MoDE: CLIP Data Experts via Clustering

Jiawei Ma, Po-Yao Huang, Saining Xie, Shang-Wen Li,
Luke Zettlemoyer, Shih-Fu Chang, Wen-Tau Yih, Hu Xu

Project: https://github.com/facebookresearch/MetaCLIP/tree/main/mode

Mixture of Data Expert (MoDE)

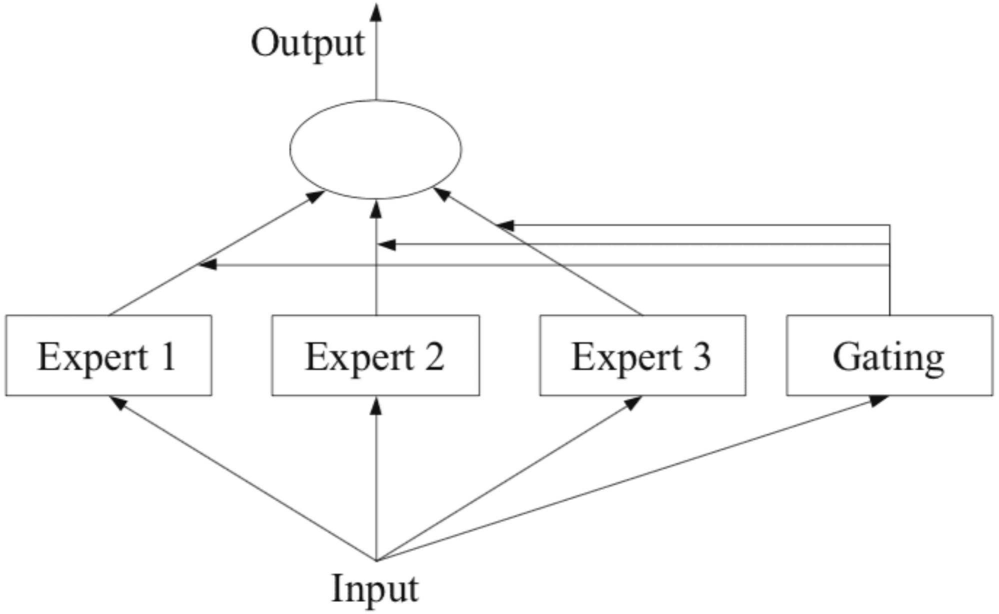
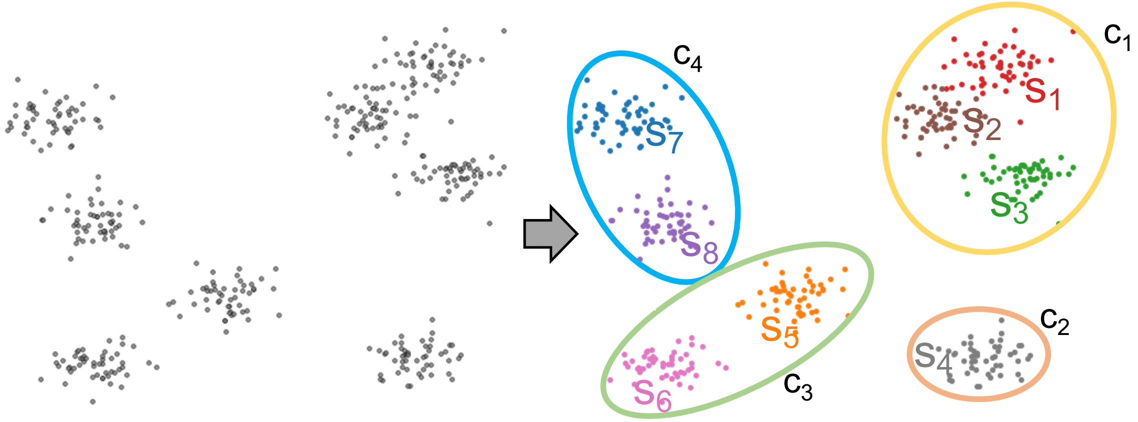
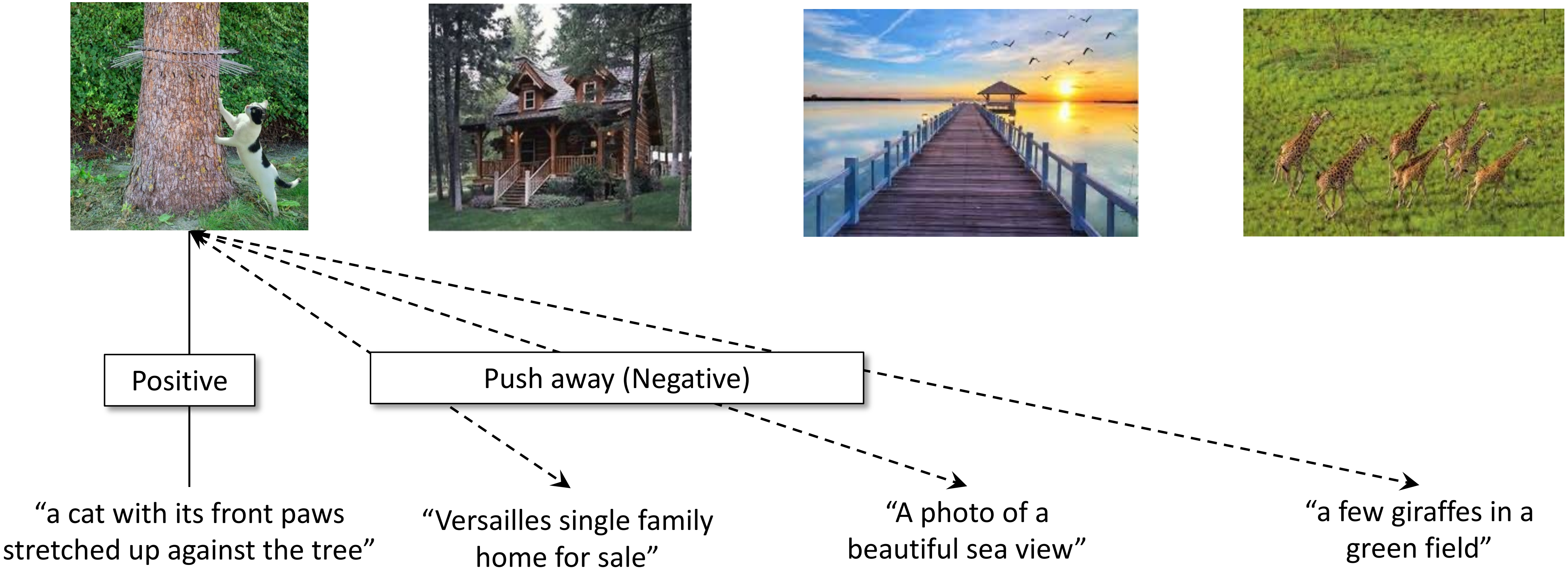


