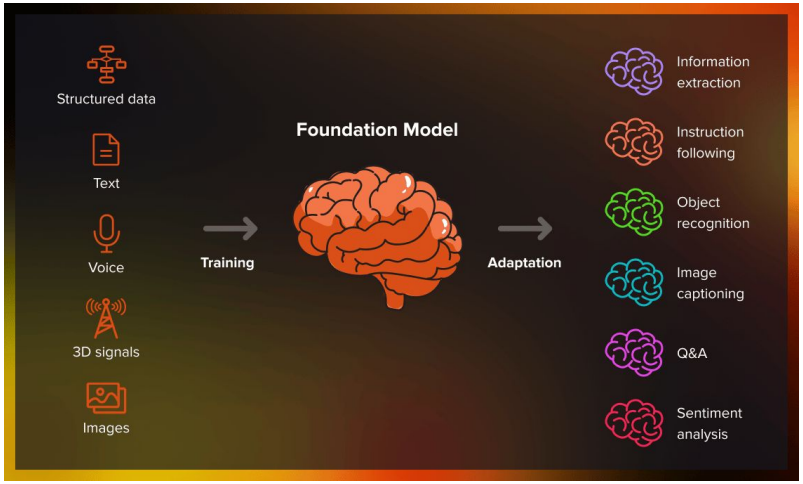
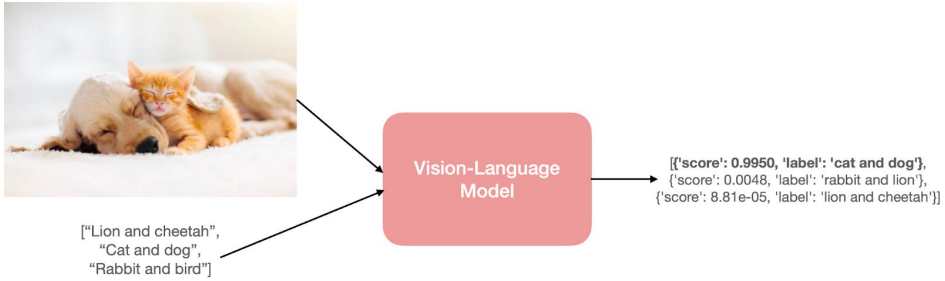
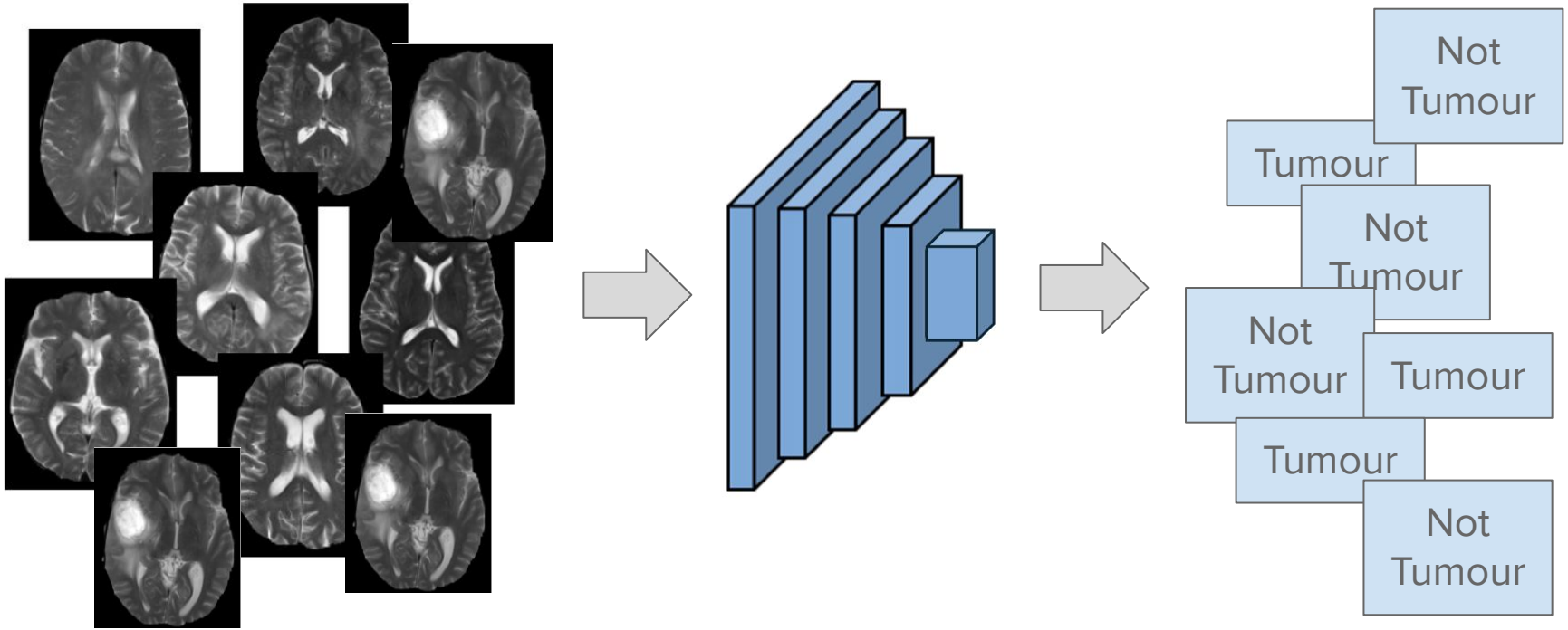
Advanced concepts in deep learning II



