Google Research

Three Towers: Flexible Contrastive Learning with Pretrained Image Models

Jannik Kossen*, Mark Collier*, Basil Mustafa, Xiao Wang, Xiaohua Zhai, Lucas Beyer, Andreas Steiner, Jesse Berent, Rodolphe Jenatton, Efi Kokiopoulou

OATML Oxford, Google DeepMind, Google Research

Contact: jannik.kossen@cs.ox.ac.uk & markcollier@google.com

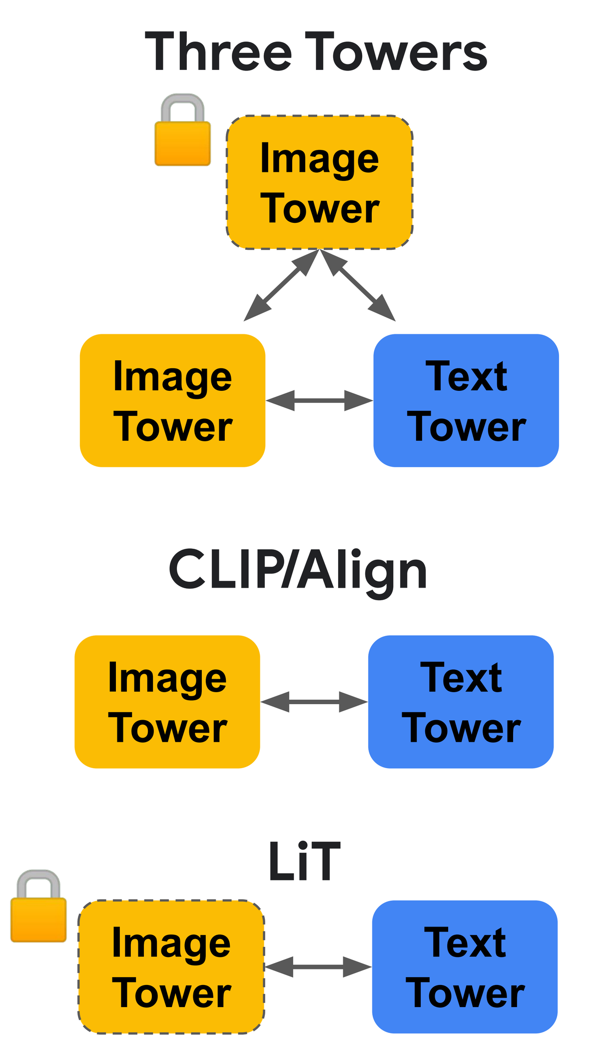
arXiv:2305.16999

Summary

- **Contrastive vision-language models** are usually trained **from scratch.**
- LiT (Zhai et al., 2022) has shown performance gains by replacing the learned image tower with **frozen embeddings** from a **pretrained classifier**.
- With **Three Towers (3T)** we introduce a **third tower** that **contains** the **frozen pretrained embeddings**.
- We **encourage alignment** between the third tower and the main image-text towers with additional losses.
- This is a more **flexible strategy** that allows the **image tower** to **benefit from** both **pretrained embeddings** and **contrastive training.**
- For **retrieval tasks**, 3T **consistently improves over LiT and** the **CLIP**-style from-scratch baseline.
- For **classification tasks**, 3T **reliably improves over CLIP** and while it underperforms relative to LiT for JFT-pretrained models, it outperforms LiT for ImageNet-21k and Places365 pretraining.

Method

- The pretrained model is **locked** in a third tower.
- Objective = Average of **contrastive learning objectives between all towers.**
- The **main** image and text **towers** are **unlocked**. ⟶ They benefit from both **contrastive learning** and the **pretrained model**.
- The third tower is usually **discarded at test time** ⟶ No additional inference costs.
- The contrastive objectives to the third tower effectively perform **transfer learning**.