Illustration of Mixture of Expert (MoE)



MoDE: You Reap What You Sow

Contrastive Language-Image Pretraining (CLIP)



Positive

Push away (Negative)

“a cat with its front paws stretched up against the tree”

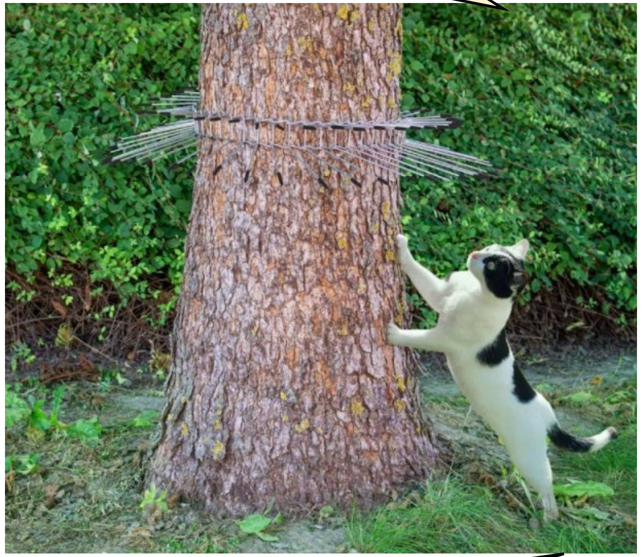
“Versailles single family home for sale”

“A photo of a beautiful sea view”

“a few giraffes in a green field”

