Masked Autoencoding Does Not Help Natural Language Supervision At Scale

Floris Weers, Vaishaal Shankar, Angelos Katharopoulos, Yinfei Yang, Tom Gunter

THU-PM-270

Summary

Contributions

- A baseline that combines masked auto-encoders (MAE) and contrastive language-image pre-training (CLIP): MAE-CLIP
- We study the performance of MAE, M3AE, CLIP and MAE-CLIP in both a “low-sample” (11.3M) and a “high-sample” (1.4B) regime
- We analyze whether the addition of MAE improves visual grounding: the ability to localize objects in images

Summary

Conclusions

- MAE-CLIP provides a benefit over CLIP alone for relatively small training datasets (e.g. CC12M)
- CLIP outperforms MAE-CLIP when training on a large dataset of 1.4B image-text pairs
- Although the addition of MAE does slightly improve visual grounding, changing pooling operator has a much larger effect

Summary

Related Work

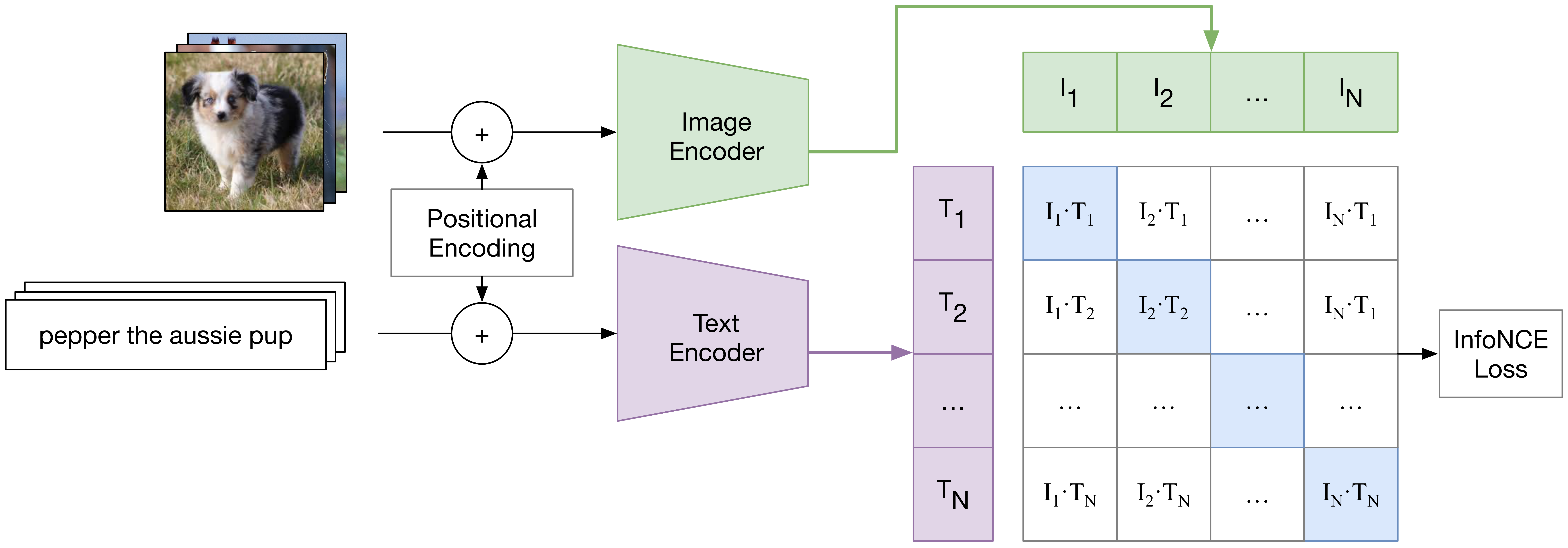
1. When Does Contrastive Visual Representation Learning Work?
2. Transfer Learning or Self-supervised Learning? A Tale of Two Pretraining Paradigms
3. Scaling and Benchmarking Self-Supervised Visual Representation Learning

- We explore the benefits of incorporating within-modality SSL in addition to natural language supervision
- They consider different ‘large’ vs ‘small’ scale data regimes

Does a combination of self supervision and natural language supervision actually lead to higher quality visual representations?