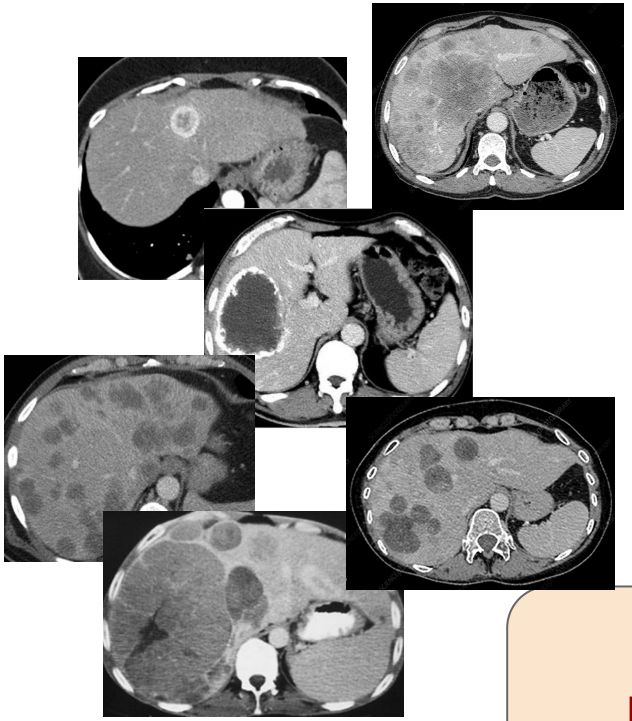


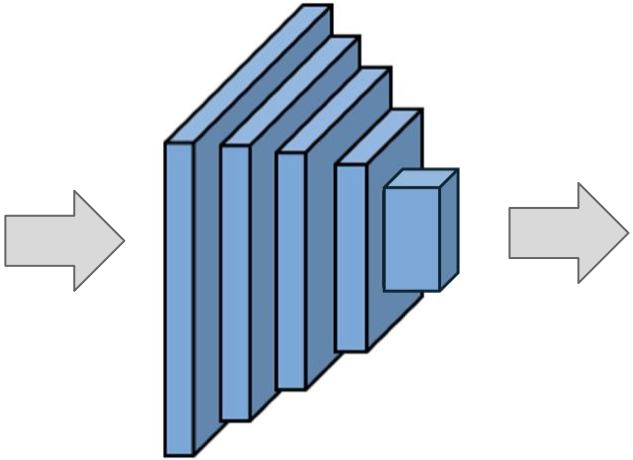
Jose Dolz

Traditional (task-specific) learning



Traditional (task-specific) learning

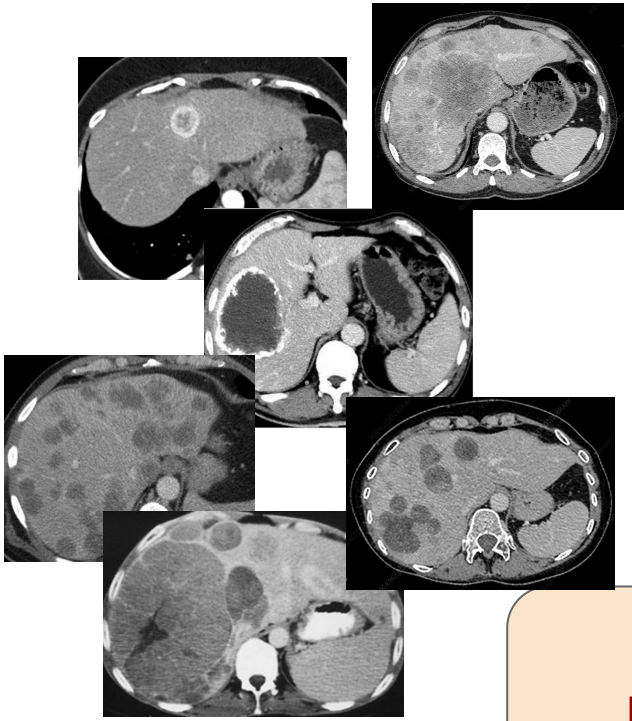


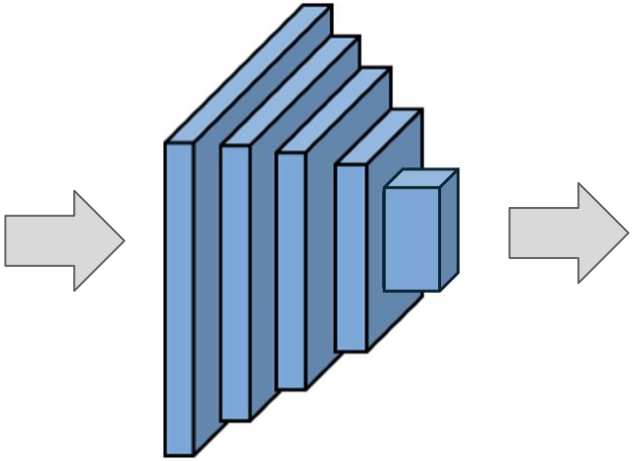


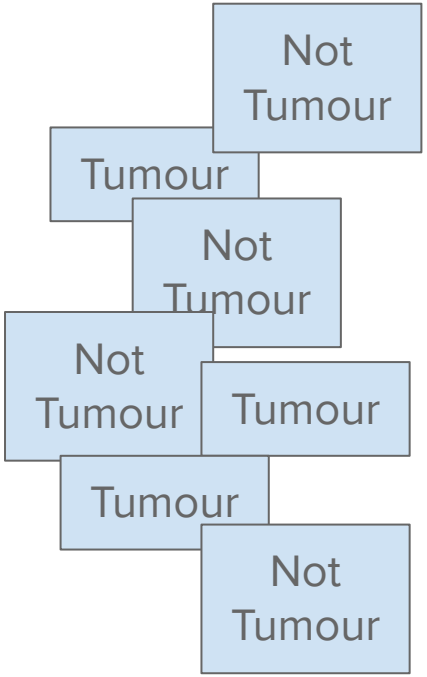
How to adapt to novel tasks?

Traditional (task-specific) learning

Train the model from *scratch*



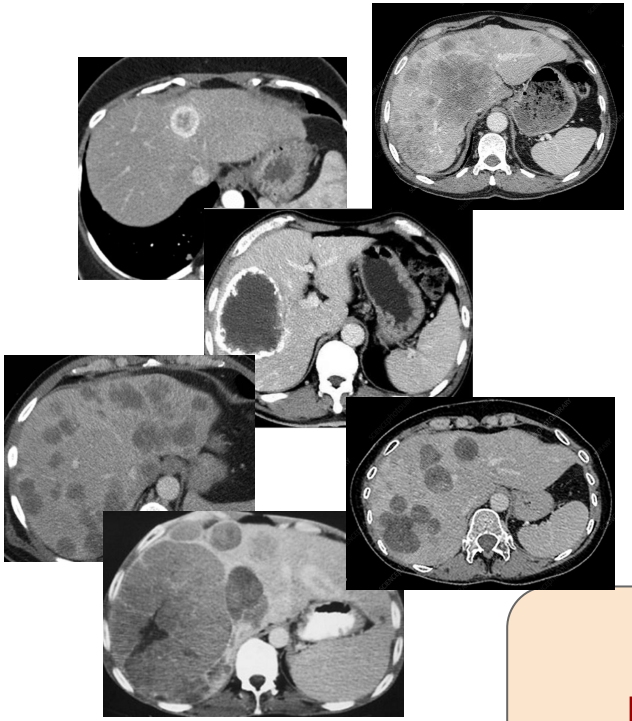


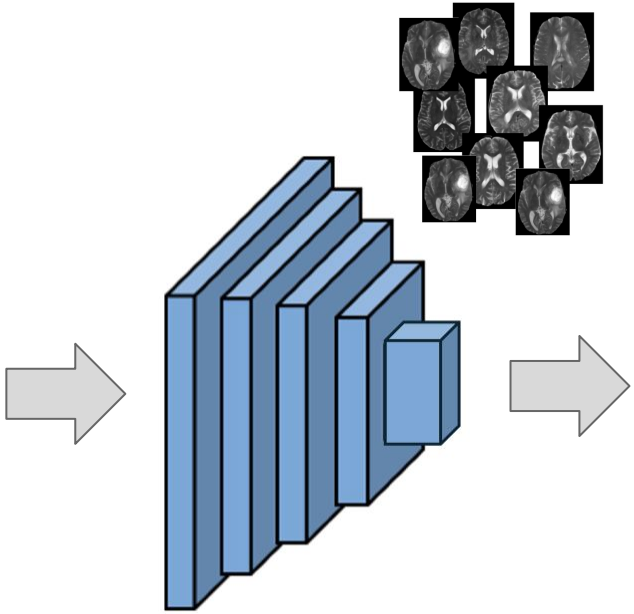


How to adapt to novel tasks?

Traditional (task-specific) learning

Adapt from a model pre-trained on another dataset





Not Tumour

Tumour

Not Tumour

Not Tumour

Tumour

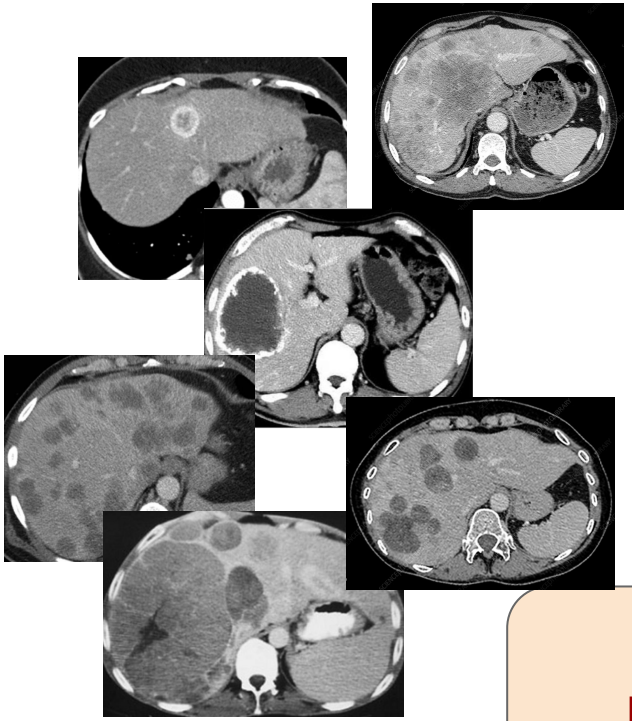
Tumour

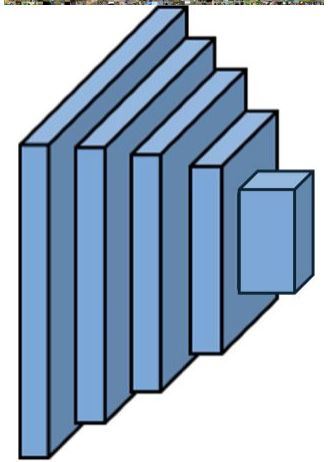
Not Tumour

How to adapt to novel tasks?

Traditional (task-specific) learning

Adapt from a model pre-trained on another dataset





Not Tumour

Tumour

Not Tumour

Not Tumour

Tumour

Tumour

Not Tumour

How to adapt to novel tasks?