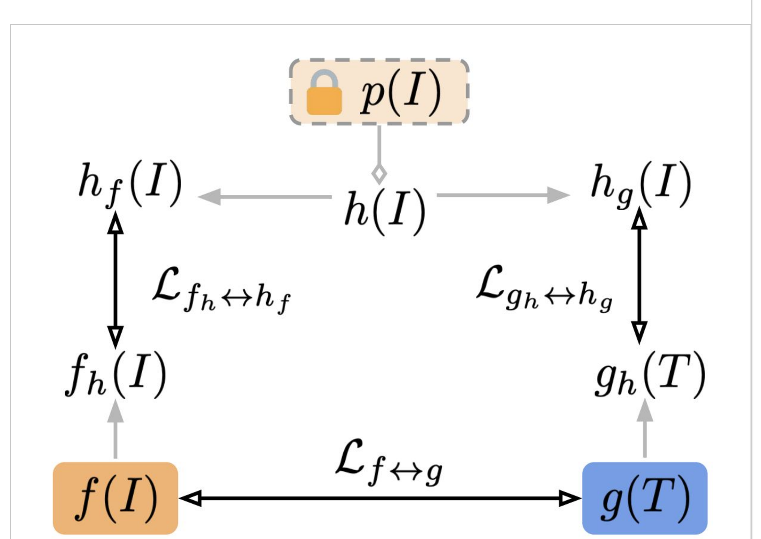


$$\mathcal{L}_{3T} = \frac{1}{3} \cdot (\mathcal{L}_{f\leftrightarrow g} + \mathcal{L}_{f_h\leftrightarrow h_f} + \mathcal{L}_{g_h\leftrightarrow h_g})$$

$$\mathcal{L}_{f\leftrightarrow g} = \frac{1}{2}(\mathcal{L}_{f\to g} + \mathcal{L}_{g\to f})$$

$$\mathcal{L}_{f\to g} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(f(I_i)^\top g(T_i)/\tau)}{\sum_{j=1}^{N}\exp(f(I_i)^\top g(T_j)/\tau)}$$

$$\mathcal{L}_{g\to f} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(f(I_i)^\top g(T_i)/\tau)}{\sum_{j=1}^{N}\exp(f(I_j)^\top g(T_i)/\tau)}$$



Results

Retrieval

g scale – JFT pre-training

Method	Basel.	LiT	3T
Flickr img2txt	85.0	83.9	87.3
Flickr txt2img	67.0	66.5	72.1
COCO img2txt	60.0	59.5	64.1
COCO txt2img	44.7	43.6	48.5

L scale

Pretraining	–	IN-21k		JFT	
Method	Basel.	LiT	3T	LiT	3T
Flickr* img2txt	75.6	71.7	80.0	78.7	80.0
Flickr* txt2img	57.1	49.3	60.9	58.8	61.4
COCO img2txt	51.0	46.1	54.4	52.7	54.4
COCO txt2img	34.2	27.8	37.7	36.7	37.9

Robustness

L scale

Setup	Mismatched			Places365		
Method	Basel.	LiT	3T	Basel.	LiT	3T
IN-1k	69.5	69.5	71.5	45.6	24.5	47.4
CIFAR-100	73.5	78.6	75.6	48.3	27.4	52.4
Pets	84.2	84.7	87.4	61.5	30.3	60.2
...		...			...	
Full Average	66.4	61.7	69.8	47.5	29.4	49.3