Background

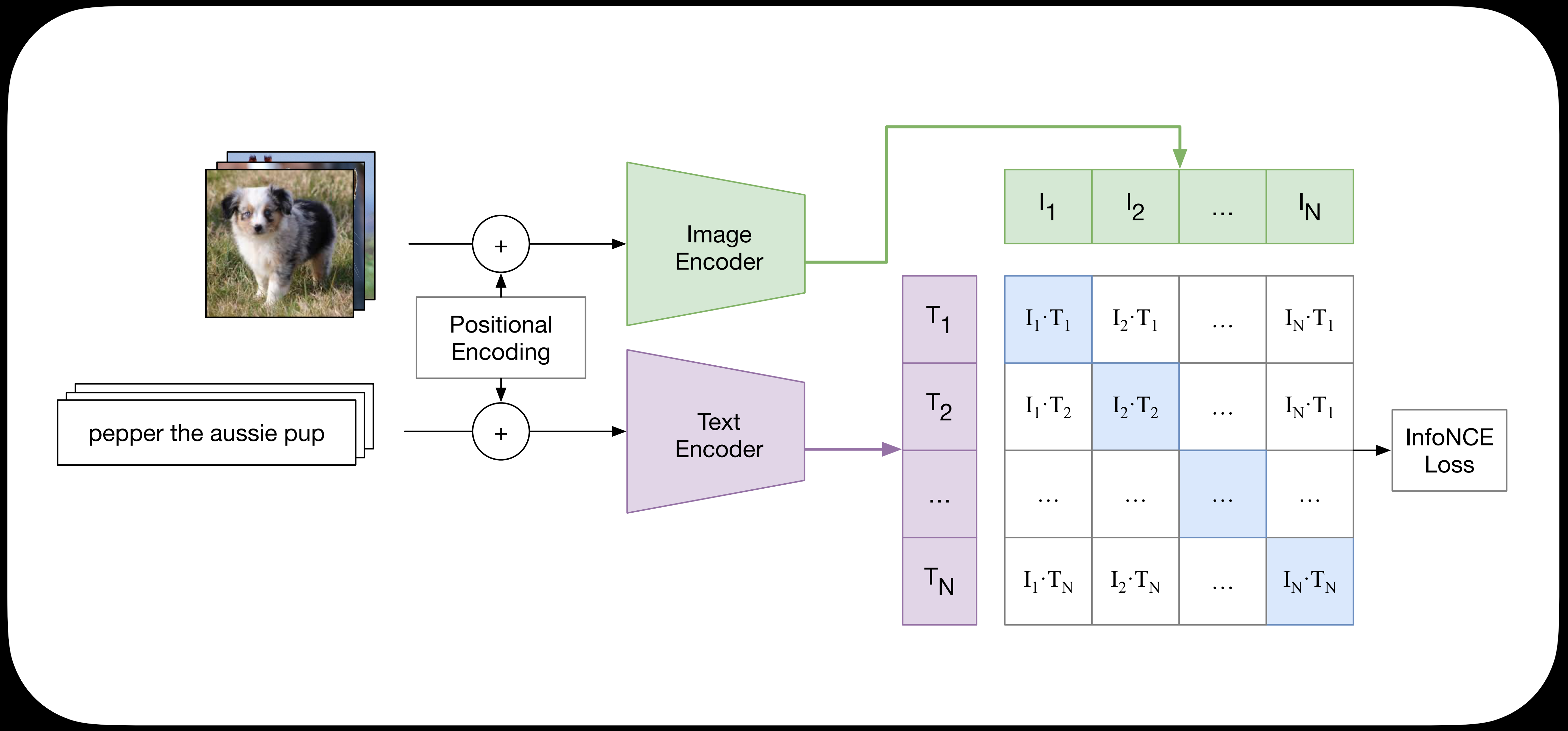
CLIP



Background

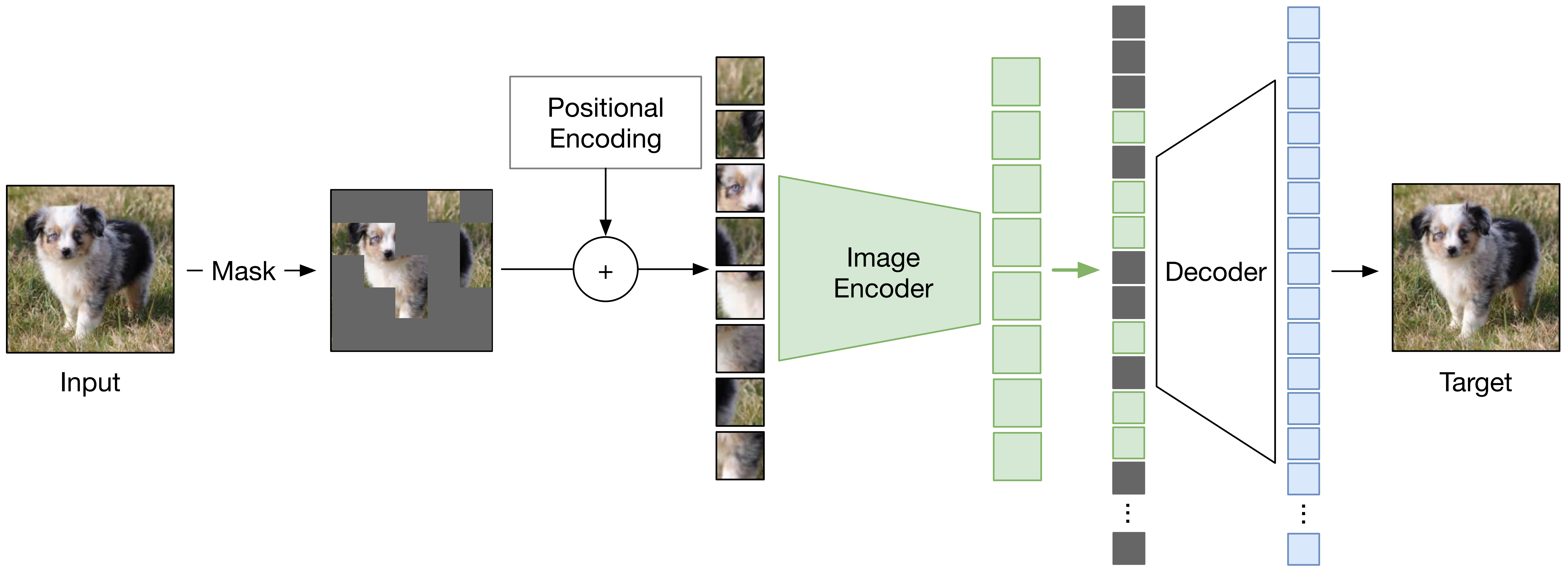
CLIP

- Task: contrastive
- Low visual grounding
    - Whole image, whole text matching