Traditional (task-specific) learning

How to adapt to novel tasks?
Limitations

- A substantial amount of target samples need to be labeled

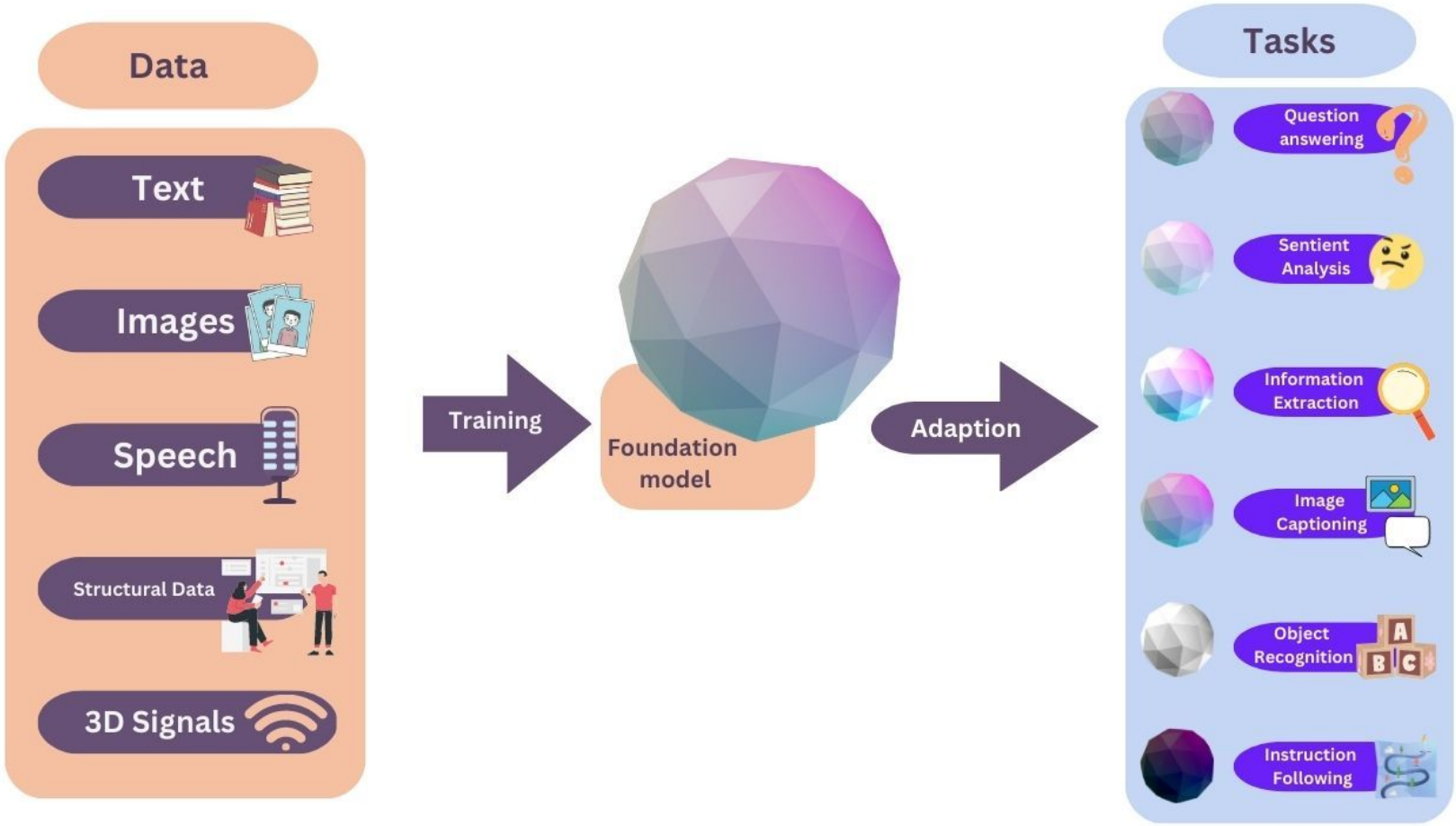
- Large scale labeled datasets may be significantly different, hindering transferability

Source

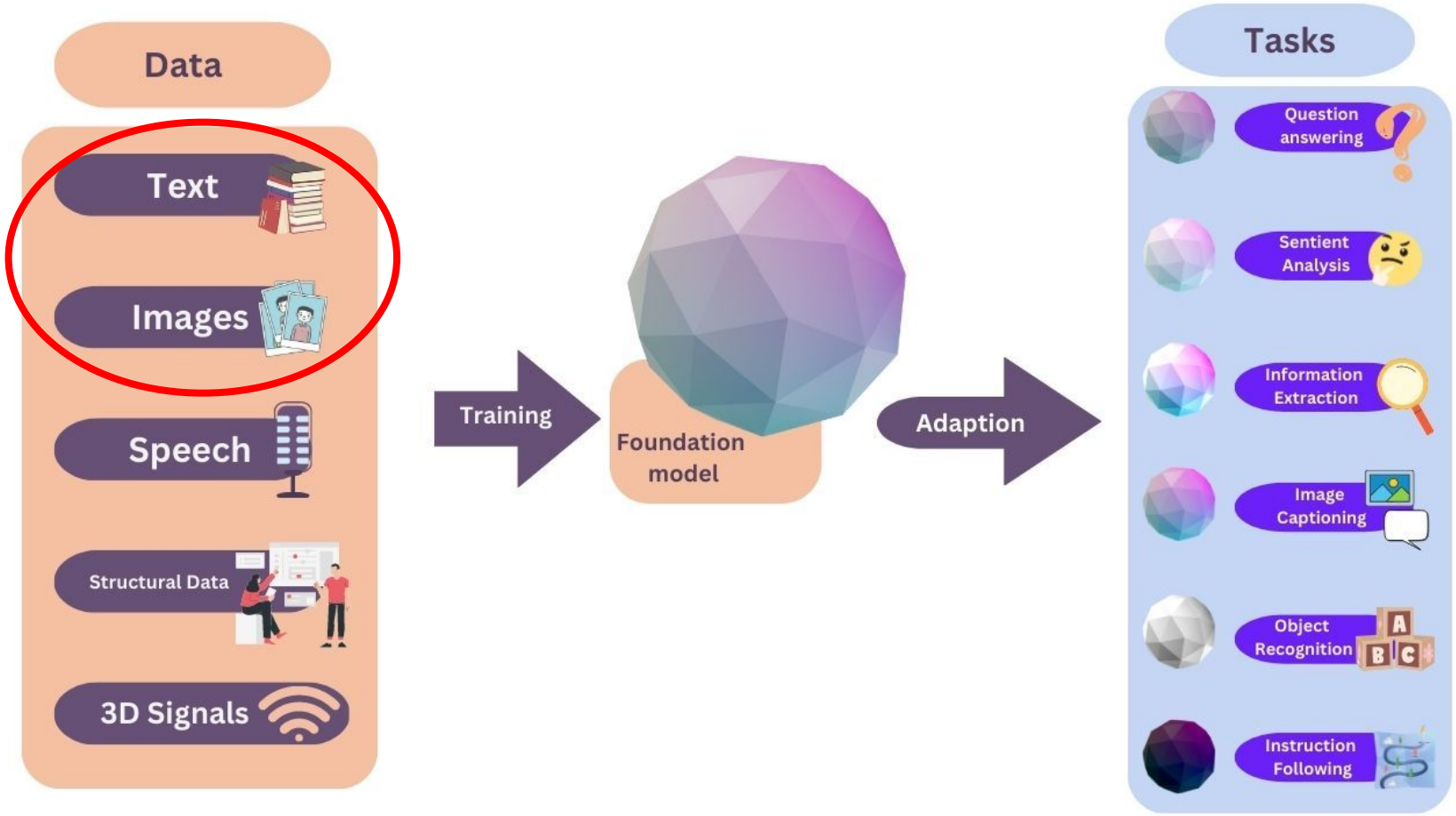
Target

- Adaptation still requires fine-tuning of the whole model.
  This increases the computational complexity.

What are foundation models?



What are foundation models?



Why to pair Images with Language?

- Language is a universal interface
- Language represents a natural way to describe the world (even for non specialists)
- It is highly semantic and rich
- It is domain-agnostic
- Many methods from NLP

“Car”

Why to pair Images with Language?

- Language is a universal interface
- Language represents a natural way to describe the world (even for non specialists)
- It is highly semantic a
- It is domain-agnostic
- Many methods from NLP

How to connect these two worlds?

“Car”

What is a language model?

DISCLAIMER

The following slides do not intend to be (not even close) a tutorial on language models

- A probabilistic model of a natural language

- It provides probabilities for a set of words

  "given all the words I have seen, which word can be **next**?"

  "To be, or not to be, that is the _____"

What is a language model?

DISCLAIMER

The following slides do not intend to be (not even close) a tutorial on language models

- A probabilistic model of a natural language

- It provides probabilities for a set of words

  "given all the words I have seen, which word can be **next**?"

  "To be, or not to be, that is the _question_ "

- It is trained on a plethora of text-corpus (Wikipedia, CNN, Oxford,etc)

- All in all, we can say that it mimics how people write/talk.

How text (words) can be used?

(in the era of deep learning)

DISCLAIMER

The following slides do not intend to be (not even close) a tutorial on language models

Feature #1

Feature #2

Feature #n

output/prediction

"To be or not to be that is the _____"

How text (words) can be used?

(in the era of deep learning)

DISCLAIMER

The following slides do not intend to be (not even close) a tutorial on language models

Feature #1

Feature #2

Feature #n

output/prediction

"To be or not to be that is the _____"

- We need an embedding of each word in a different space, which is typically extracted by other model.

How text (words) can be used?

