Joint Learning of Visual and Text Representations

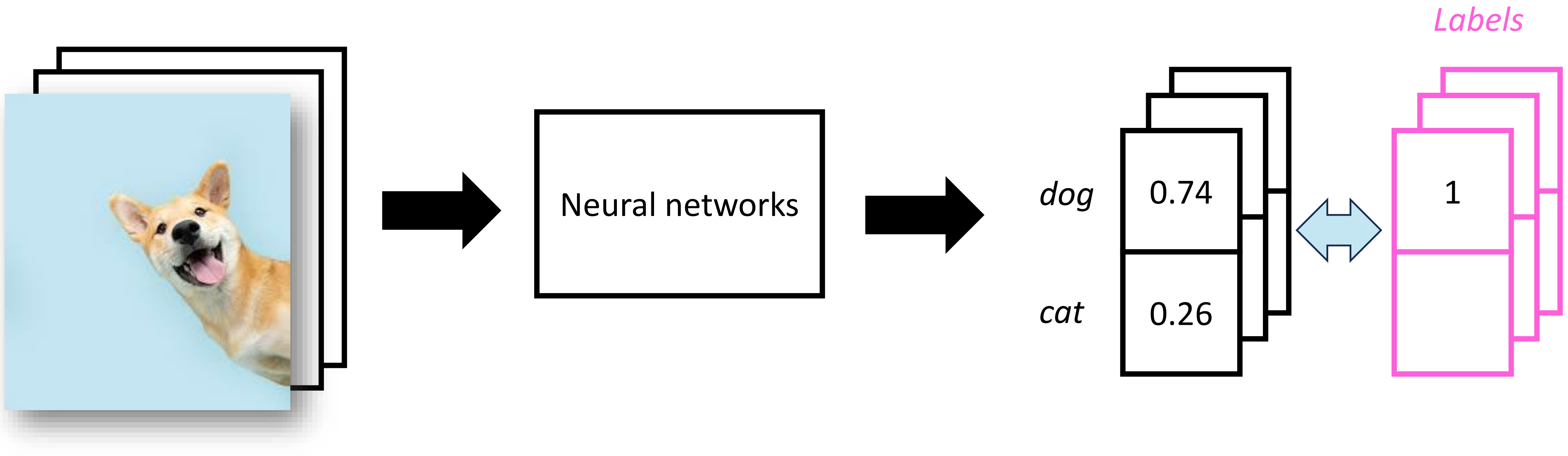
Ph.D candidate in Computational Science and Engineering
Yonsei Univ.

Jin-Duk Park

Reading group material

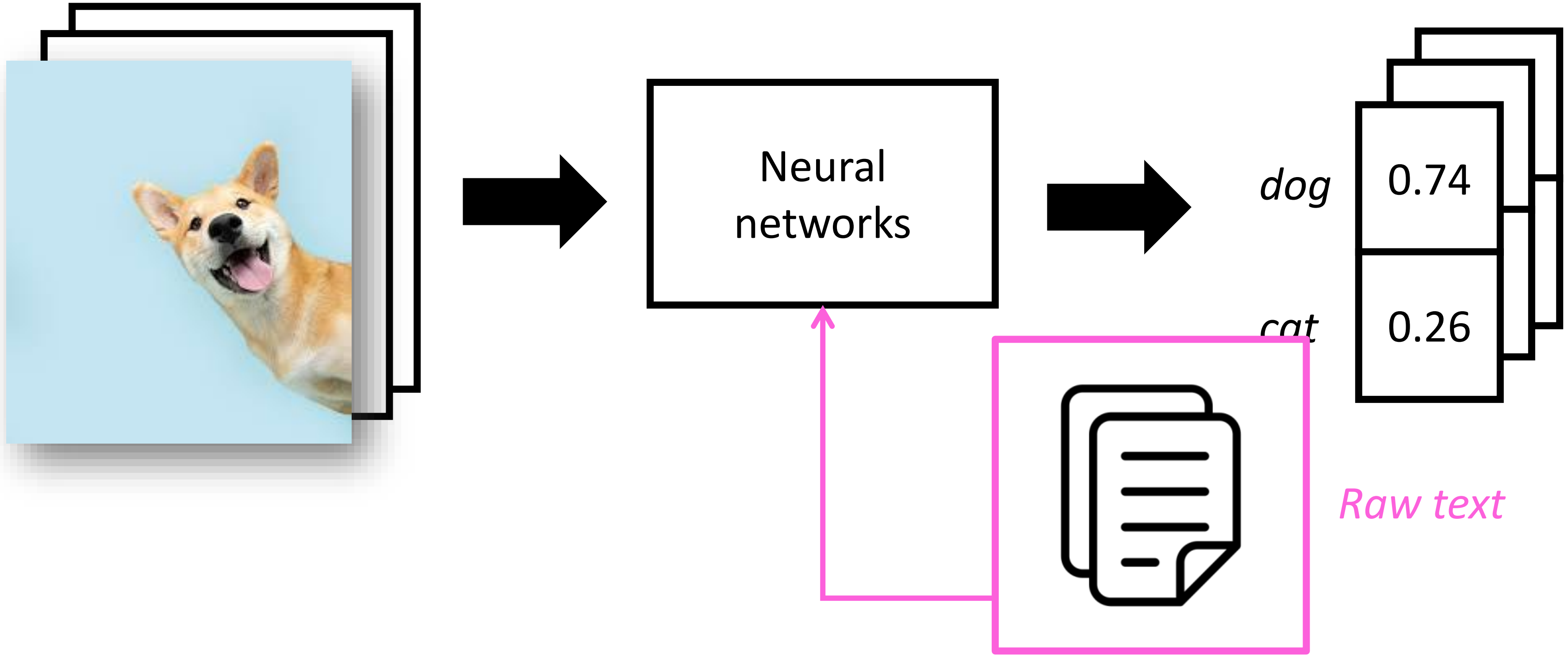
Conventional Supervision in Vision Tasks

- Conventional supervision typically requires label annotation
  - However, label annotation is expensive
  - E.g.) According to OpenAI, **+25,000 workers** for 14M images



Natural Language Supervision

- What if we use **raw text** for improving visual representations?
  - **Vast amount of data available** on web
  - It **does not require** labor-intensive annotations
  - Improvement of **quality of visual representation**



INDEX

Joint learning of visual and text information

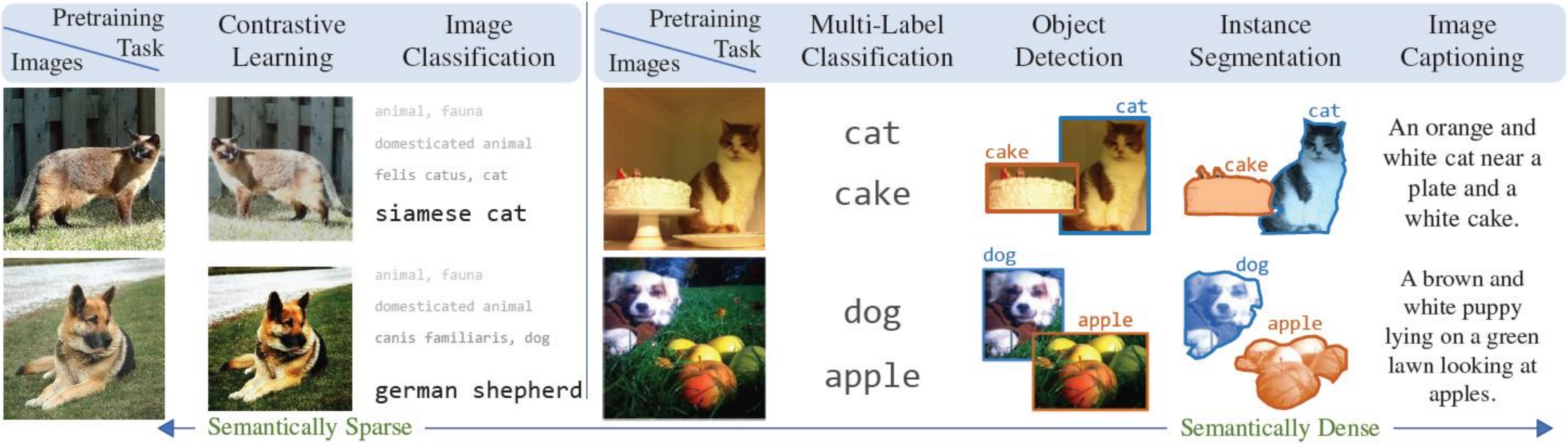
How to get
High-quality dataset?