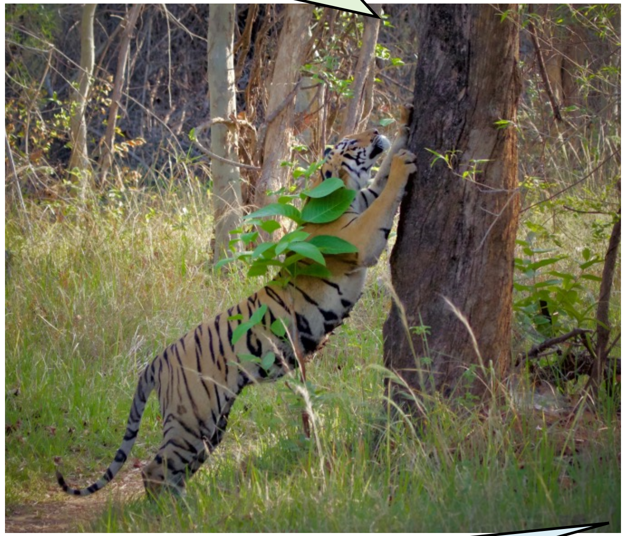
An Image is Worth A Thousand Words

Tree guard to stop the cats



a cat with its front paws stretched up against the tree

The tiger reaches up to a tree trunk in a wooded area

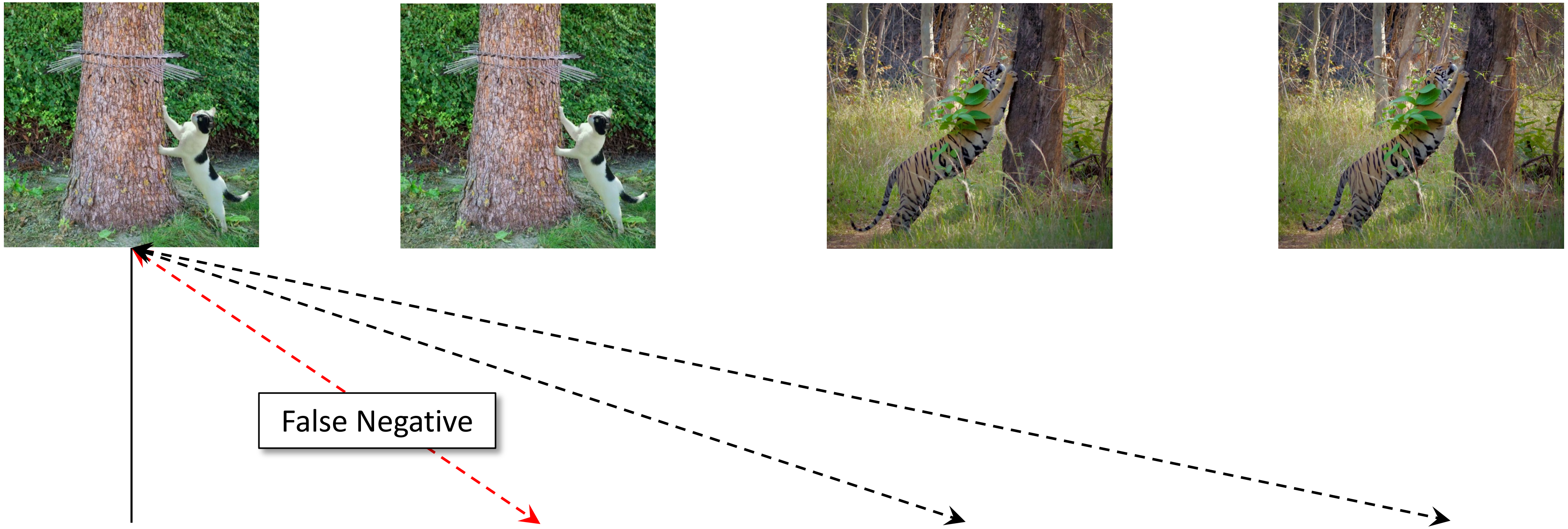


A picture took in a national park

Topics are color-coded

Negative Quality in Web-Crawled Data

The annotation noise/conflict in language may result in false negative in CLIP training.



False Negative

“a cat with its front paws stretched up against the tree”