(in the era of deep learning)

DISCLAIMER

The following slides do not intend to be (not even close) a tutorial on language models

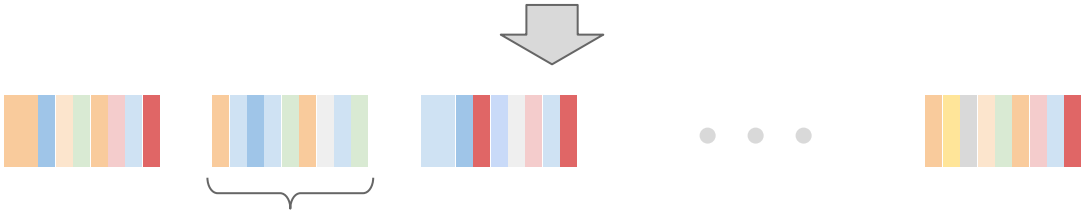
Feature #1

Feature #2

Feature #n

output/prediction

"To be or not to be that is the _____"



Individual d-dimensional embedding

Embedding space

- We need an embedding of each word in a different space, which is typically extracted by other model.

How text (words) can be used?

(in the era of deep learning)

DISCLAIMER

The following slides do not intend to be (not even close) a tutorial on language models

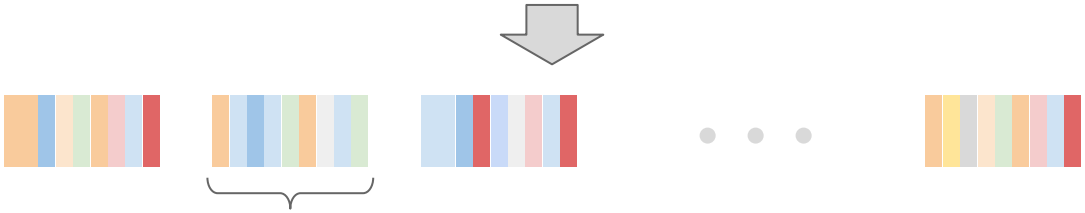
Feature #1

Feature #2

Feature #n

output/prediction

"To be or not to be that is the _____"



Individual d-dimensional embedding

Embedding space

Transformers learn their own embeddings

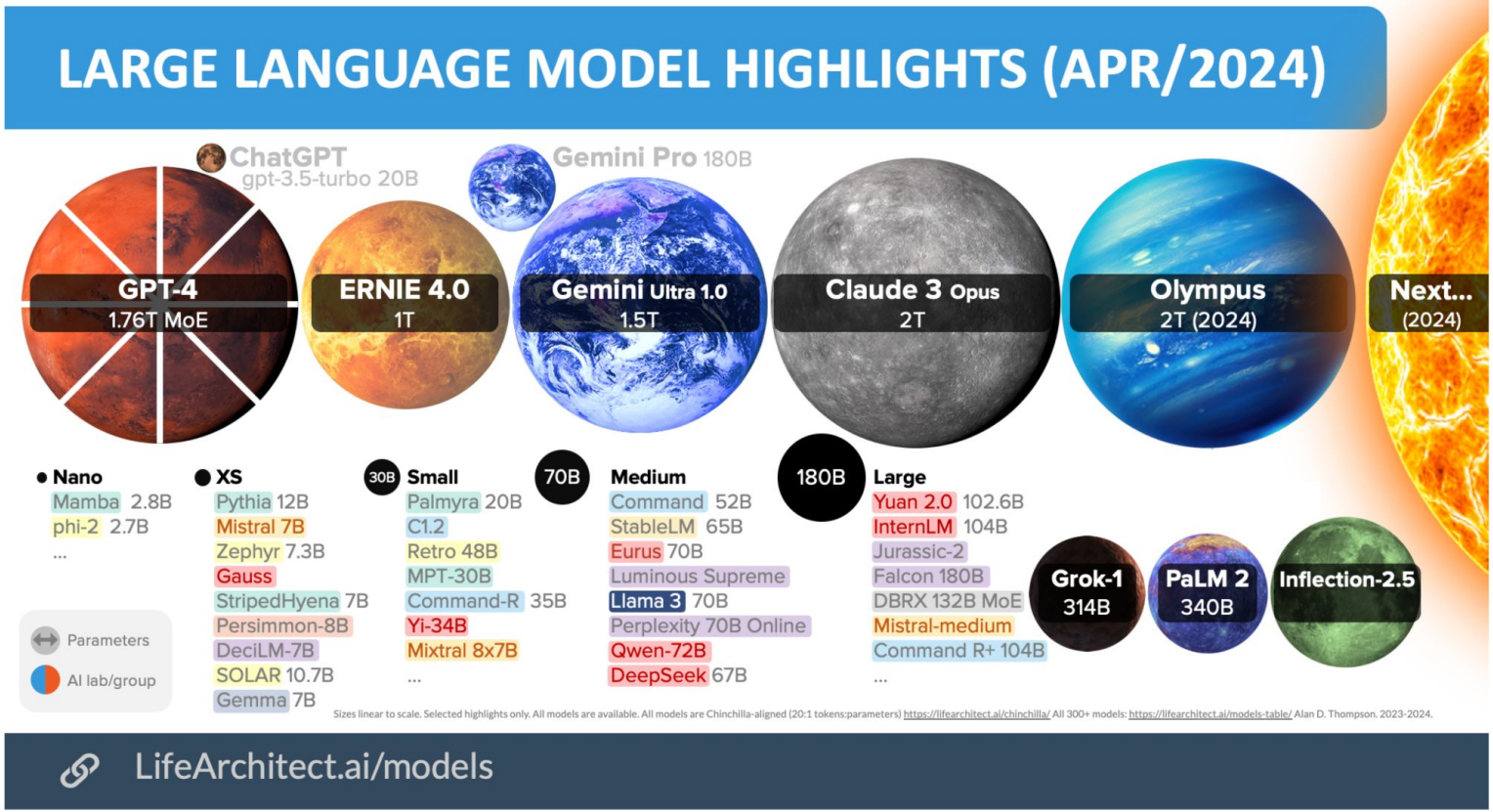
- We need an embedding of each word in a different space, which is typically extracted by other model.

How text (words) can be used?

(in the era of deep learning)

DISCLAIMER

The following slides do not intend to be (not even close) a tutorial on language models



How text (words) can be used?

(in the era of deep learning)

DISCLAIMER

The following slides do not intend to be (not even close) a tutorial on language models

To keep it simple:

1. Original text is split into independent tokens

"Hello world!"

How text (words) can be used?

(in the era of deep learning)

DISCLAIMER

The following slides do not intend to be (not even close) a tutorial on language models

To keep it simple:

1. Original text is split into independent tokens

"Hello world!"

2. These tokens are assigned a unique ID (one ID = one word in a dictionary)

['Hello', 'world','!']

[1951, 3023, 134]

How text (words) can be used?

(in the era of deep learning)

DISCLAIMER

The following slides do not intend to be (not even close) a tutorial on language models

To keep it simple:

1. Original text is split into independent tokens

"Hello world!"

2. These tokens are assigned a unique ID (one ID = one word in a dictionary)

['Hello', 'world','!']

[1951, 3023, 134]

3. Each ID corresponds to an embedding (e.g., 512-d vector) (the feature embedding is **learned**)

1951 = [.....]

Embedding of the word 'Hello'

Revisit language-vision models

(the main topic of this talk)

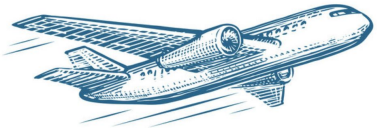
Main idea

Visual domain

Language domain

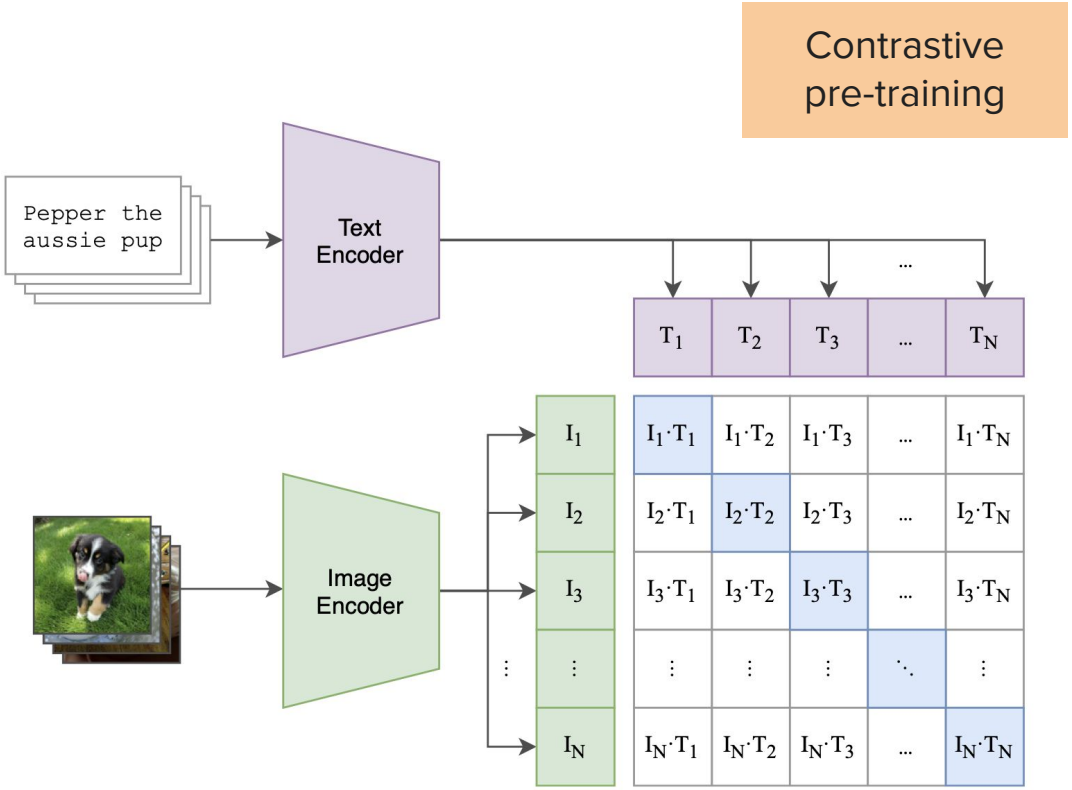
"An orange sports car"