VirTex [CVPR 2021]

- Leveraging sementically dense (text) information
  - Training with **10x fewer data points**

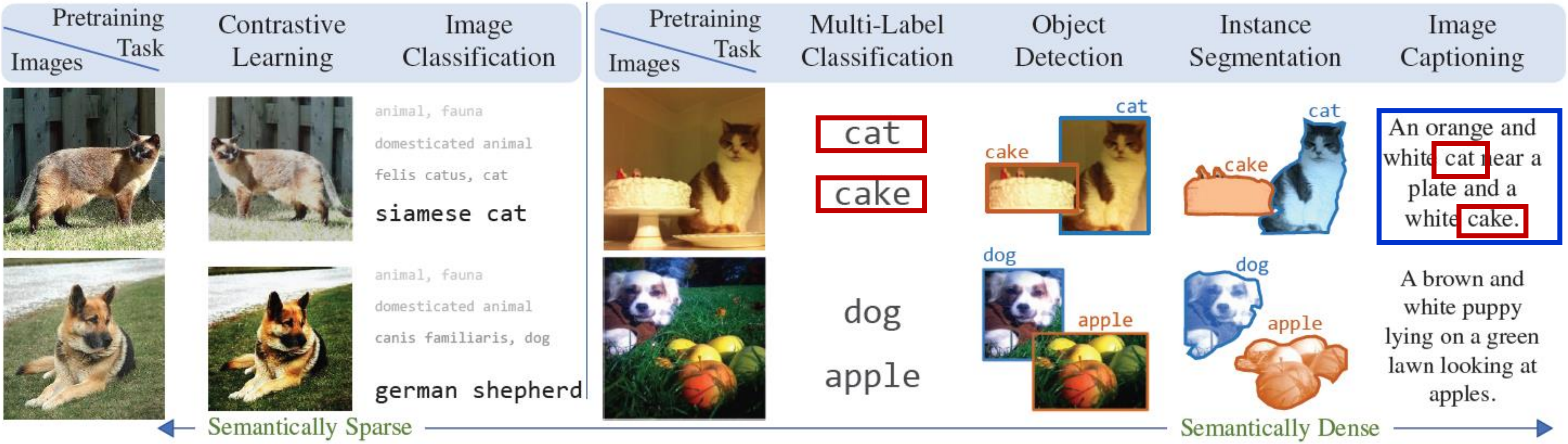
Semantically Sparse vs. Dense

- When use conventional supervision, model doesn't know ***dense* semantics**



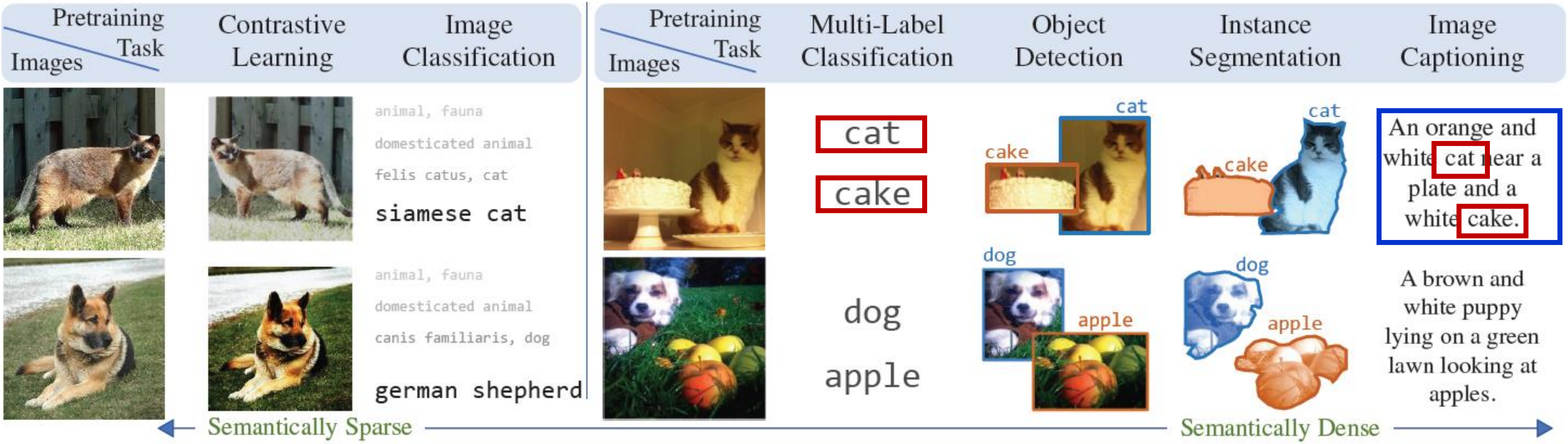
Semantically Sparse vs. Dense

- When use conventional supervision, model doesn't know ***dense* semantics**
  - Image captions provides **additional information:**
    "orange and white **cat** near a plate and a white **cake**"



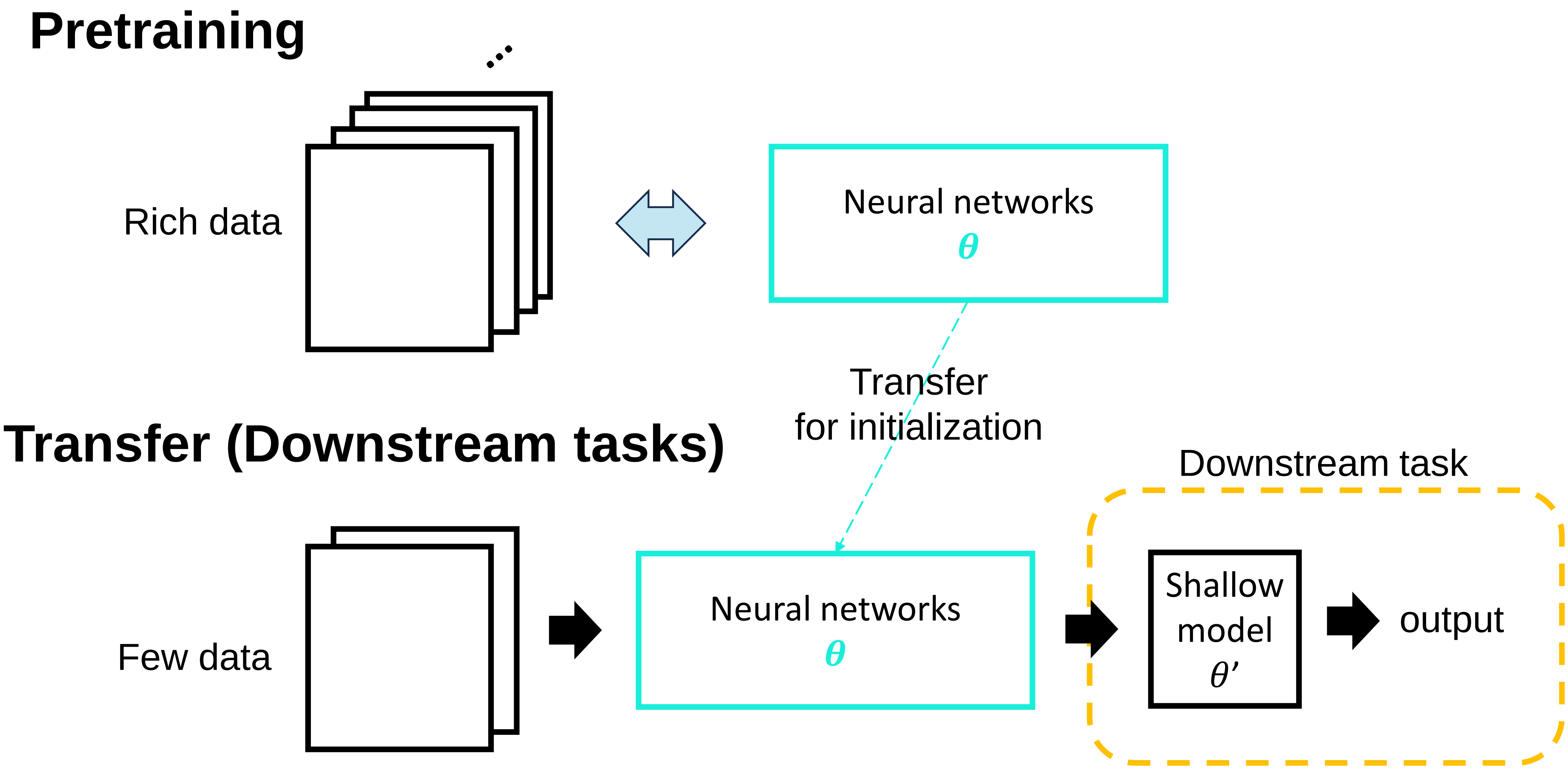
Semantically Sparse vs. Dense

- When use conventional supervision, model doesn't know ***dense* semantics**
  - Image captions provides **additional information:**
    "orange and white **cat** near a plate and a white **cake**"