Background

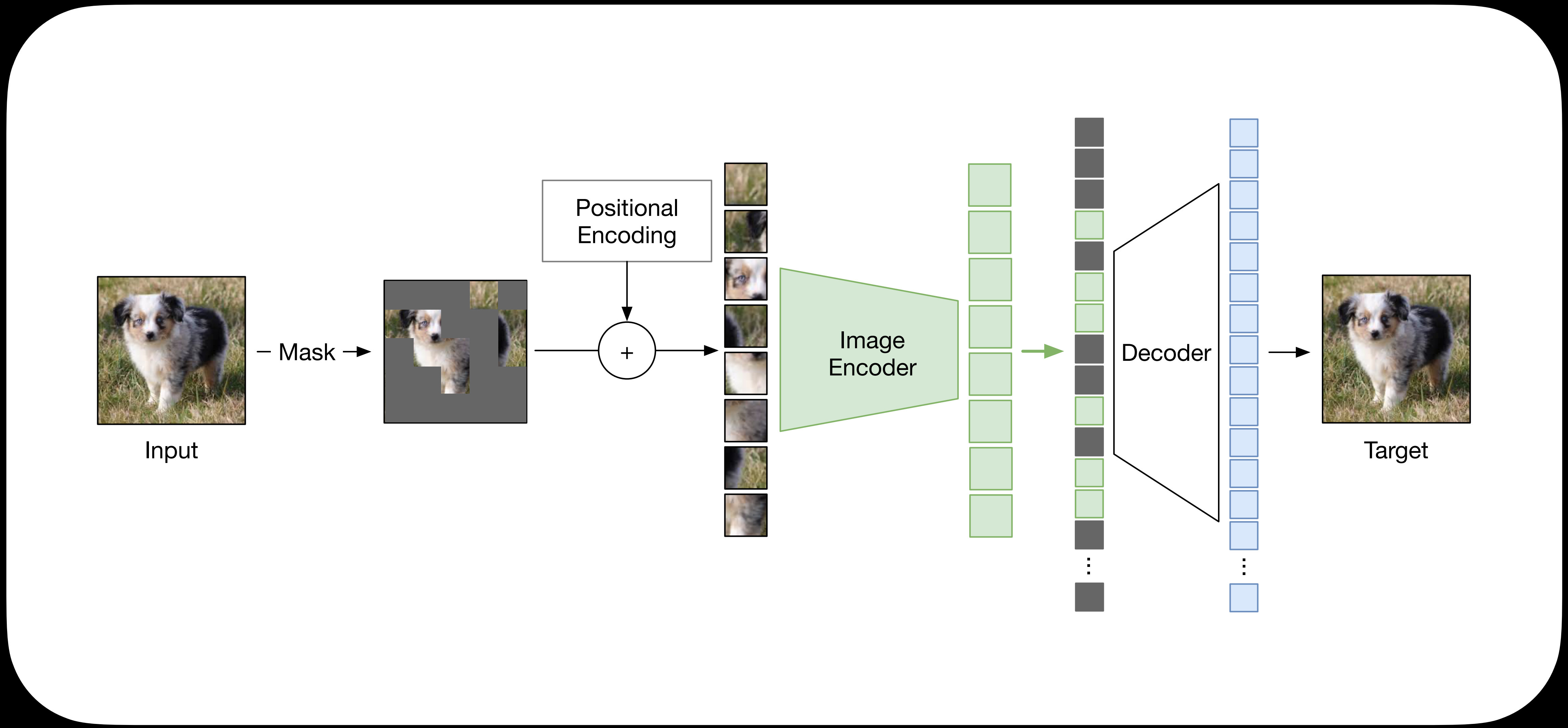
MAE



Background

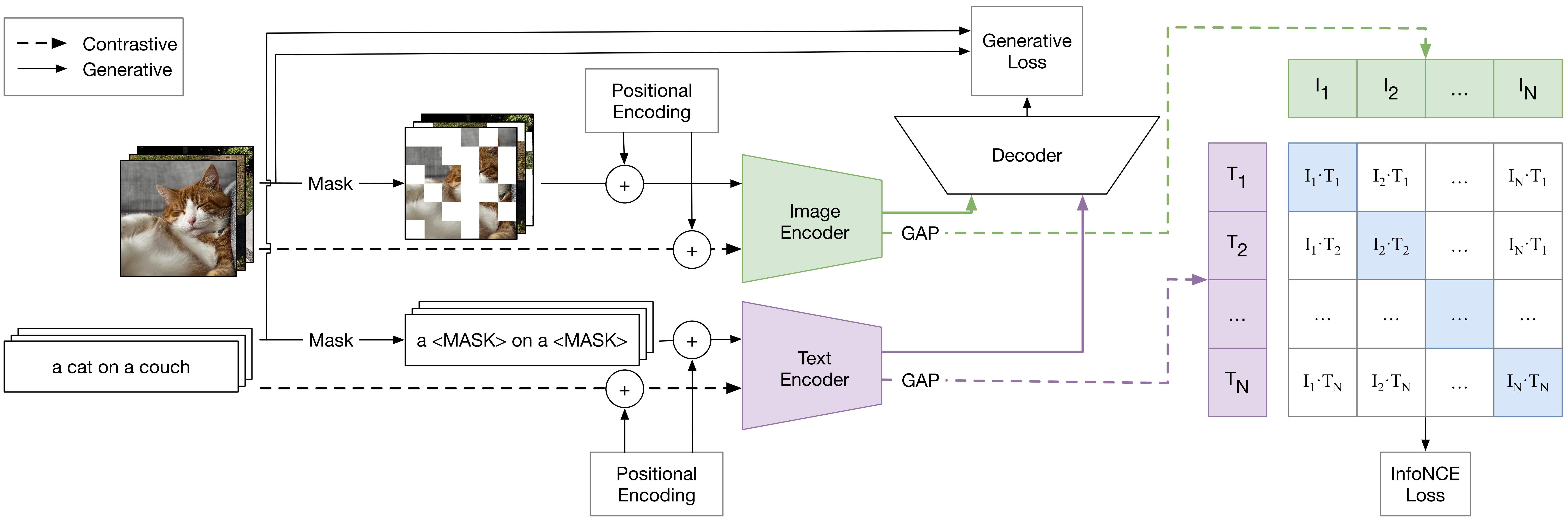
MAE

- Task: generative, predict raw pixels
- High local attention
    - Objective function only considers within-example information
    - High masking ratio