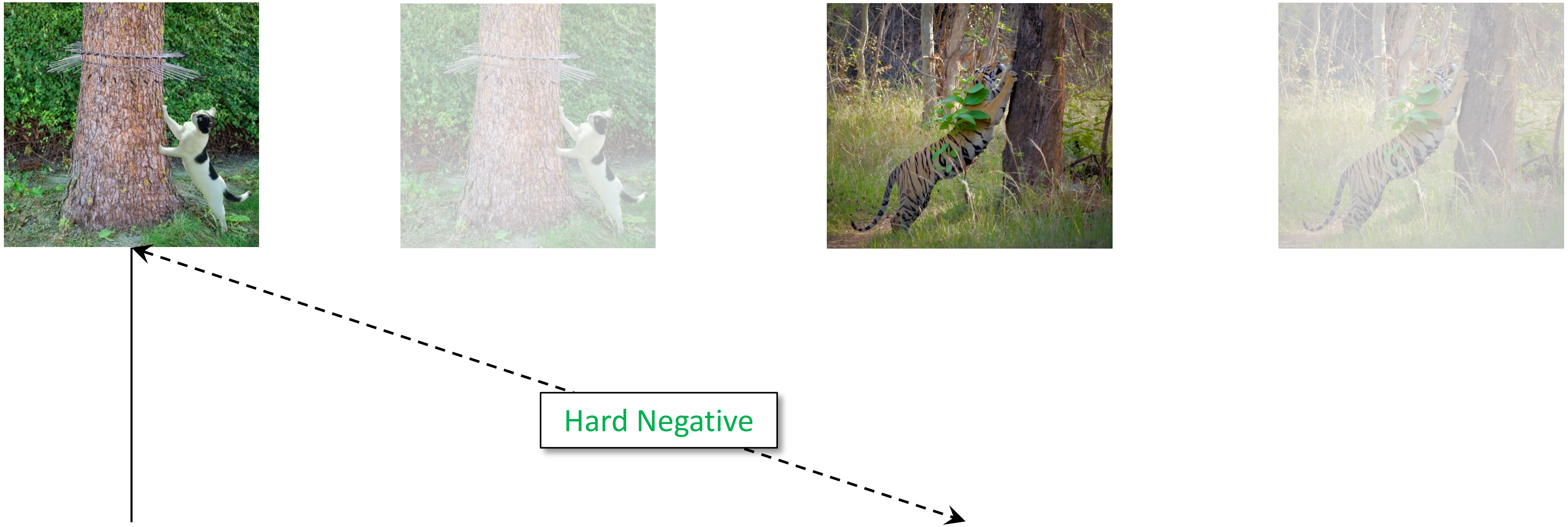
“Tree guard stops the cat”

“The tiger reaches up to a tree trunk in a wooded area”

“A picture took in a national park”

Negative Quality in Web-Crawled Data

Contrasting with hard negative can improve CLIP training effectiveness



Hard Negative

"a cat with its front paws stretched up against the tree"

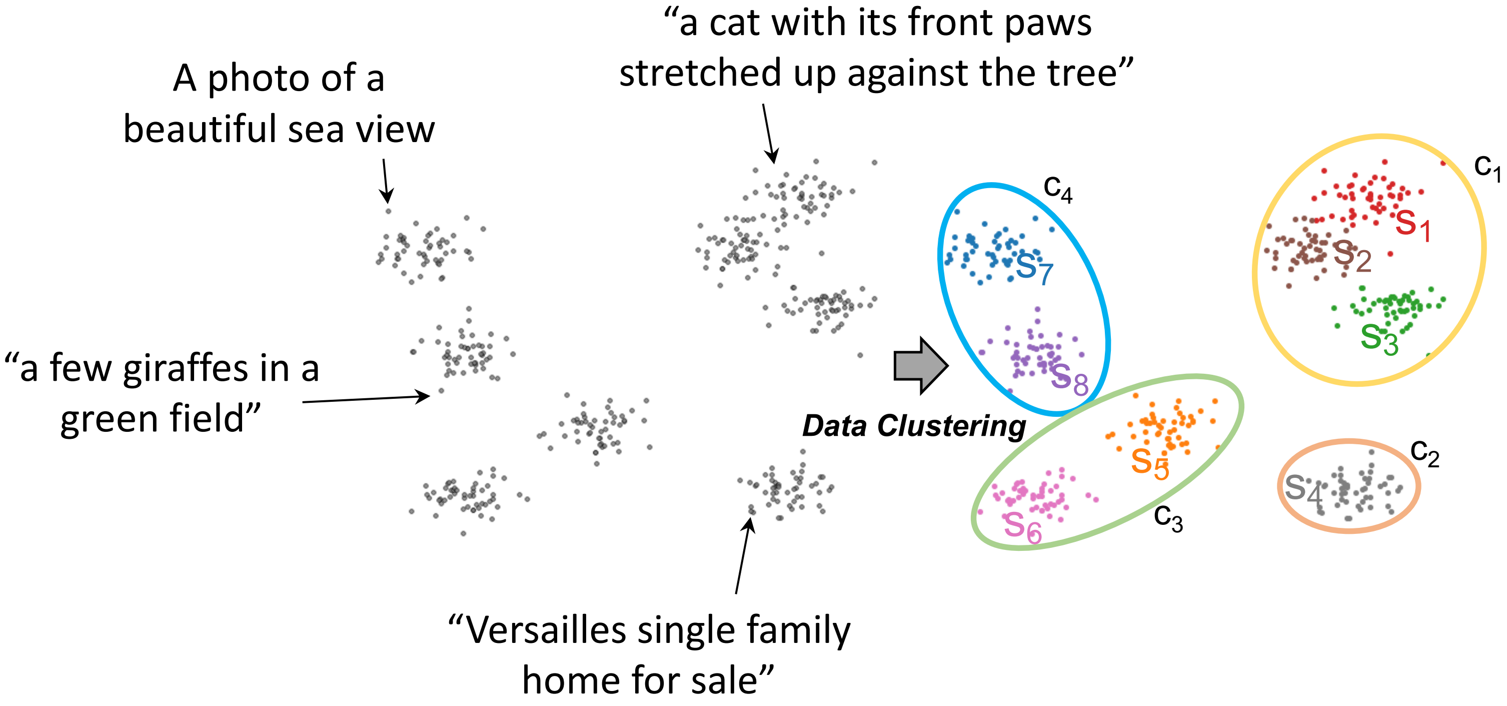
"Tree guard stops the cat"