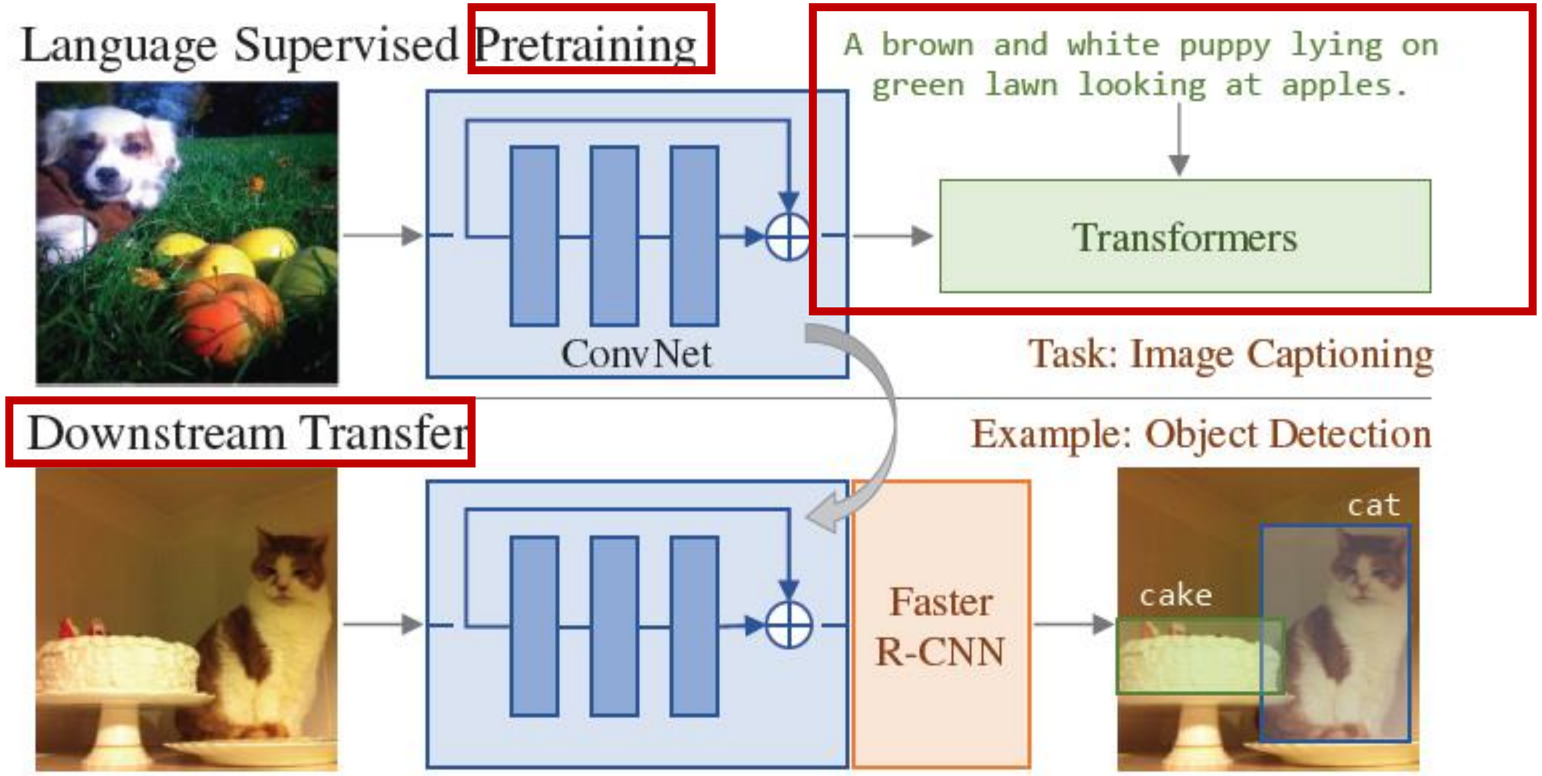
- **How to leverage dense semantics for visual representation learning?**