MAE-CLIP

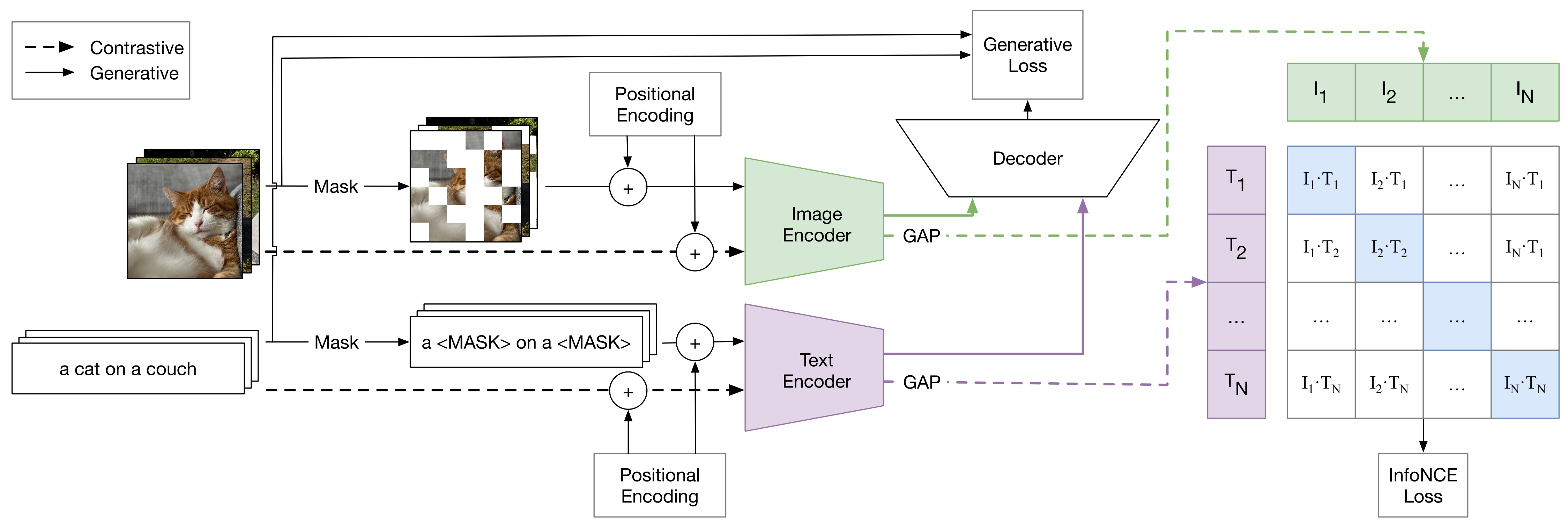
Architecture



MAE-CLIP

Architecture

- Task
  - Contrastive (unmasked)
  - Generative (masked)
- Shared encoders, separate forward passes
  - Compute weighted combination of task objectives



MAE-CLIP

Motivation

- M3AE and SLIP show promising results
- Lack evaluations in “high accuracy” regime or clean ablations

Results

Linear Probing

- At 11.3M training examples it provides clear benefit

Linear-probing on ImageNet

	11.3M
MAE	33.9
M3AE	52.5
CLIP	52.6
MAE-CLIP	**58.9**

Results

Linear Probing

- At 11.3M training examples it provides clear benefit
- Masked self-supervision is not a useful addition to CLIP for sufficiently large datasets
  - At 1.4B examples it does not help

Linear-probing on ImageNet

	11.3M	**1.4B**
MAE	33.9	-
M3AE	52.5	69.3
CLIP	52.6	**77.5**
MAE-CLIP	**58.9**	76.6

Results

VQA finetuning

VQA finetuning after training at large scale (1.4B images), we see that while MAE does improve CLEVR performance, most tasks are not benefited, despite the additional compute

VQA Finetuning

	CLEVR	VQAv2	GQA
M3AE	**96.9**	59.9	53.3
CLIP	87.8	61.8	55.0
MAE-CLIP	92.8	**61.9**	**55.3**

Results

Pooling operation

- The pooling operator has a larger effect than the addition of MAE for improving visual grounding
- Max pooling outperforms global average pooling