"A photo of a dog"



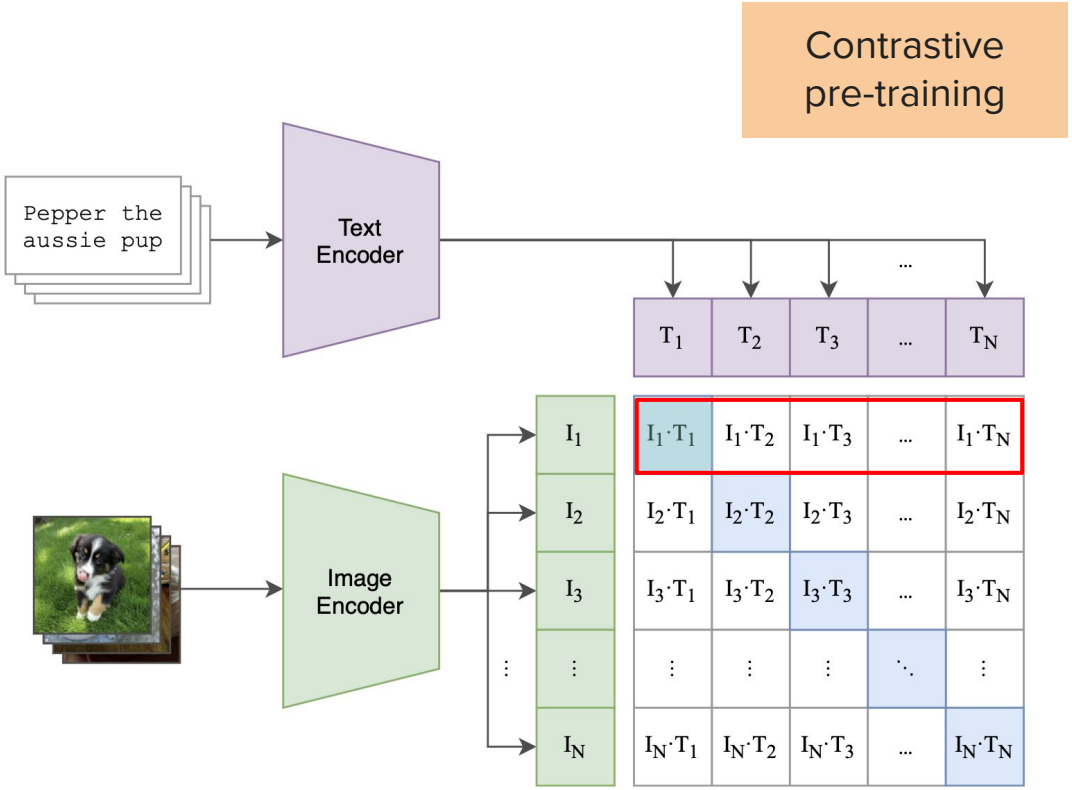
"A sketch of an aircraft"

Contrastive Language-Image Pre-training (CLIP)



Radford et *al*. Learning transferable visual models from natural language supervision. ICML'21

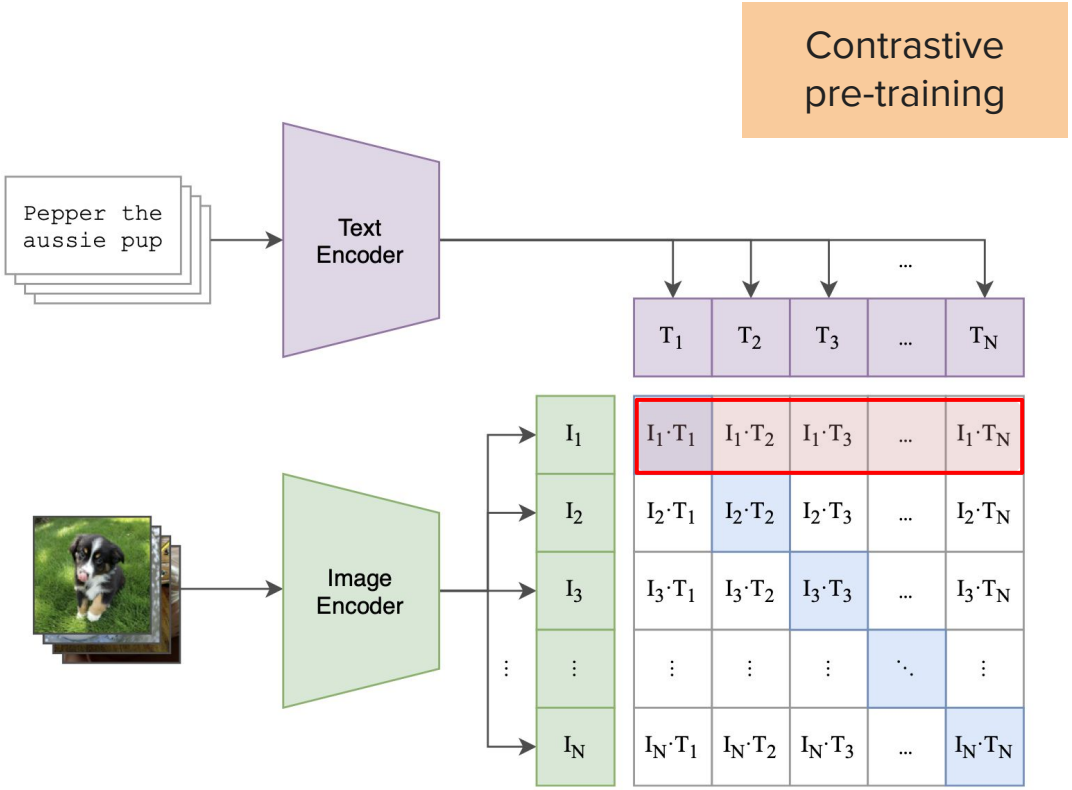
Contrastive Language-Image Pre-training (CLIP)



$$\hat{y}^{v}_{i,j} = \frac{\exp\left(\mathbf{v}_i \cdot \mathbf{z}_j^\top / \tau\right)}{\sum_{n=1}^{N} \exp\left(\mathbf{v}_i \cdot \mathbf{z}_n^\top / \tau\right)}$$

Radford et *al*. Learning transferable visual models from natural language supervision. ICML'21