Short Recap of Transfer Learning



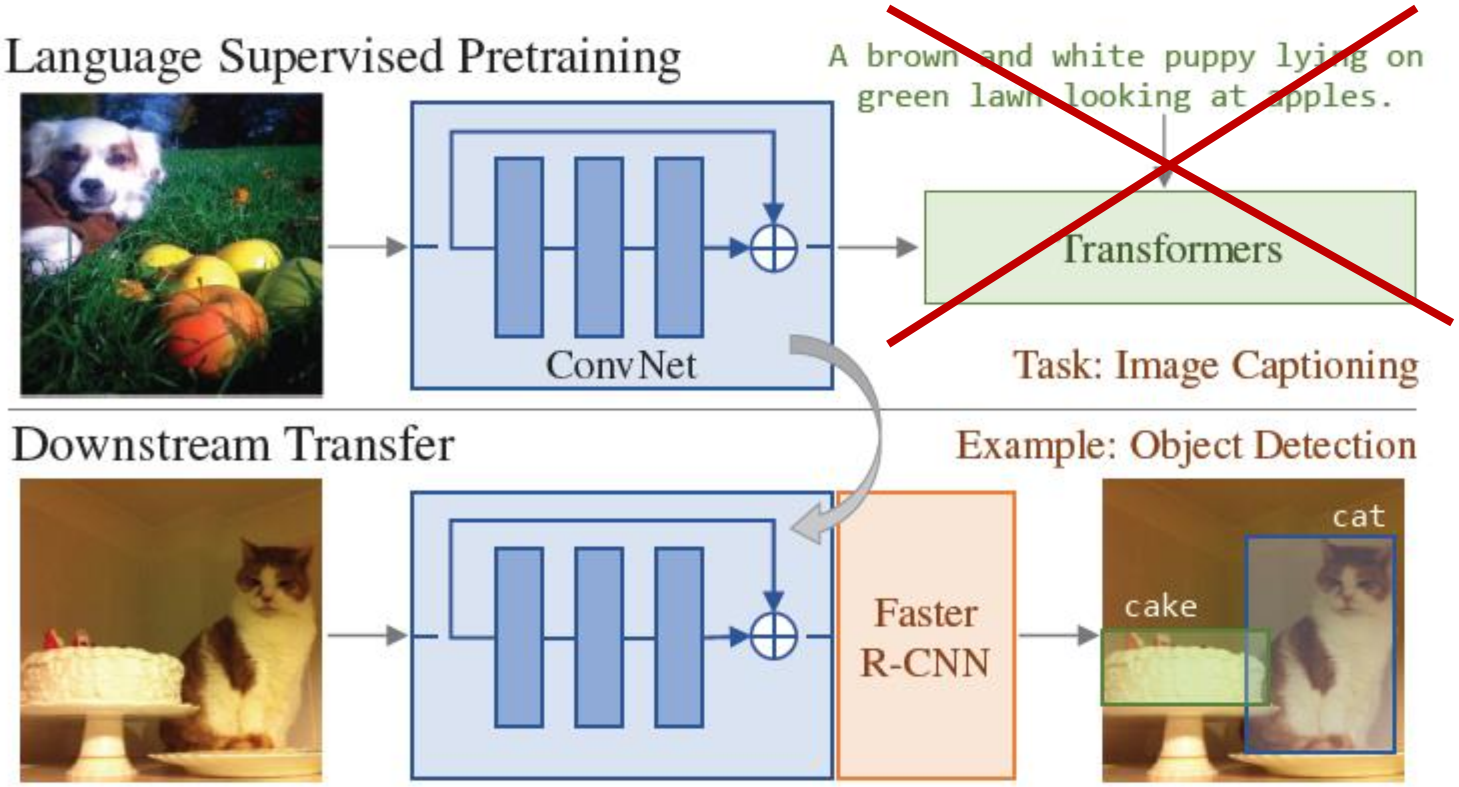
Overview of VirTex

Joint text learning

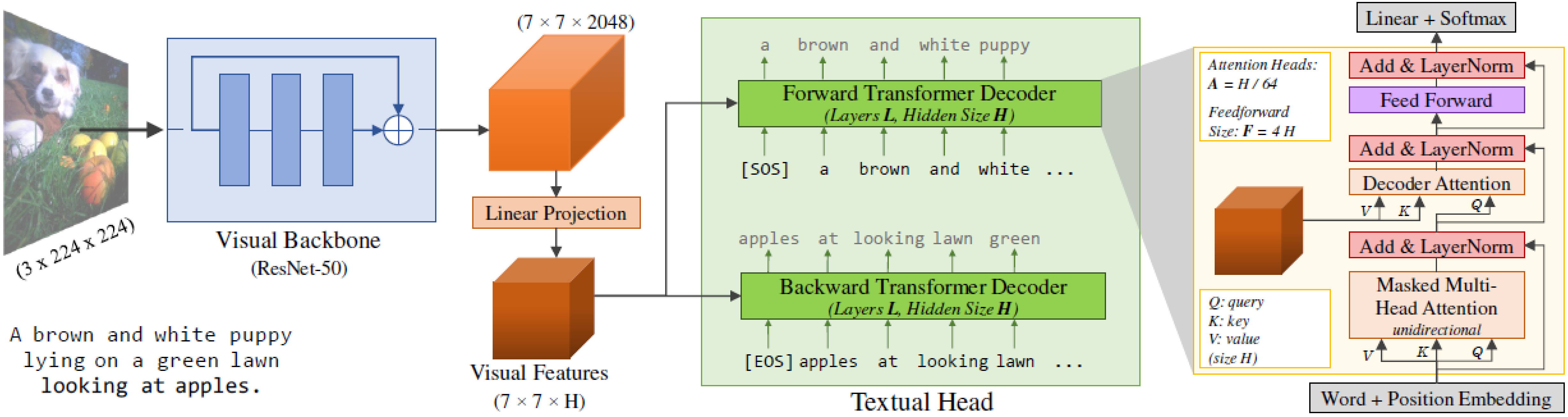


Overview of VirTex

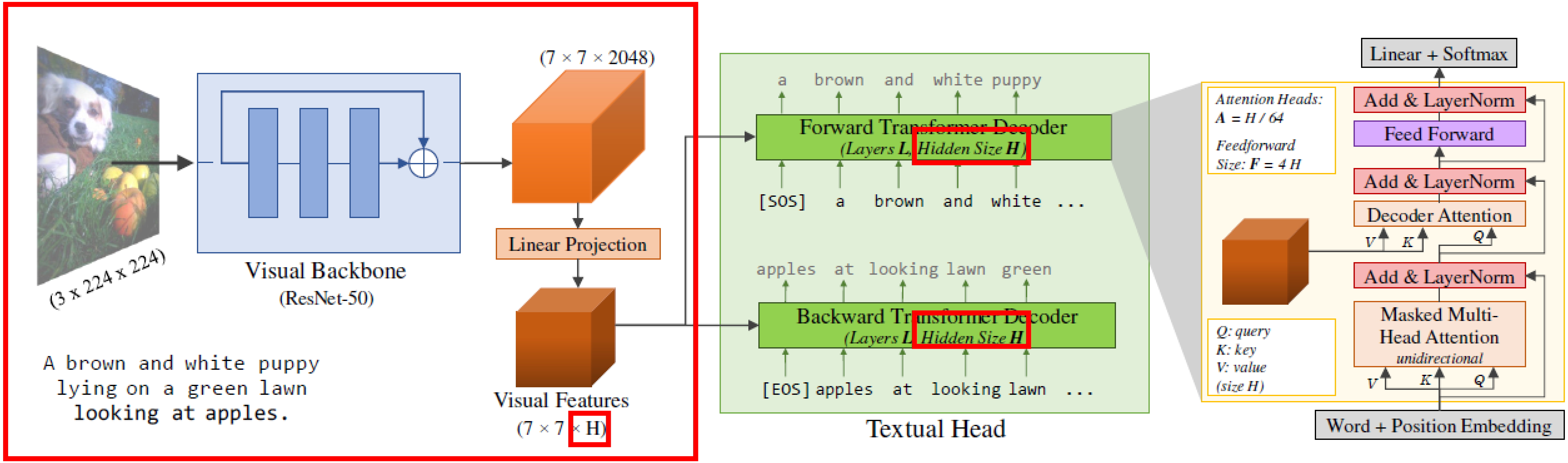
- Here, we **drop text learner (transformer) for transfer**



VirTex Architecture



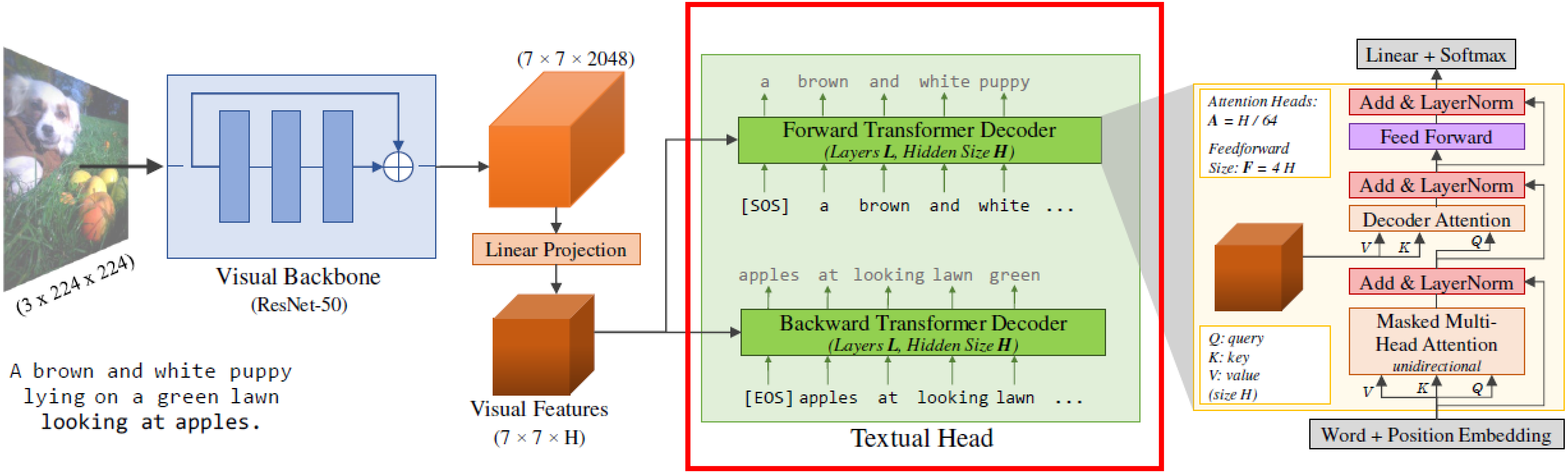
VirTex Architecture



Visual backbone

- ResNet-50 is used for visual learning backbone
- Visual features roughly have 7x7 different positions

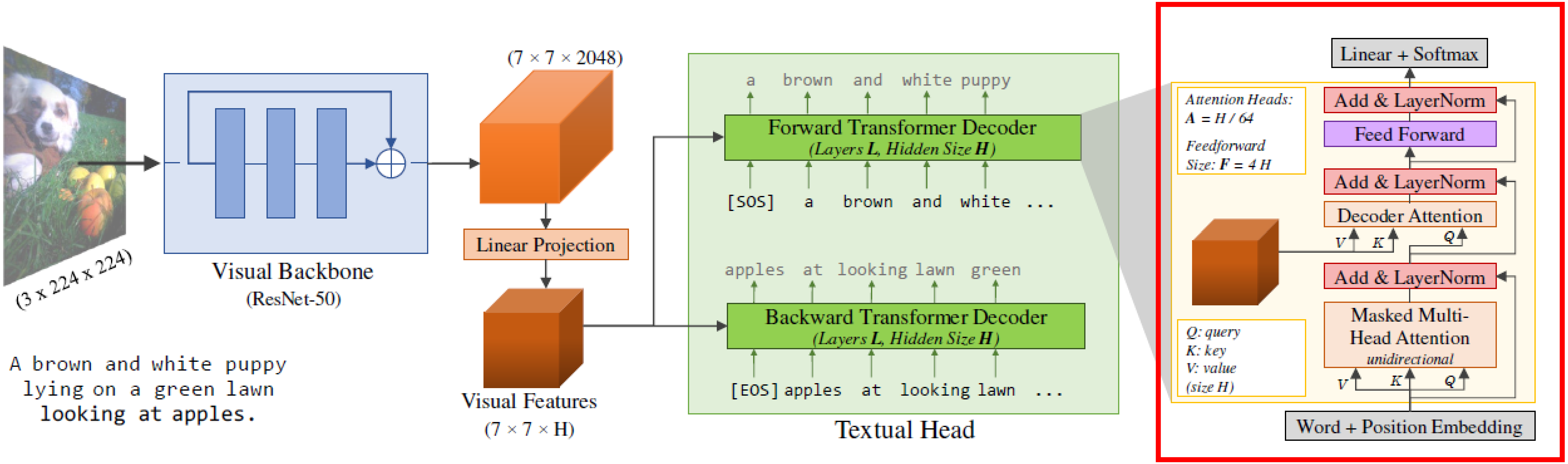
VirTex Architecture



Bidirenctional encoding

- 2 Transformers are used for training (bidirection)
- Two outputs are **not aggregated**: We don’t need inference (only training is enough)

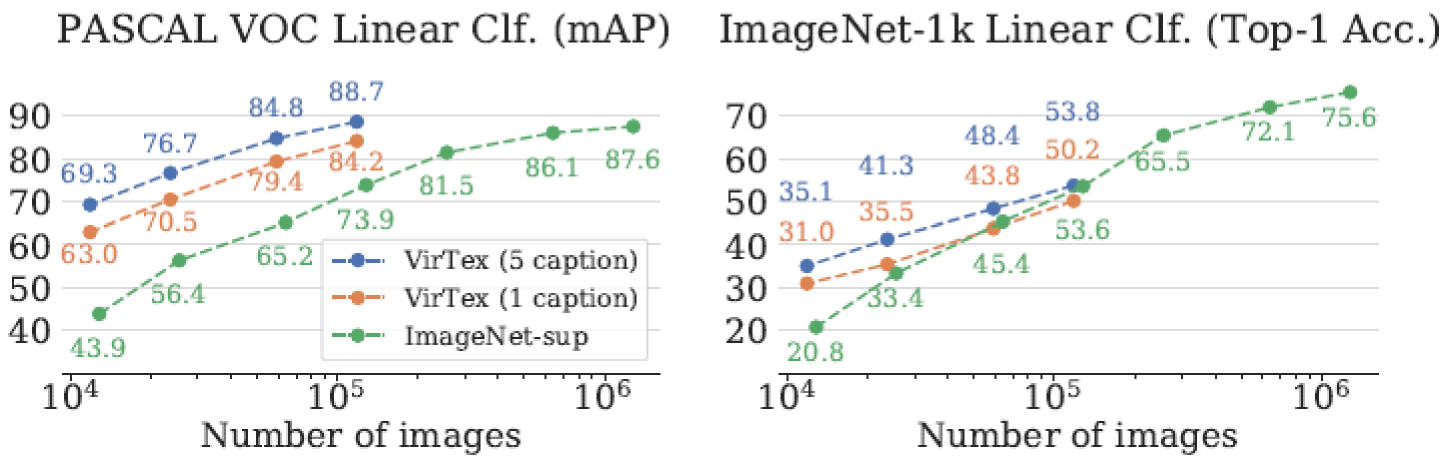
VirTex Architecture



Masked language model (MLM)