Classification

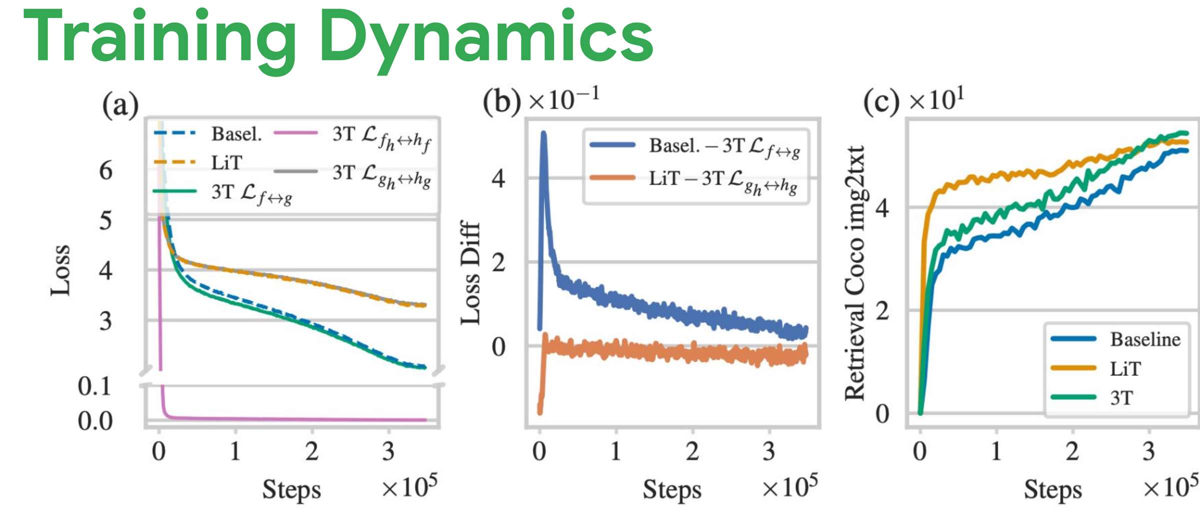
L scale – IN-21k pre-training

	Method	Basel.	LiT	3T
Few-Shot Classification	IN-1k	62.8	79.0	68.0
	CIFAR-100	70.4	83.6	72.5
	Caltech	91.0	88.4	92.3
	Pets	85.9	89.2	86.5
	DTD	70.3	69.2	73.3
	UC Merced	91.8	92.8	94.0
	Cars	81.5	41.9	84.9
	Col-Hist	71.7	86.4	76.6
	Birds	53.4	83.4	65.0
Zero-Shot Classification	IN-1k	69.5	76.0	71.7
	CIFAR-100	73.5	82.9	73.4
	Caltech	81.9	82.4	84.1
	Pets	84.2	87.1	87.0
	DTD	58.6	51.8	60.3
	IN-C	49.6	62.0	51.8
	IN-A	53.0	45.6	54.3
	IN-R	85.8	66.1	88.1
	IN-v2	62.2	67.2	64.9
	ObjectNet	56.2	41.9	58.3
	EuroSat	32.7	27.6	42.8
	Flowers	62.0	72.6	65.7
	RESISC	58.0	29.0	57.9
	Sun397	67.6	65.4	68.7
Average		68.4	68.3	71.4

g scale – JFT pre-training

	Method	Basel.	LiT	3T
Few-Shot Classification	IN-1k	62.8	81.3	67.7
	CIFAR-100	70.4	83.2	74.3
	Caltech	91.0	89.0	91.8
	Pets	85.9	96.8	88.4
	DTD	70.3	72.1	72.4
	UC Merced	91.8	95.5	93.1
	Cars	81.5	92.9	87.1
	Col-Hist	71.7	81.3	77.0
	Birds	53.4	85.6	62.4
Zero-Shot Classification	IN-1k	69.5	80.1	72.0
	CIFAR-100	73.5	80.1	75.2
	Caltech	81.9	79.5	82.5
	Pets	84.2	96.3	88.7
	DTD	58.6	59.0	59.0
	IN-C	49.6	68.1	52.8
	IN-A	53.0	69.1	56.4
	IN-R	85.8	91.7	88.4
	IN-v2	62.2	74.0	65.4
	ObjectNet	56.2	61.9	59.3
	EuroSat	32.7	36.6	54.7
	Flowers	62.0	76.7	66.6
	RESISC	58.0	58.9	60.9
	Sun397	67.6	69.7	68.1
Average		68.4	77.4	72.4