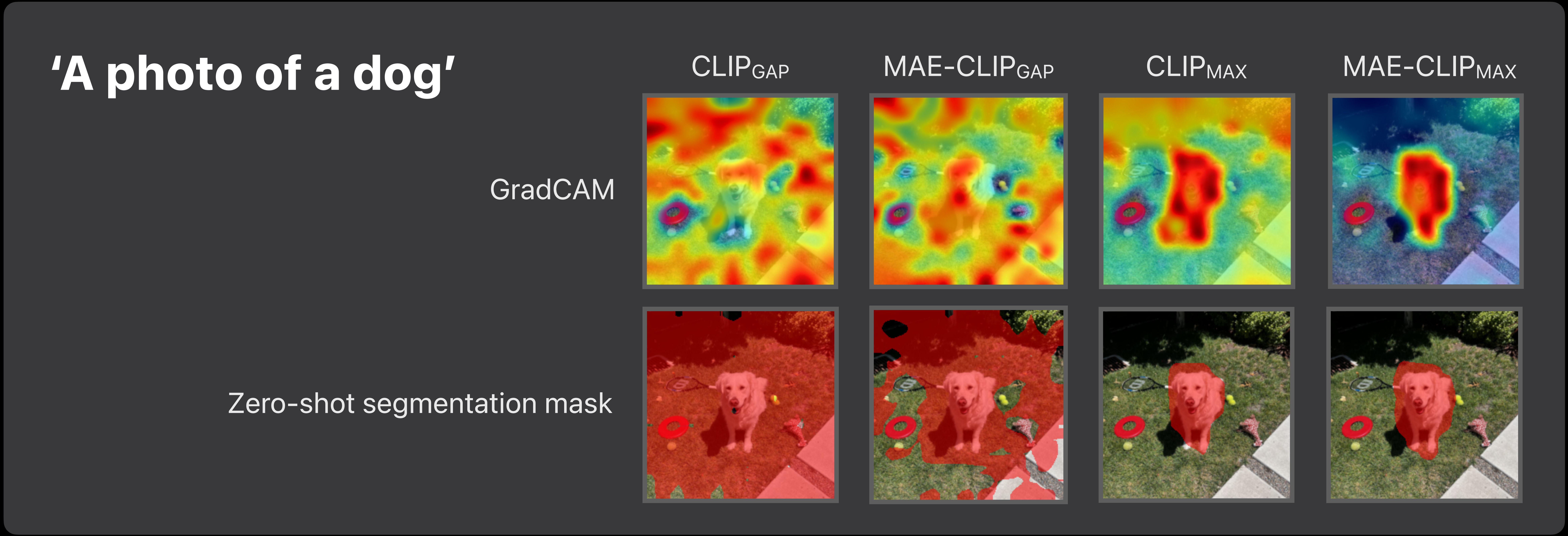
ImageNet performance

Model	Pooling	Zero-shot	Linear Probing
CLIP	GAP	61.8	75.9
	MAX	**63.7**	**77.5**
MAE-CLIP	GAP	57.4	75.7
	MAX	60.9	76.6

Results

Visual grounding

Self-supervision qualitatively improves visual grounding, but the pooling operator has the largest effect.



Future Work

Visual Grounding

- How well a presentation or network can localize objects within an image
- Only incremental improvement of localisation when self-supervision is added
- More thorough future analysis on the relationship between self supervision and visual grounding is needed.

Future Work

Dataset Diversity

- Self-supervision and natural language supervision might excel for entirely different parts of the dataset diversity-size spectrum.
- Scaling trends of self supervised methods are an interesting future line of work.

	COCO		FLICKR		COCOA	
	I→T	T→I	I→T	T→I	Top 1	Top 5
CLIP_{GAP}	51.9	36.6	78.8	62.3	24.2	46.9
CLIP_{MAX}	**55.3**	**39.0**	80.5	**65.3**	22.7	**51.6**
$\text{MAE-CLIP}_{\text{GAP}}$	53.0	37.0	77.3	62.0	20.7	39.5
$\text{MAE-CLIP}_{\text{MAX}}$	54.4	37.7	**81.2**	64.2	**24.6**	41.4