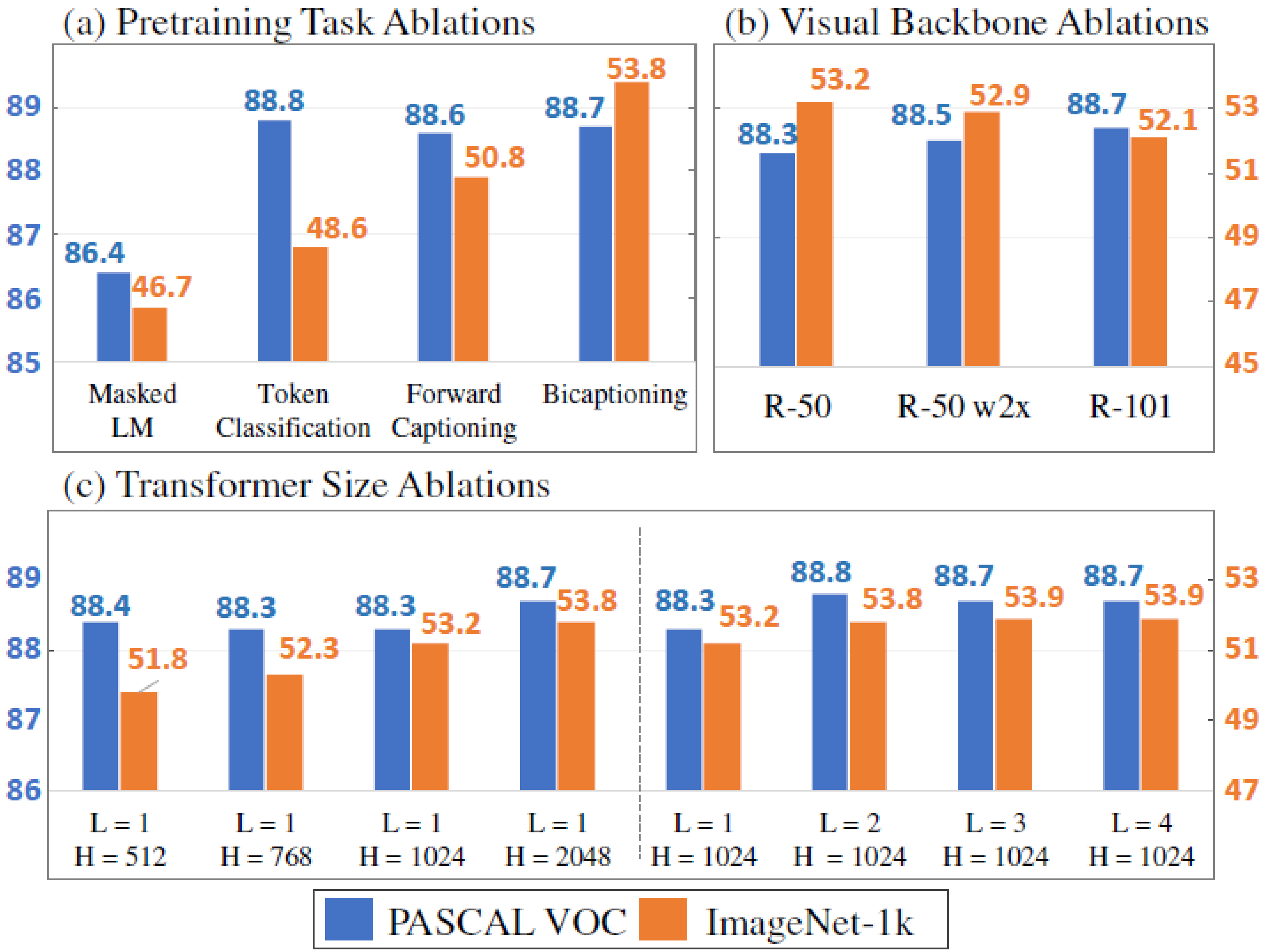
- Basically, same architecture as original transformer decoder
- **Cross-attention** between visual features and text instead
- **Shallow** transformer layer (1-2 layer): as visual part is important

SOTA Performance w/ Fewer Data



- Caption: hot many captions for each image
- ImageNet-sup: accuracy based on conventional supervision
- Can it **exceeds performance of supervision**?

Ablation Study



- Bi-direction is important
- **L=1 (very shallow)** is enough for Transformer

Visualization of Attention Map

VirTex predicted captions (R-50, $L = 1$, $H = 512$), forward transformer decoder

a cat laying on a pair of blue **shoes** | a laptop computer sitting on top of a **desk** | an orange is sitting on the side of a **road** | a dog **riding** on a surfboard in the ocean

- Upscale attention map & overlap on image
- Visual attention aligns well with text part

INDEX

Joint learning of visual and text information

How to get
High-quality dataset?

How to achieve
zero-shot transfer for
downstream tasks?

VirTex [CVPR 2021]

- Leveraging sementically dense (text) information
  - Training with **10x fewer data points**

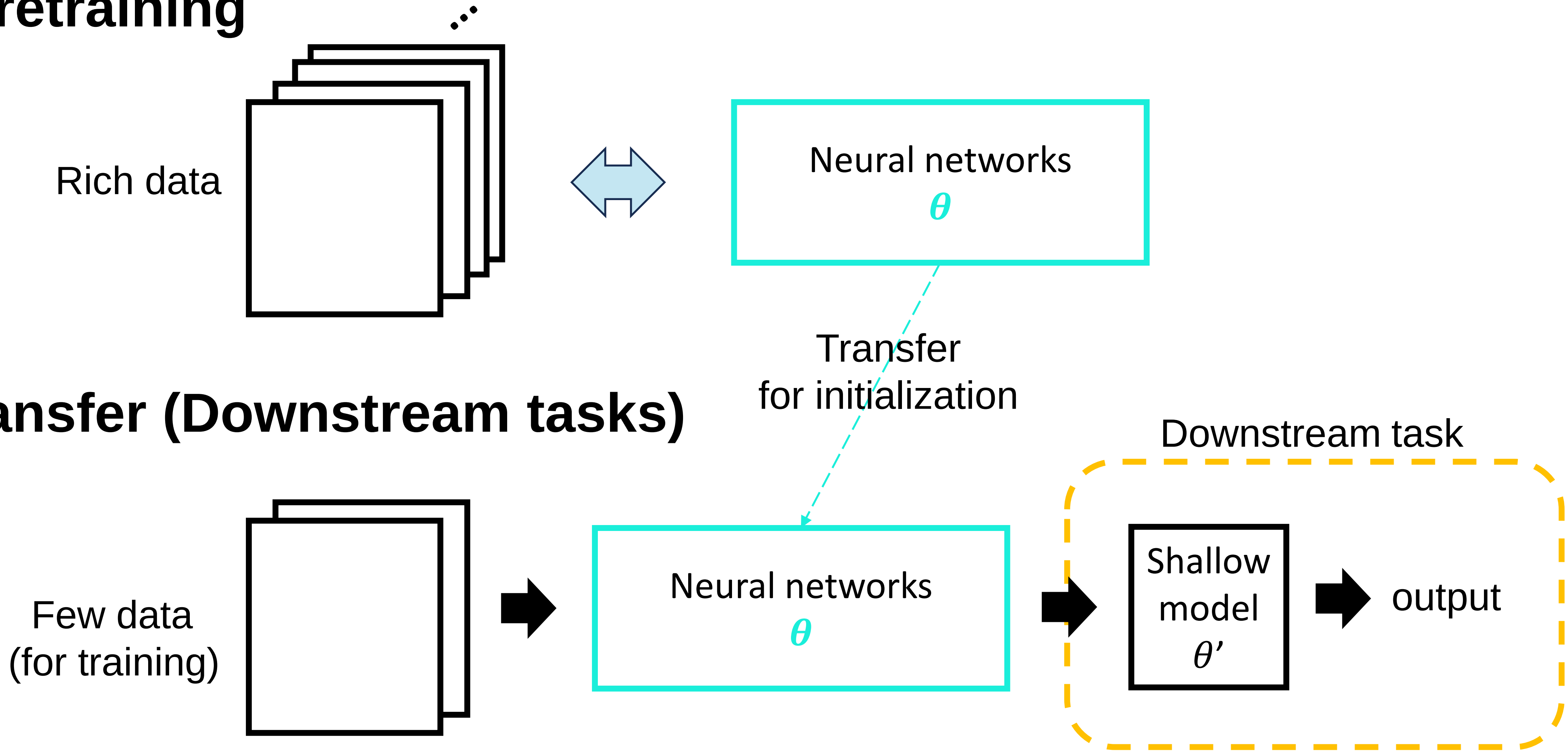
CLIP [ICML 2021]

- Utilize text information for learning
  - For **zero-shot** prediction

Desai et al., VirTex: Learning Visual Representations from Textual Annotations, CVPR 2021
Radford et al., Learning Transferable Visual Models From Natural Language Supervision, ICML 2021

What is Zero-Shot Learning?