Training Dynamics

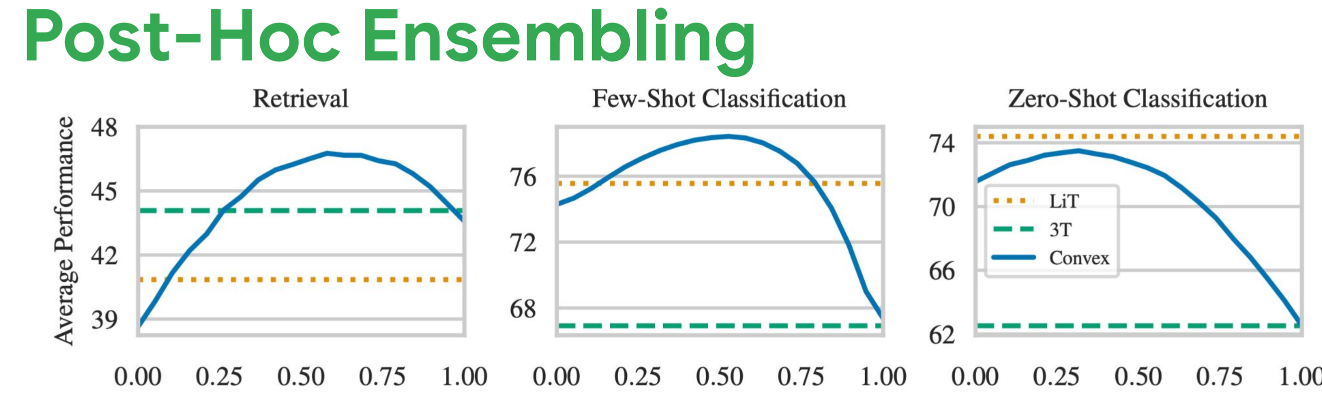


Ablation

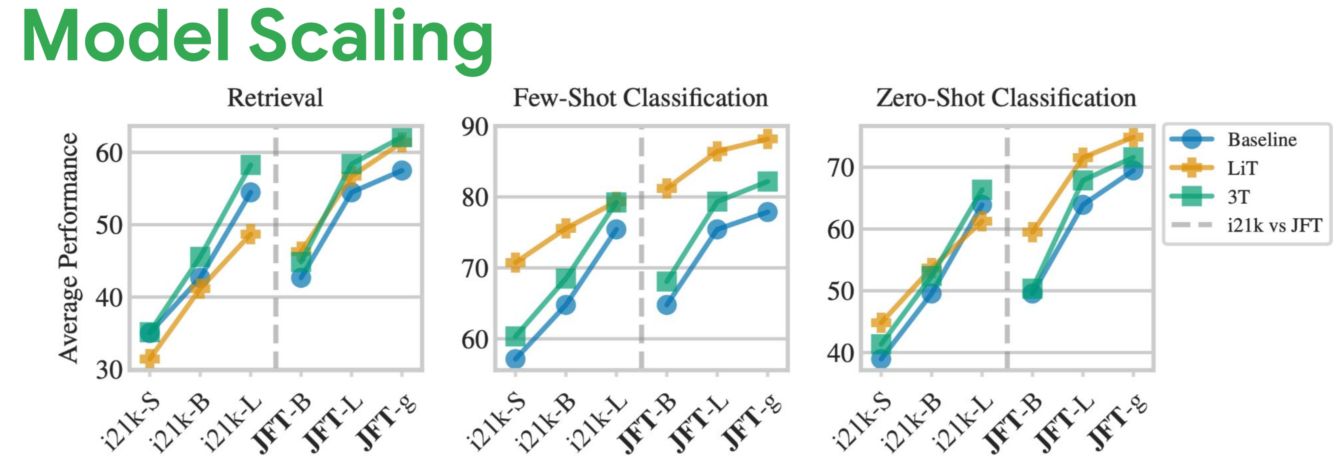
	Difference to 3T
Rerun	-0.22 ± 0.25
No $\mathcal{L}_{f\leftrightarrow g}$ Loss	-26.63 ± 10.61
No $\mathcal{L}_{f_h\leftrightarrow h_f}$ Loss	-1.19 ± 0.75
No $\mathcal{L}_{g_h\leftrightarrow h_g}$ Loss	-2.77 ± 0.91
Head Variants	0.09 ± 0.35
MLP Embedding	-0.08 ± 0.35
More Temperatures	-0.26 ± 0.48
Loss Weights	0.17 ± 0.53
L2 Transfer	-3.80 ± 1.13
3T Finetuning	1.85 ± 1.27

	Difference to LiT
Rerun	-0.10 ± 0.22
LiT Finetune	-14.99 ± 6.09
FlexiLiT1	-4.63 ± 1.36
FlexiLiT2	-5.04 ± 1.54

Post-Hoc Ensembling



Model Scaling



Data Scaling