"The tiger reaches up to a tree trunk in a wooded area"

"A picture took in a national park"

Learning Data Experts via Clustering

Clustering/Splitting along captions to remove false negative and increase hard negative, improving the effectiveness of CLIP training.



A photo of a beautiful sea view

“a cat with its front paws stretched up against the tree”

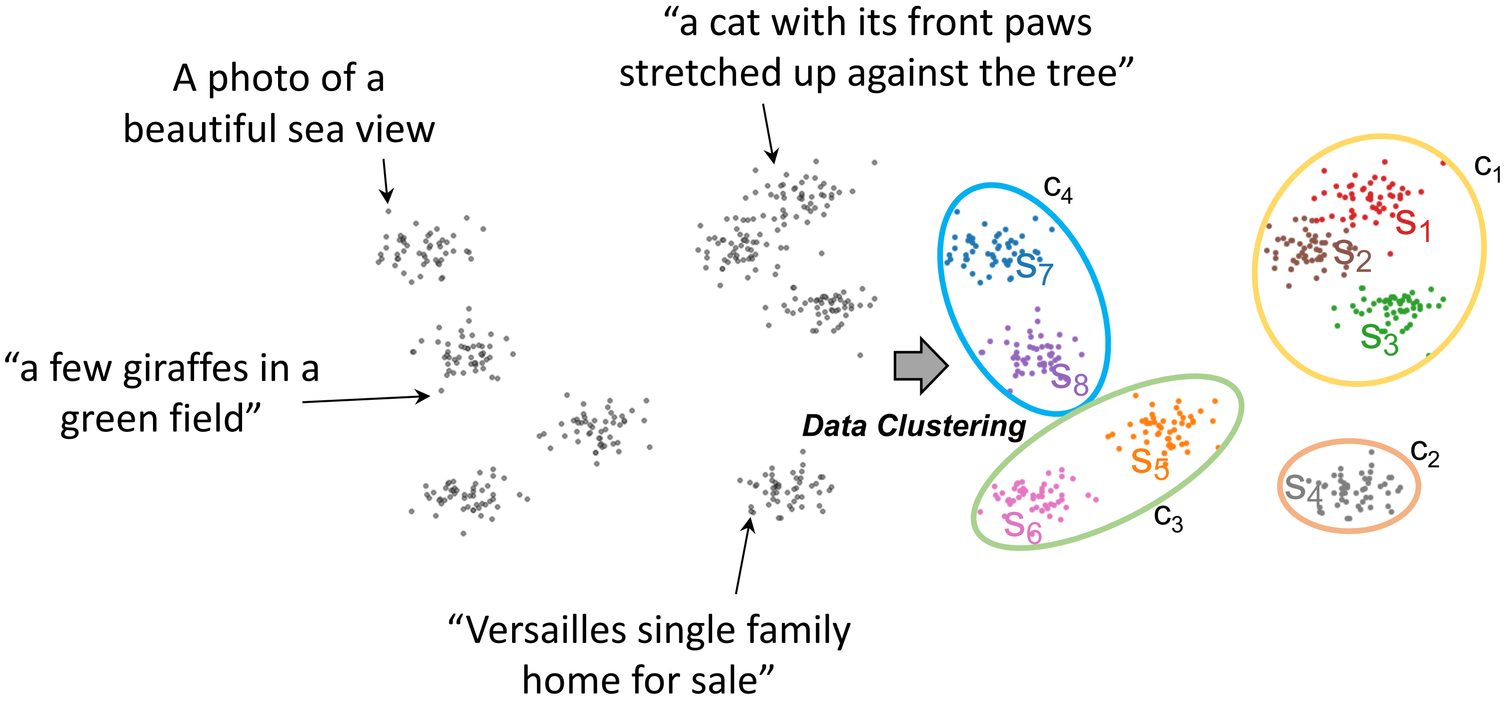
“a few giraffes in a green field”

“Versailles single family home for sale”