Pretraining



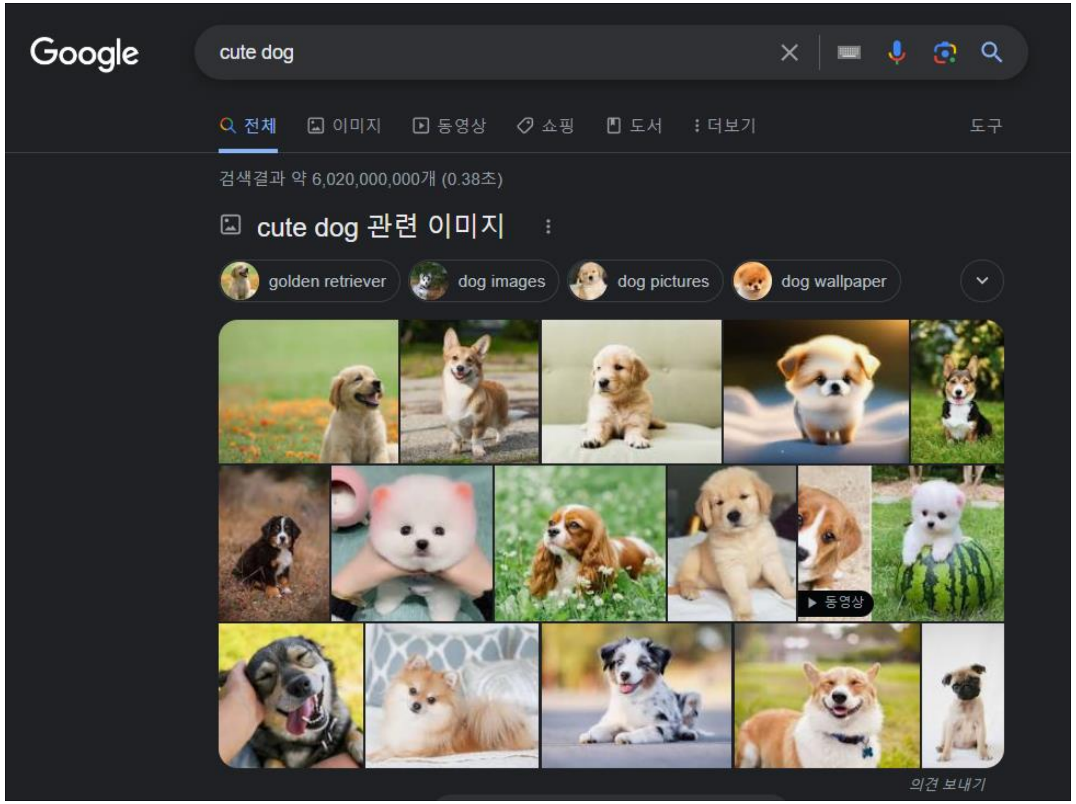
Transfer (Downstream tasks)

Zero shot: No fine-tuning

Dataset Collection

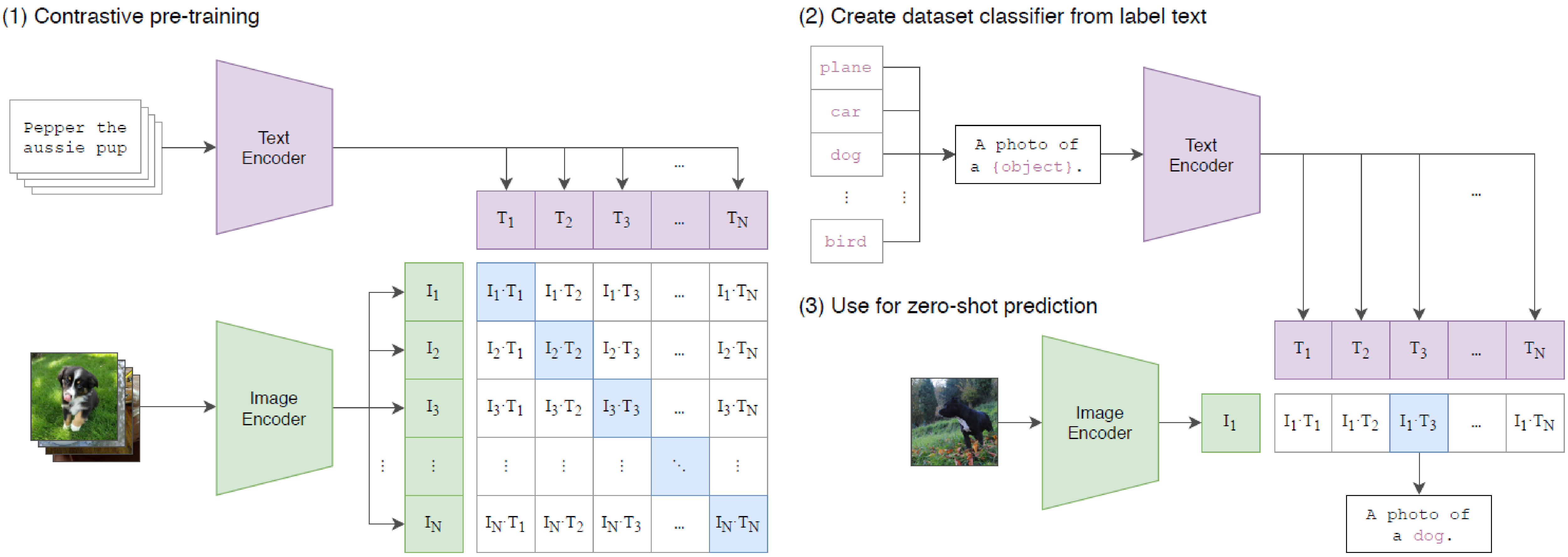
Typical image dataset size: 3.5 billion, while 100K for MS-COCO (not enough)

OpenAI collects 400M (image, text) pair via Web querying

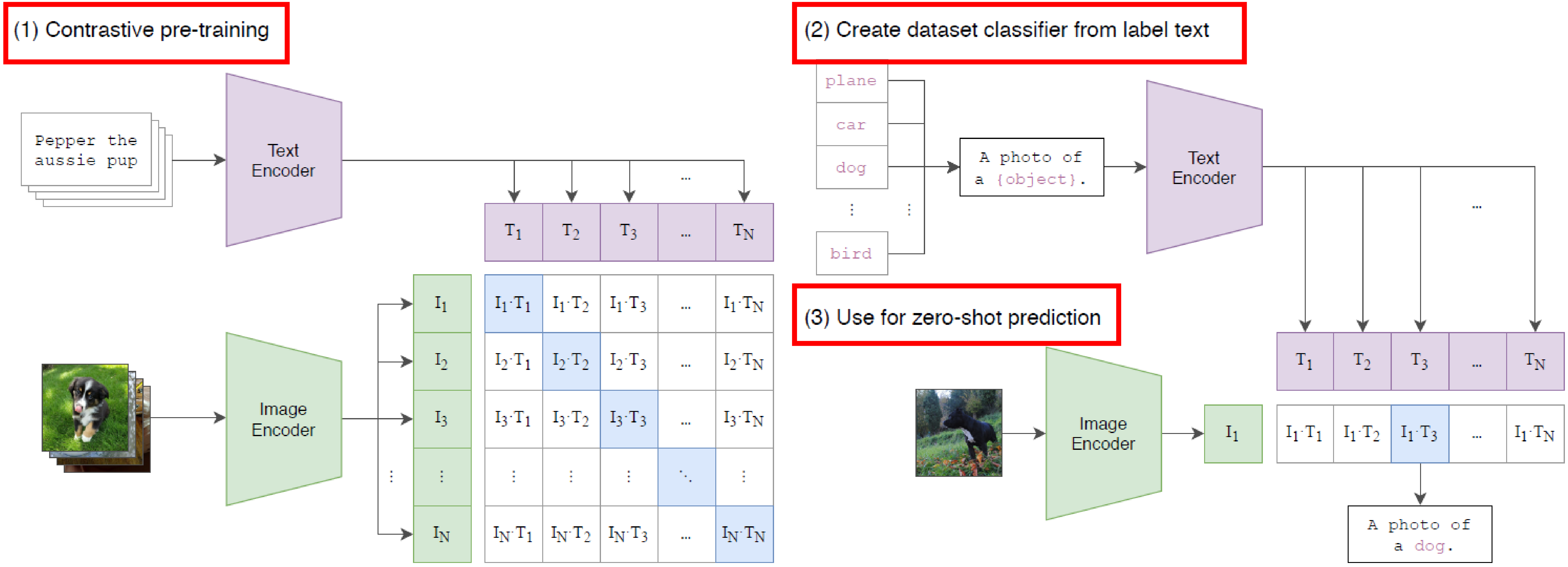


- Note: we don’t leverage dense text now
- Worries about data quality?

