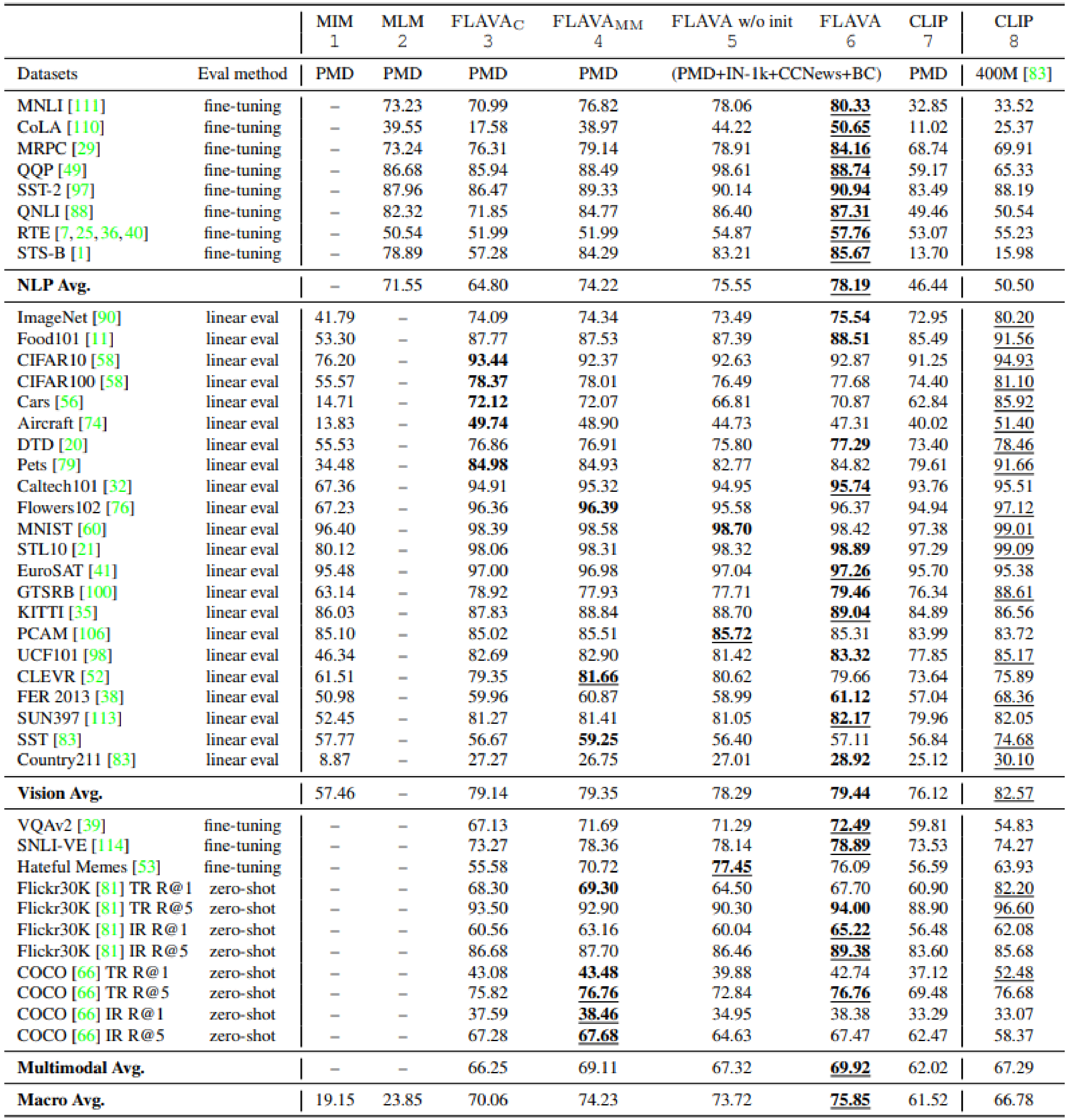
"Finally, we add an image-text matching loss LITM following prior vision-and-language pretraining literature. During pretraining, we feed a batch of samples including both matched and unmatched image-text pairs. Then, on top of $h_{CLS,M}$ from the multimodal encoder, we apply a classifier to decide if an input image and text match each other."

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 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) on image-text pairs."

FLAVA RESULTS

Table: Comparing full FLAVA pretraining with other settings, where FLAVA gets the highest macro average score.