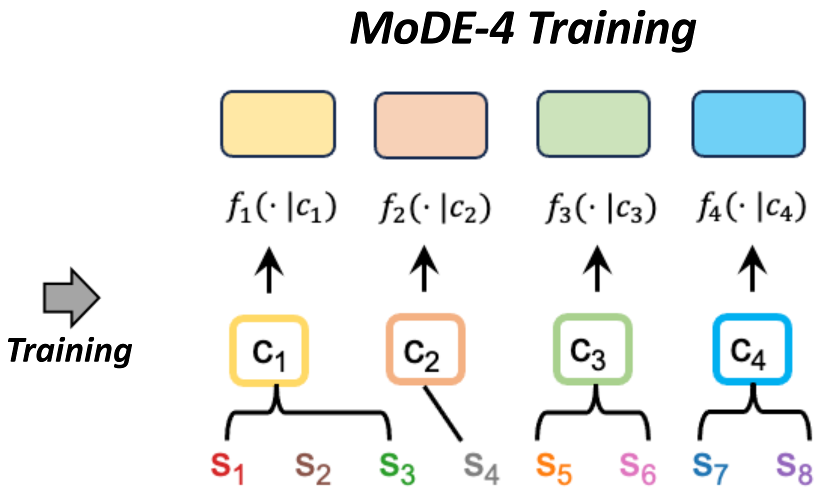
Language Embedding Space by SimCSE (EMNLP’21)

Learning Data Experts via Clustering

On each cluster, a model (termed as a Data Expert) is trained with more quality negative.


A photo of a beautiful sea view
"a cat with its front paws stretched up against the tree"
"a few giraffes in a green field"
"Versailles single family home for sale"
Data Clustering
c_4
c_1
c_3
c_2
S_1
S_2
S_3
S_4
S_5
S_6
S_7
S_8

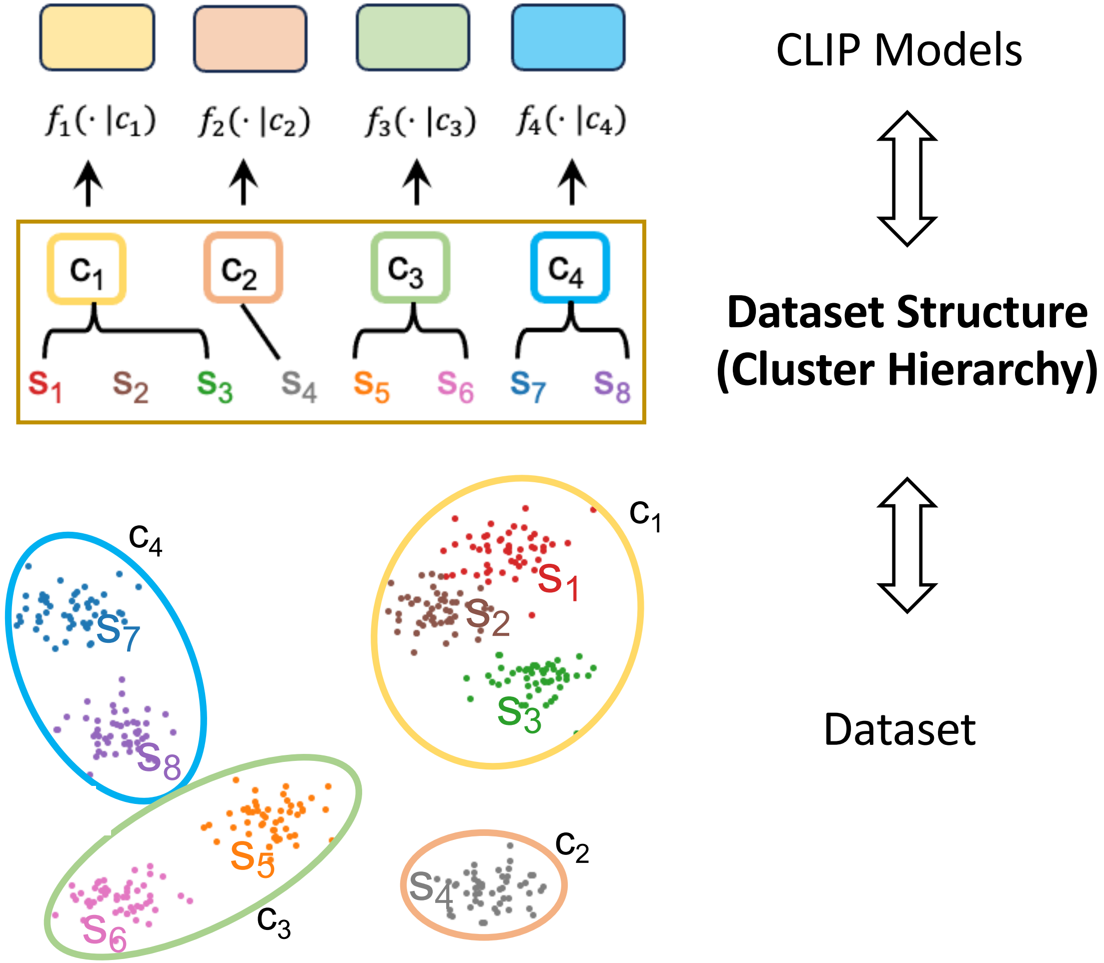
Language Embedding Space by SimCSE (EMNLP'21)