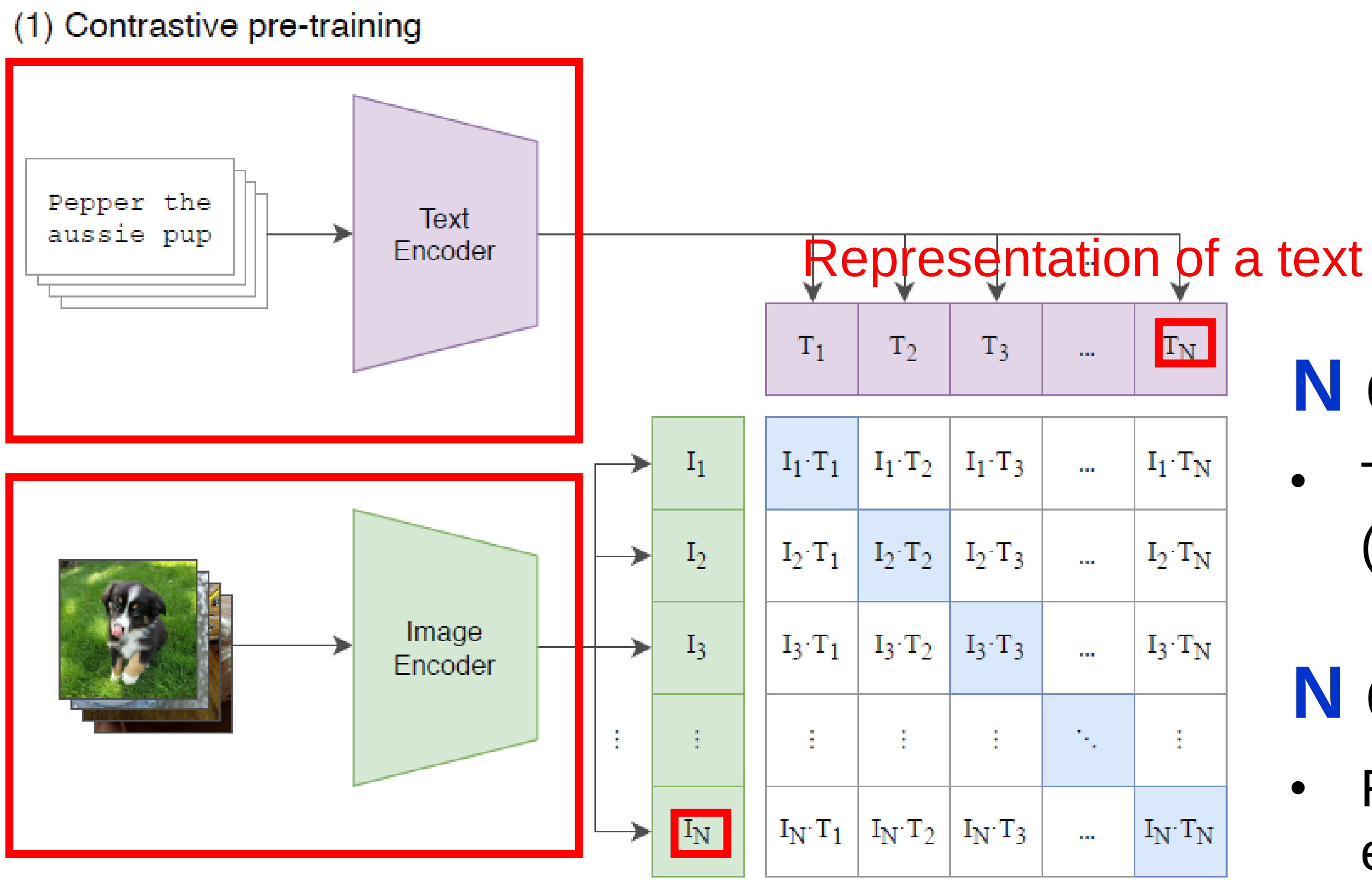
Overview of CLIP



Overview of CLIP



CLIP: Contrastive Pre-Training



N different texts

- Transformer: encode each text sentence (word or sentence)