		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 [83]
MNLI [111]	fine-tuning	–	73.23	70.99	76.82	78.06	**80.33**	32.85	33.52
CoLA [110]	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 [97]	fine-tuning	–	87.96	86.47	89.33	90.14	**90.94**	83.49	88.19
QNLI [88]	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 [90]	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 [100]	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 [106]	linear eval	85.10	–	85.02	85.51	**85.72**	85.31	83.99	83.72
UCF101 [98]	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 [113]	linear eval	52.45	–	81.27	81.41	81.05	**82.17**	79.96	82.05
SST [83]	linear eval	57.77	–	56.67	**59.25**	56.40	57.11	56.84	74.68
Country211 [83]	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 [114]	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 [81] TR R@1	zero-shot	–	–	68.30	**69.30**	64.50	67.70	60.90	82.20
Flickr30K [81] TR R@5	zero-shot	–	–	93.50	92.90	90.30	**94.00**	88.90	96.60
Flickr30K [81] IR R@1	zero-shot	–	–	60.56	63.16	60.04	**65.22**	56.48	62.08
Flickr30K [81] 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

FLAVA RESULTS

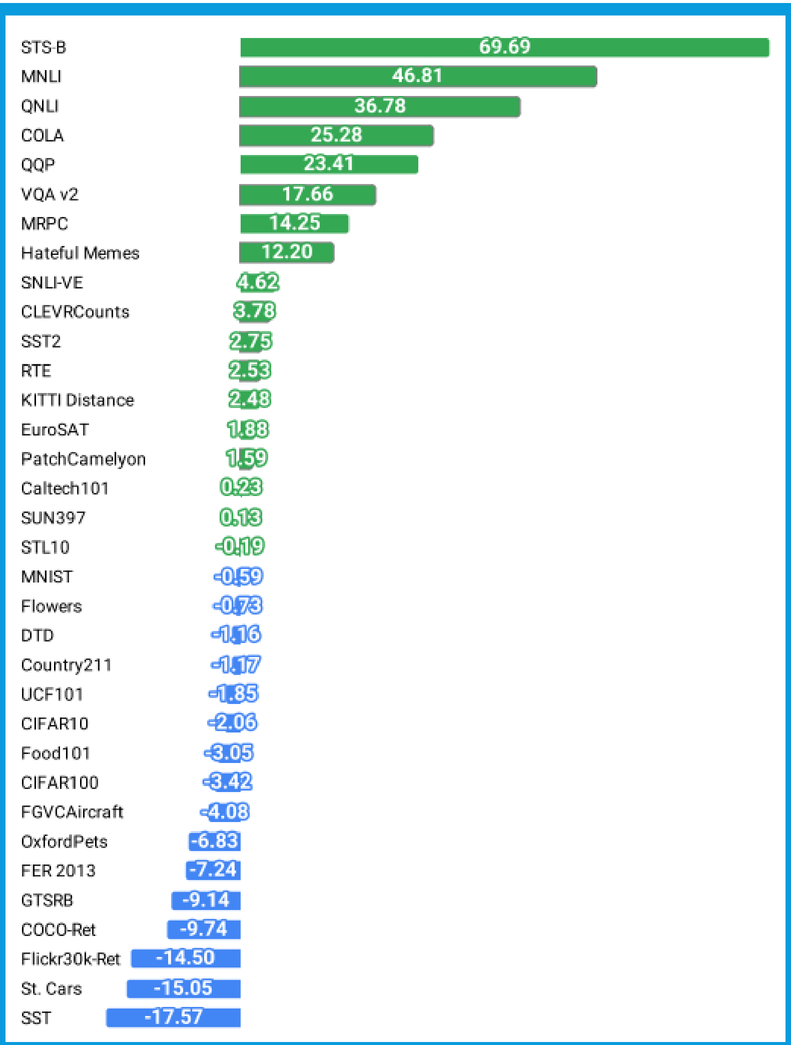


Figure: The performance difference between FLAVA and CLIP-ViT-B/16 (400M) on vision, language and multimodal tasks (positive means FLAVA is better).

CONCLUSION

FLAVA is a new model

Perhaps

REFERENCES

I. Singh, Amanpreet, et al. "FLAVA: A Foundational Language And Vision Alignment Model." *arXiv preprint arXiv:2112.04482* (2021).