Data Expert: You Reap What You Sow

(Condition each model
with the training data.)

Represent Data Expertise via Fine-Grained Clusters



CLIP Models

Dataset Structure (Cluster Hierarchy)

Dataset

Condition Model Training

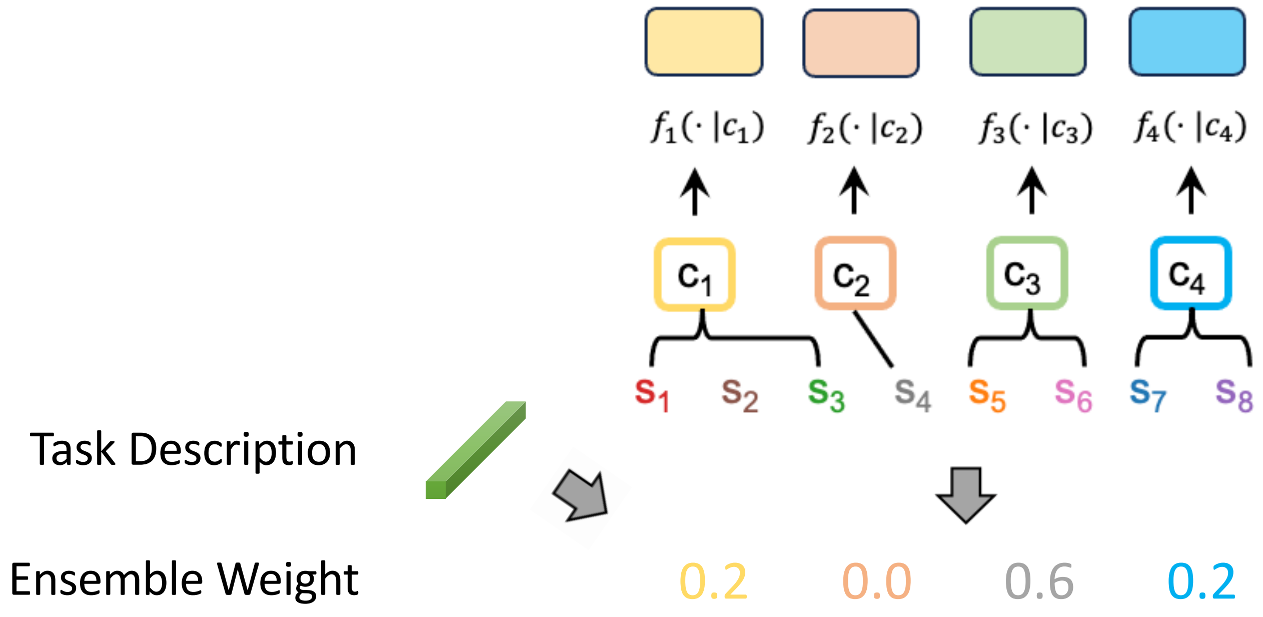
Cluster centers provide a global view of the full train set.

Inference-Time Task Adaptation

Use Cluster centers to guide the ensemble for multi-modal prediction.