N different images

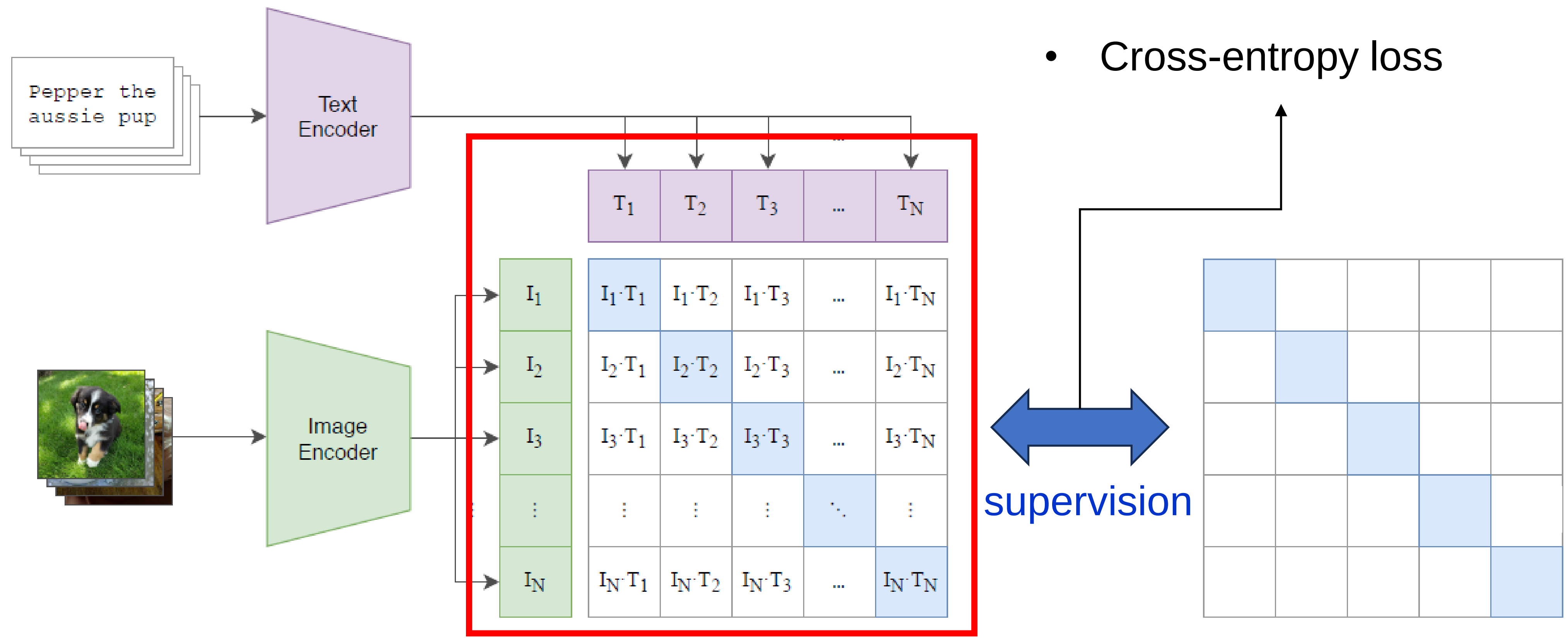
- ResNet50 for backbone visual encoder

CLIP: Contrastive Pre-Training

(1) Contrastive pre-training

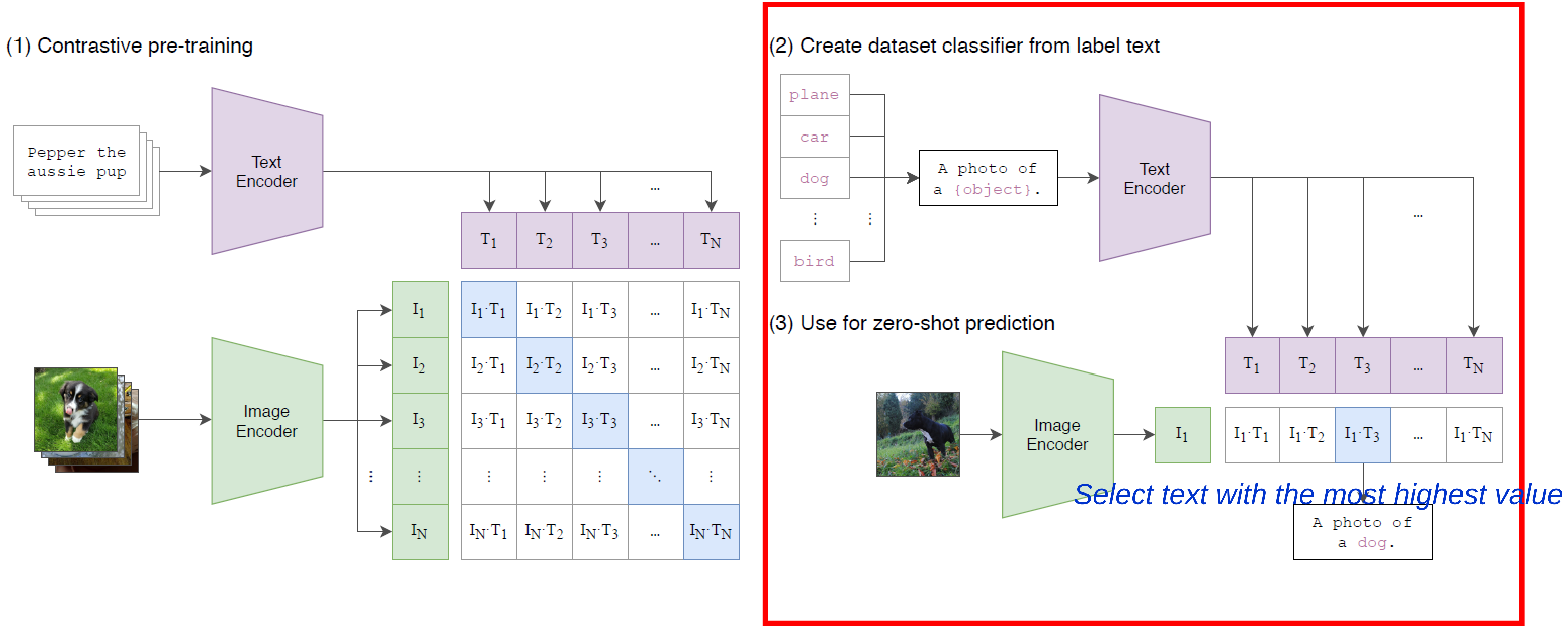
Supervised training

- Cross-entropy loss



CLIP: Create Dataset

- Label to text: To achieve zero-shot transfer, formats should be matched (dataset should be created)



- Perform zero-shot prediction with unseen data

Evaluations

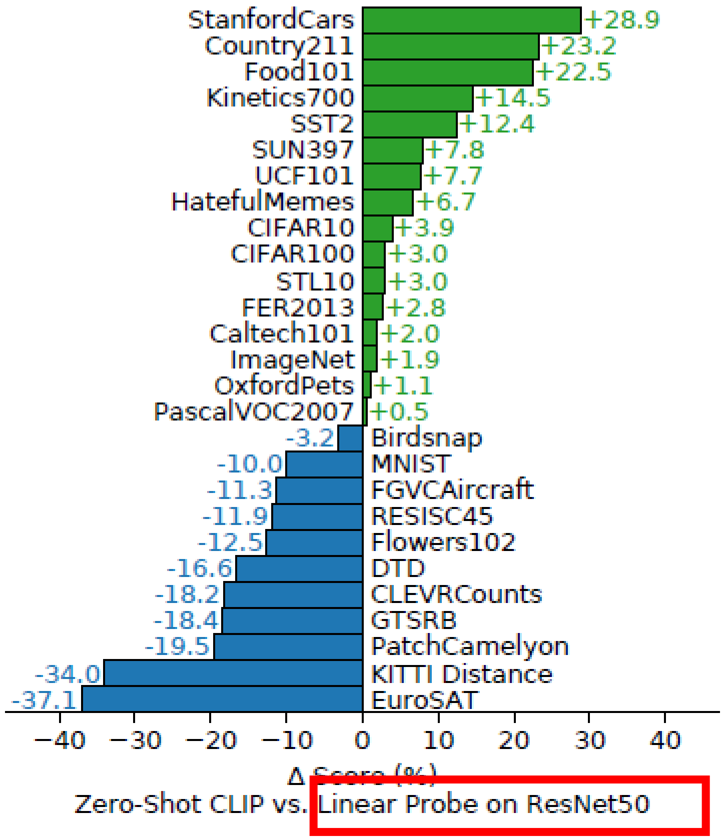


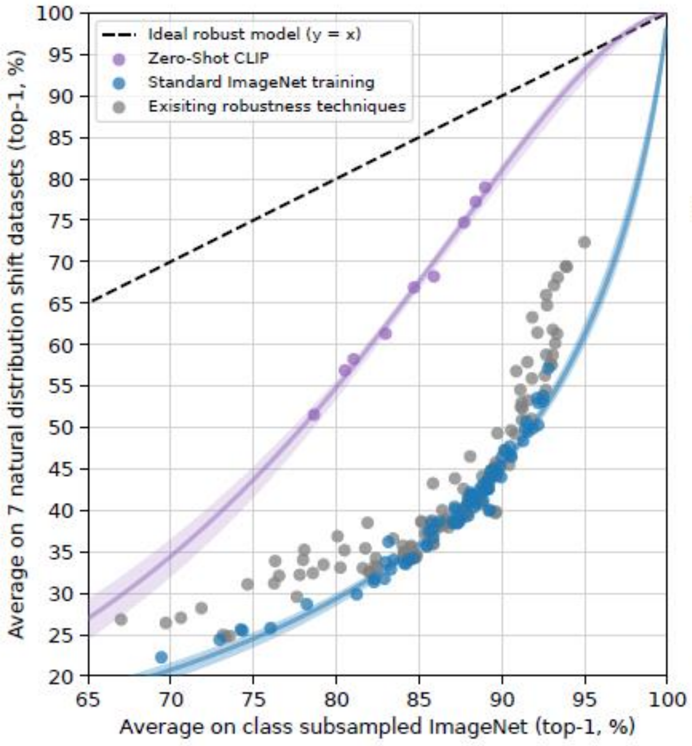
Figure 4. **Zero-shot CLIP is competitive with a fully supervised baseline.** Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet50 features on 16 datasets, including ImageNet.

- Fine-tuning on ResNet50 vs. CLIP
  - 4-shot is used for the baseline
  - CLIP (zero-shot) **even outperforms** few-shot learning
  - Outperforms in 16/27 datasets

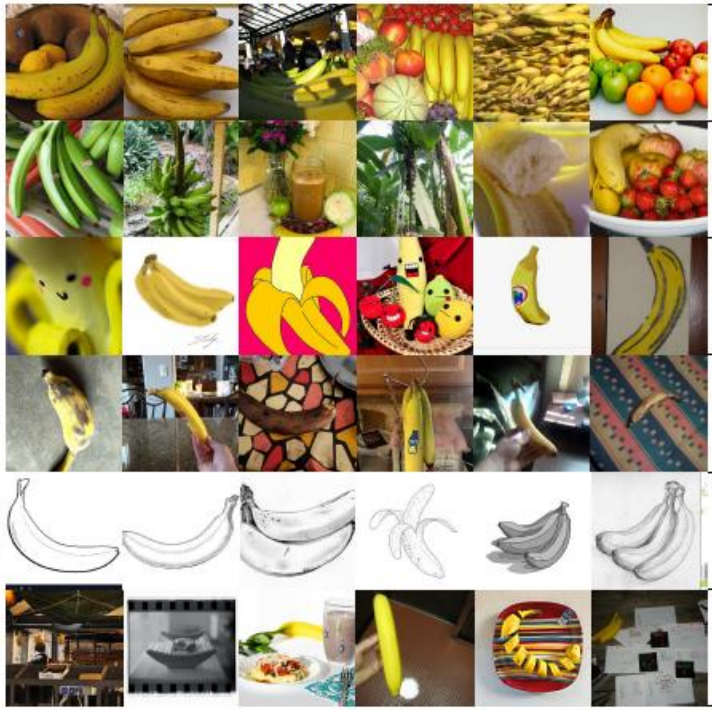
- Weak performance on several specialized, complex or abstract tasks
  - Satellite image classification (EuroSAT and RESISC45), lymph node tumor detection (PatchCamelyon), counting objects in synthetic scenes (CLEVRCounts), ...

Evaluations

- Robustness to **natural distribution shift**
  - Reduce robustness gap by up to 75%
  - zero-shot model should not be able to exploit spurious correlations or patterns that hold only on a specific distribution, since it is not trained on that distribution



Dataset Examples



Dataset	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet	**76.2**	**76.2**	0%
ImageNetV2	64.3	**70.1**	+5.8%
ImageNet-R	37.7	**88.9**	+51.2%
ObjectNet	32.6	**72.3**	+39.7%
ImageNet Sketch	25.2	**60.2**	+35.0%
ImageNet-A	2.7	**77.1**	+74.4%

Takeaways

- Jointly learning visual representation with text information is very helpful
- (VirTex) Exploiting dense semantics via text sentence is much helpful
- CLIP (a zero-shot model) is good for learning domain-agonostic, general feature of images.

"Success is not final, failure is not fatal:
it is the courage to continue that counts."
- Winston Churchill

Thank you!

jindeok6@yonsei.ac.kr