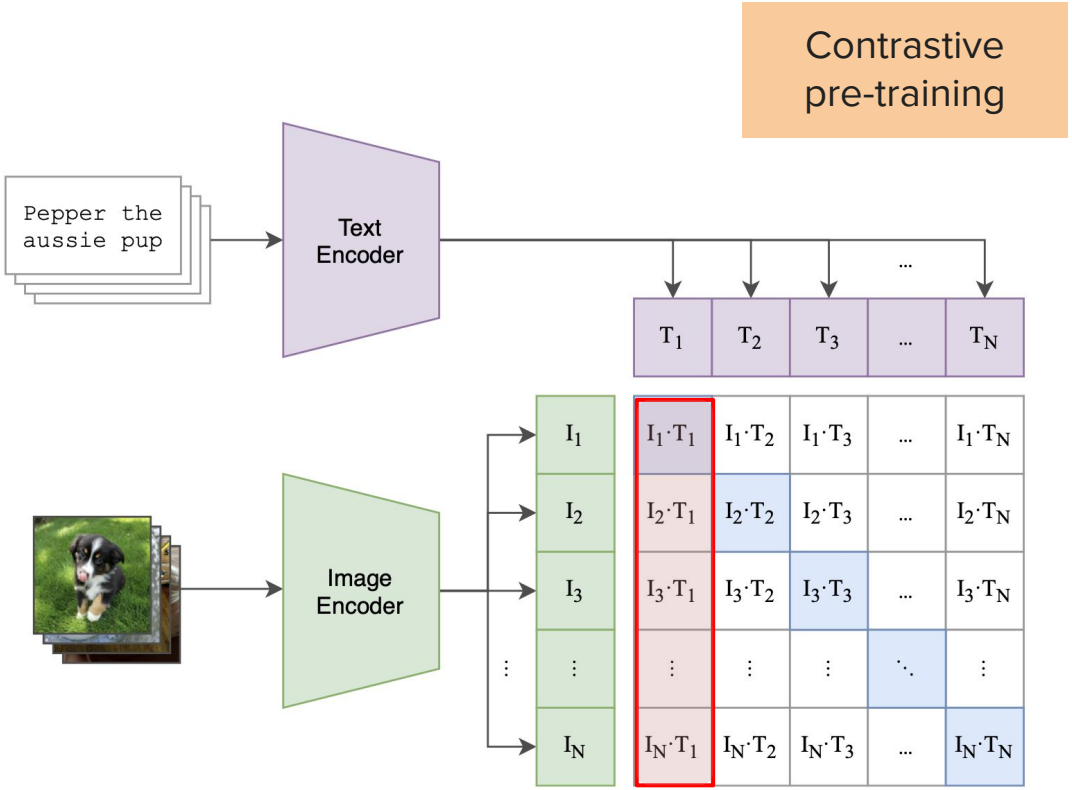
Contrastive Language-Image Pre-training (CLIP)



$$\hat{y}^v_{i,j} = \frac{\exp\left(\mathbf{v}_i \cdot \mathbf{z}_j^\top / \tau\right)}{\sum_{n=1}^{N} \exp\left(\mathbf{v}_i \cdot \mathbf{z}_n^\top / \tau\right)}$$

Radford et *al*. Learning transferable visual models from natural language supervision. ICML'21

Contrastive Language-Image Pre-training (CLIP)

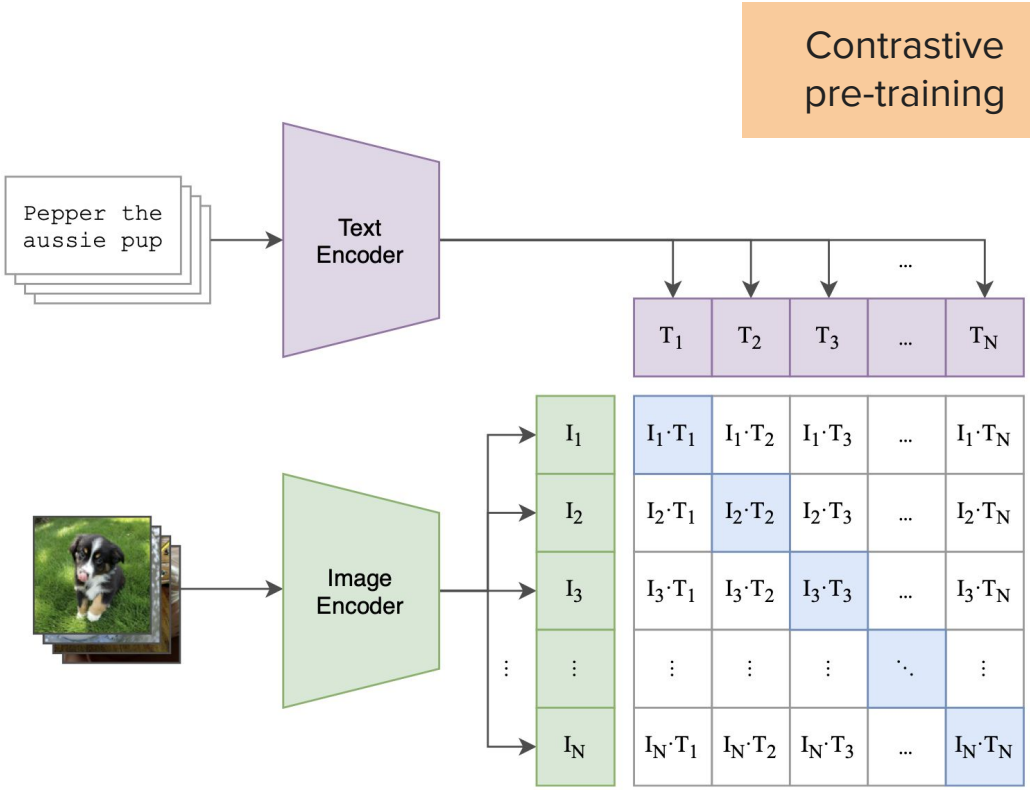


$$\hat{y}^{v}_{i,j} = \frac{\exp\left(\mathbf{v}_i \cdot \mathbf{z}_j^\top / \tau\right)}{\sum_{n=1}^{N} \exp\left(\mathbf{v}_i \cdot \mathbf{z}_n^\top / \tau\right)}$$

$$\hat{y}^{t}_{j,i} = \frac{\exp\left(\mathbf{v}_j \cdot \mathbf{z}_i^\top / \tau\right)}{\sum_{i=1}^{N} \exp\left(\mathbf{v}_j \cdot \mathbf{z}_i^\top / \tau\right)}$$

Radford et *al*. Learning transferable visual models from natural language supervision. ICML'21

Contrastive Language-Image Pre-training (CLIP)



$$\hat{y}^v_{i,j} = \frac{\exp\left(\mathbf{v}_i \cdot \mathbf{z}_j^\top / \tau\right)}{\sum_{n=1}^{N} \exp\left(\mathbf{v}_i \cdot \mathbf{z}_n^\top / \tau\right)}$$

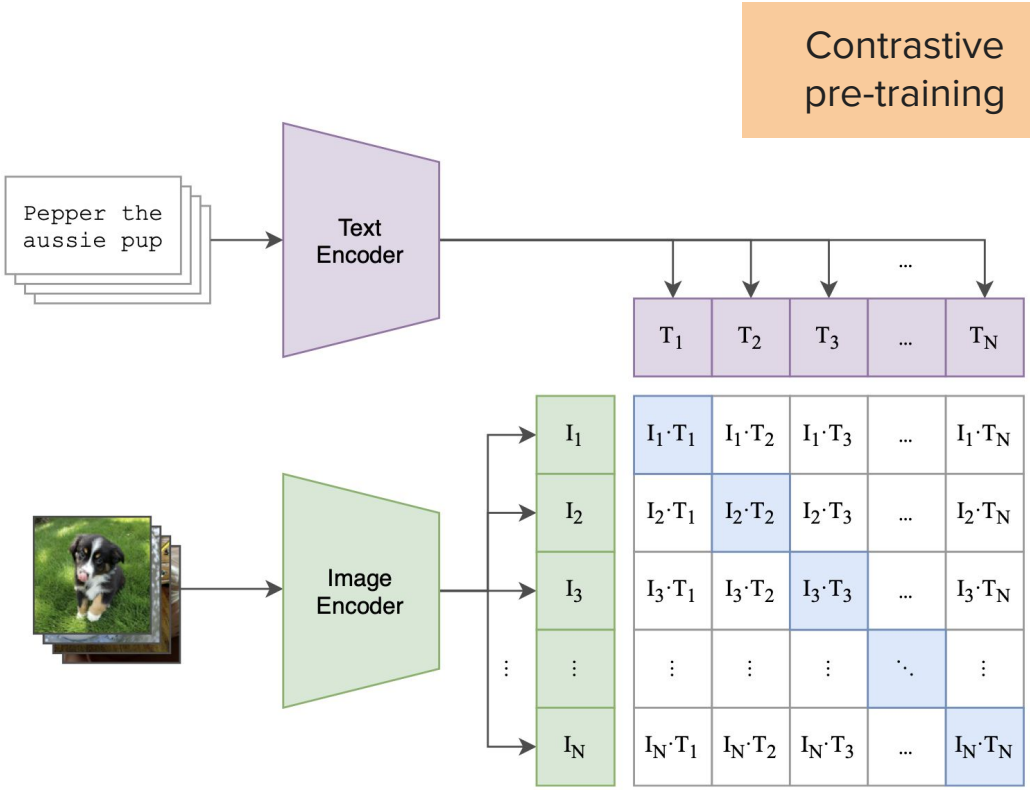
$$\hat{y}^t_{j,i} = \frac{\exp\left(\mathbf{v}_j \cdot \mathbf{z}_i^\top / \tau\right)}{\sum_{i=1}^{N} \exp\left(\mathbf{v}_j \cdot \mathbf{z}_i^\top / \tau\right)}$$

Ground truth for *i*-th element

$$\mathbf{y}_i = [0, 0, 1, ..., 0]$$

Radford et *al*. Learning transferable visual models from natural language supervision. ICML'21