$$\sum_{i=\{1,2,3,4\}} f(\cdot\,|c_i)p(c_i|\boldsymbol{T})$$

$$p(c_i|\boldsymbol{T}) \propto \sum_{l\in L} \boldsymbol{A}(l,c)$$



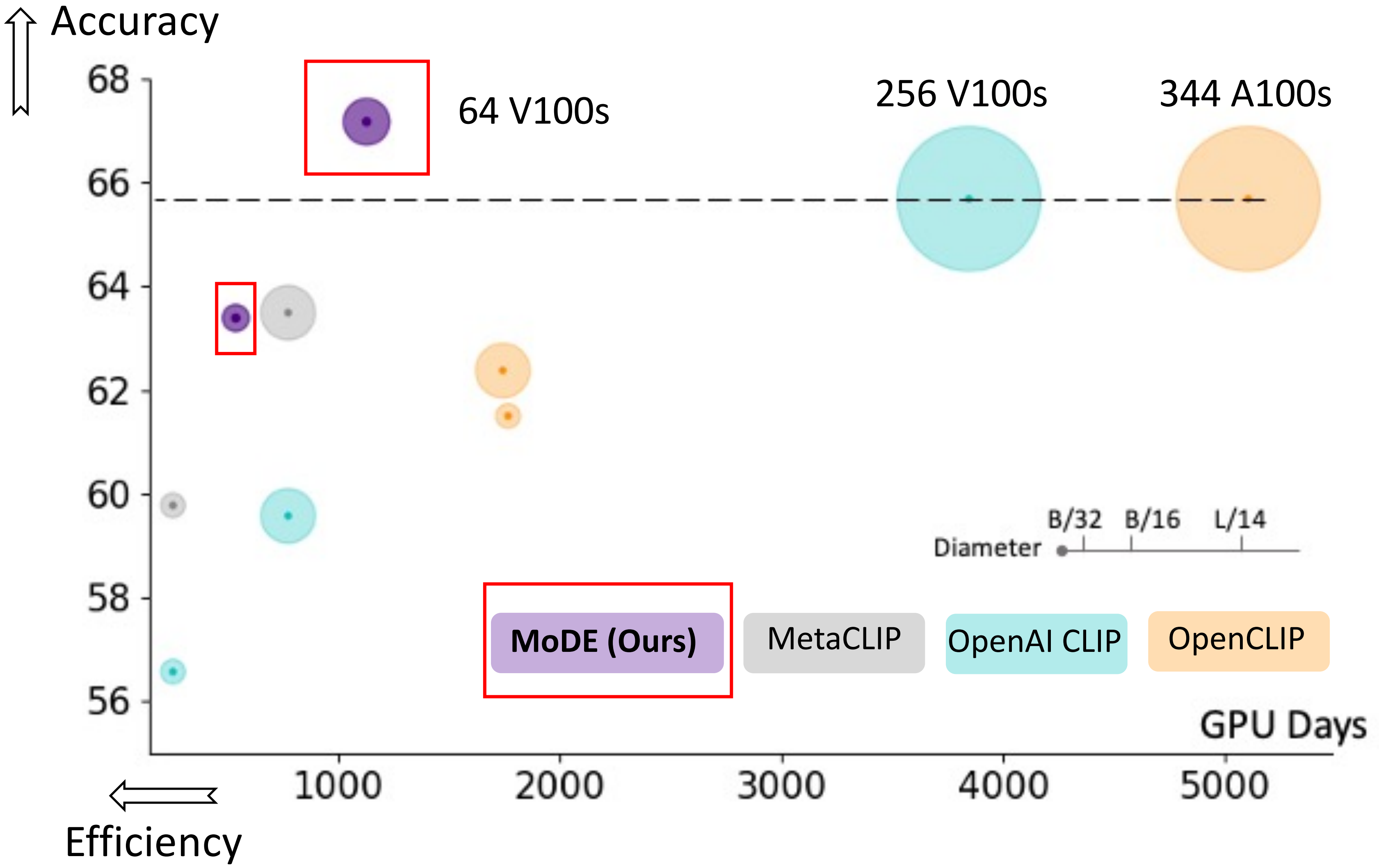
Task Description

Ensemble Weight 0.2 0.0 0.6 0.2

$\boldsymbol{T}$="ImageNet"
$\boldsymbol{L}$={‘dog’, ‘cat’, …, ’salmon’}

Efficiency & Effectiveness

CLIP Benchmark (26)
Zero-Shot Classification



MoDE Provides Strong Representation

ImageNet
Linear Probing on
Concatenated Feature

Approach	ViT-B/32	ViT-B/16	ViT-L/14
MetaCLIP	67.5	73.8	82.3
MoDE-2	71.3	76.9	83.9
MoDE-4	74.1	79.6	84.7

Summary of MoDE

Data Expert

- Deep Neural Network is naturally data-driven
- Use Data to explain the capability of a model

Mixture of Data Expert for CLIP

- Scale up the “width” of CLIP System
- MoDE offers both efficiency and effectiveness in CLIP training
- MoDE can be applied in different task types flexibly

MoDE: CLIP Data Experts via Clustering

Jiawei Ma, Po-Yao Huang, Saining Xie, Shang-Wen Li,
Luke Zettlemoyer, Shih-Fu Chang, Wen-Tau Yih, Hu Xu

Project: https://github.com/facebookresearch/MetaCLIP/tree/main/mode