Contrastive Language-Image Pre-training (CLIP)



$$\hat{y}^v_{i,j} = \frac{\exp\left(\mathbf{v}_i \cdot \mathbf{z}_j^\top / \tau\right)}{\sum_{n=1}^{N} \exp\left(\mathbf{v}_i \cdot \mathbf{z}_n^\top / \tau\right)}$$

$$\hat{y}^t_{j,i} = \frac{\exp\left(\mathbf{v}_j \cdot \mathbf{z}_i^\top / \tau\right)}{\sum_{i=1}^{N} \exp\left(\mathbf{v}_j \cdot \mathbf{z}_i^\top / \tau\right)}$$

Ground truth for *i*-th element

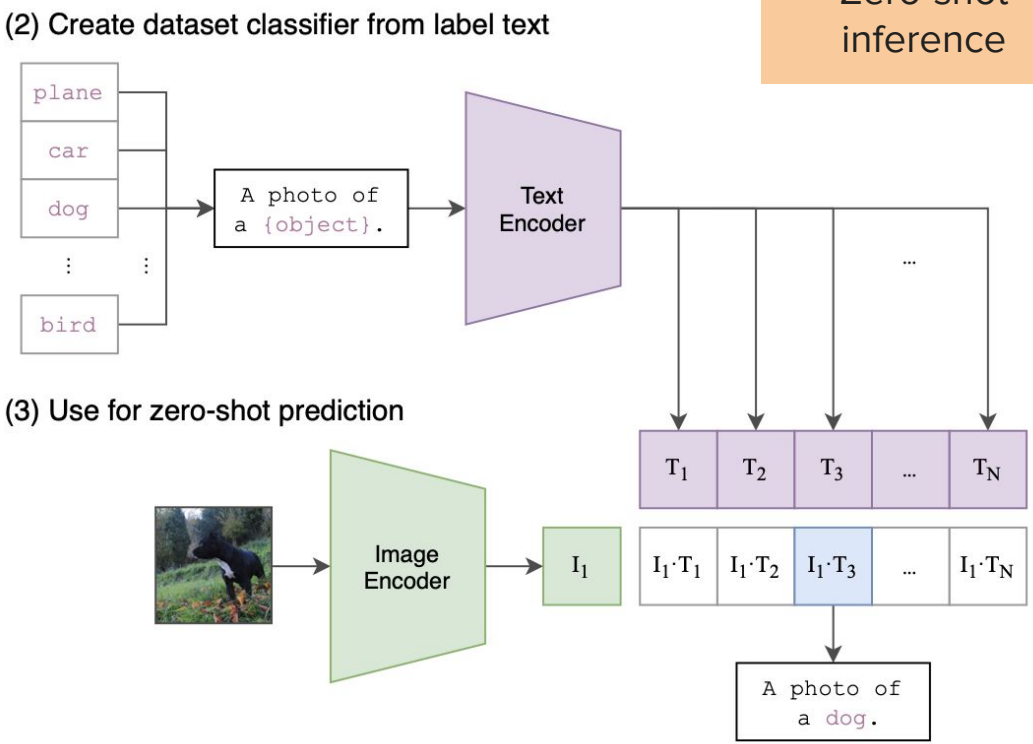
$$\mathbf{y}_i = [0, 0, 1, ..., 0]$$

$$\mathcal{L} = \frac{1}{2} \sum_{i=1}^{N} \left(\mathcal{H}(\mathbf{y}_i, \hat{\mathbf{y}}^v_i) + \mathcal{H}(\mathbf{y}_i, \hat{\mathbf{y}}^t_i)\right)$$

Radford et *al*. Learning transferable visual models from natural language supervision. ICML'21

Contrastive Language-Image Pre-training (CLIP)

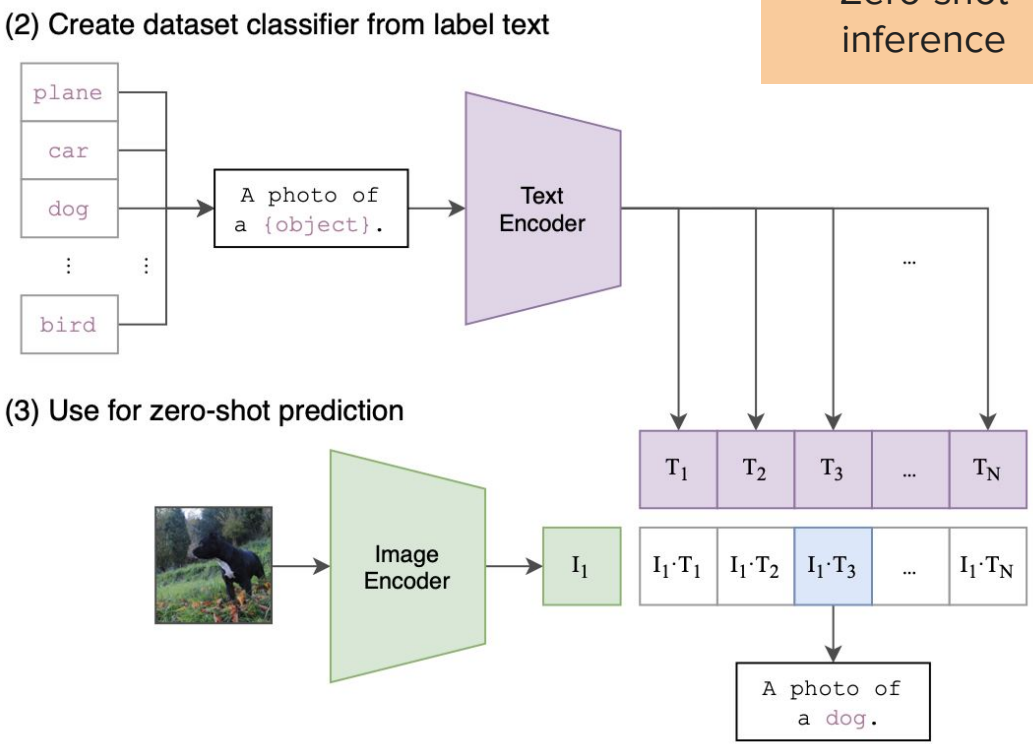
Zero-shot inference



Radford et *al*. Learning transferable visual models from natural language supervision. ICML'21

Contrastive Language-Image Pre-training (CLIP)

Zero-shot inference

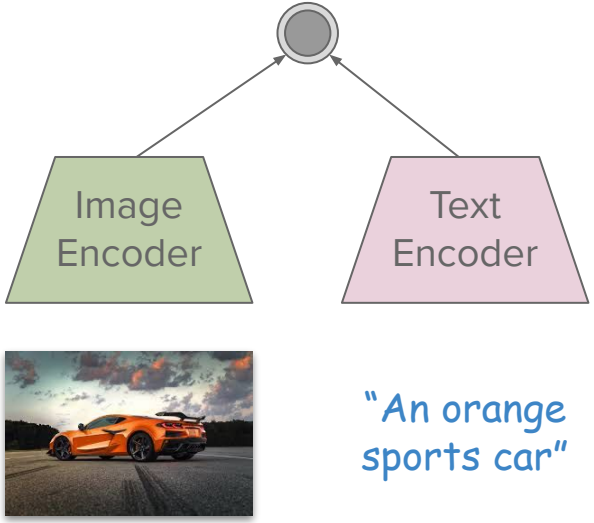


For a given class c

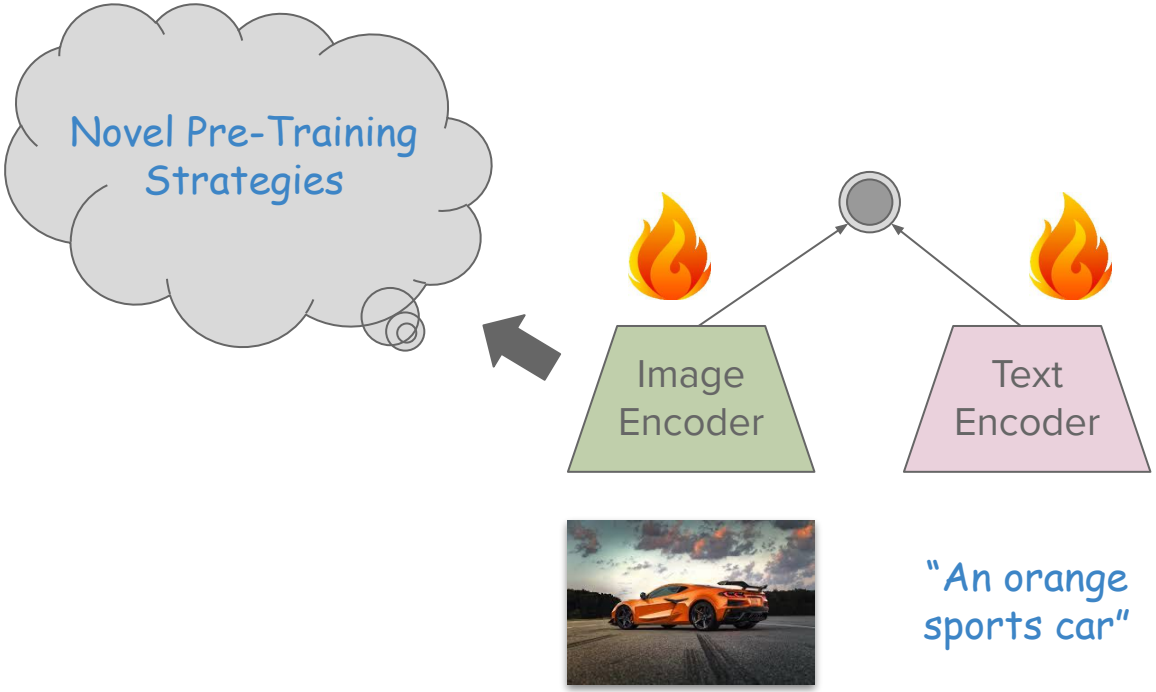
$$\hat{y}_c = \frac{\exp(\mathbf{v} \cdot \mathbf{z}_c^\top / \tau)}{\sum_{k=1}^{K} \exp(\mathbf{v} \cdot \mathbf{z}_k^\top / \tau)}$$

Radford et *al*. Learning transferable visual models from natural language supervision. ICML'21

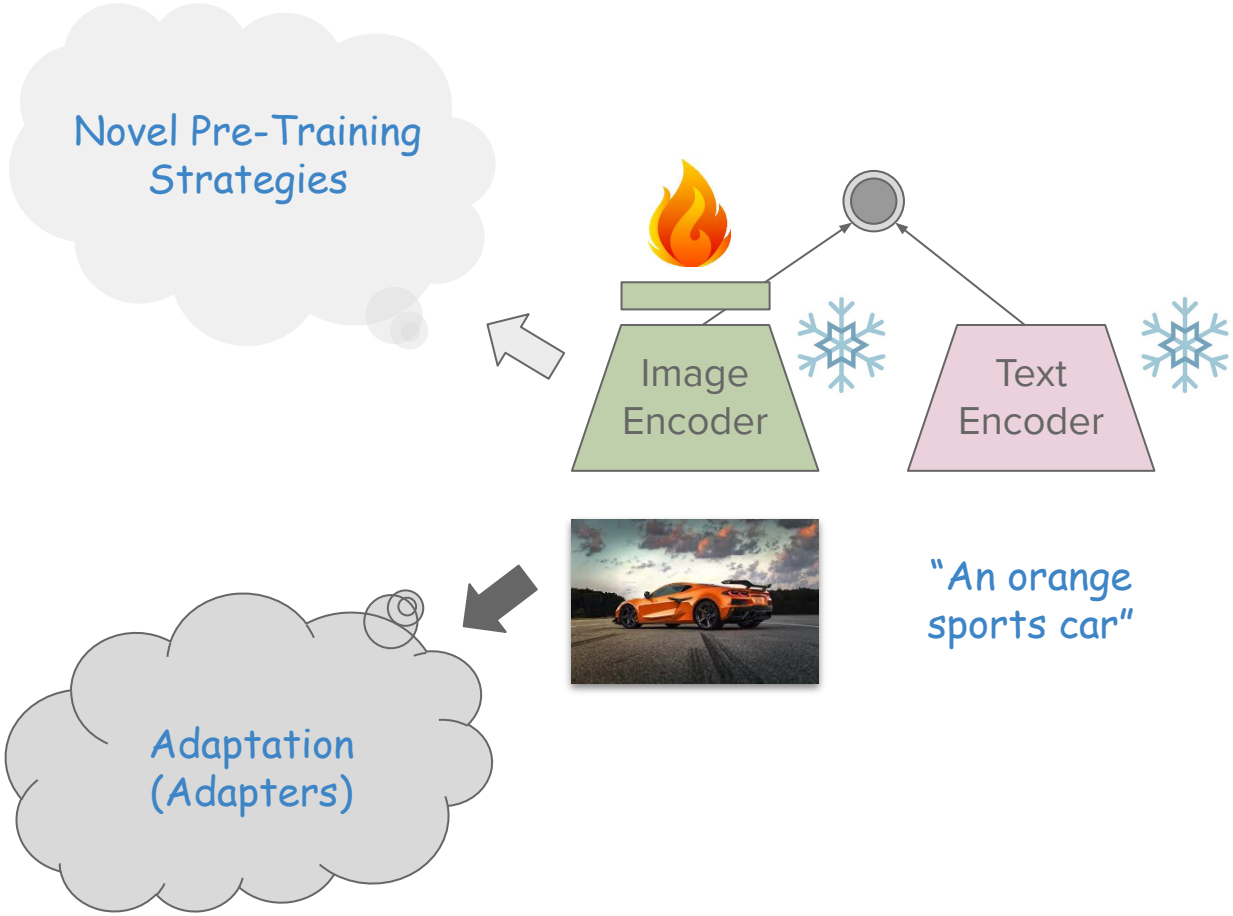
Where we go from CLIP?



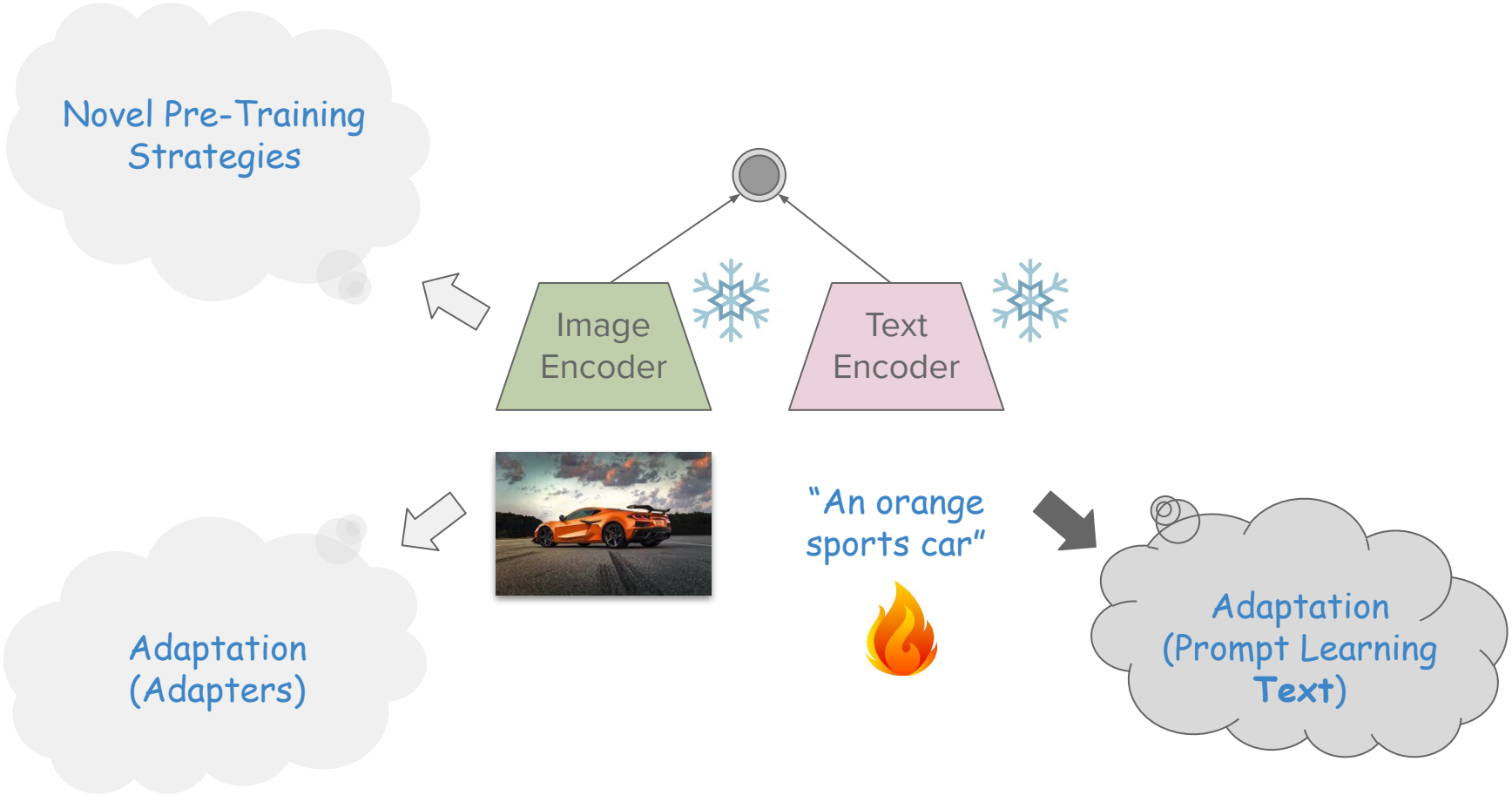
Where we go from CLIP?



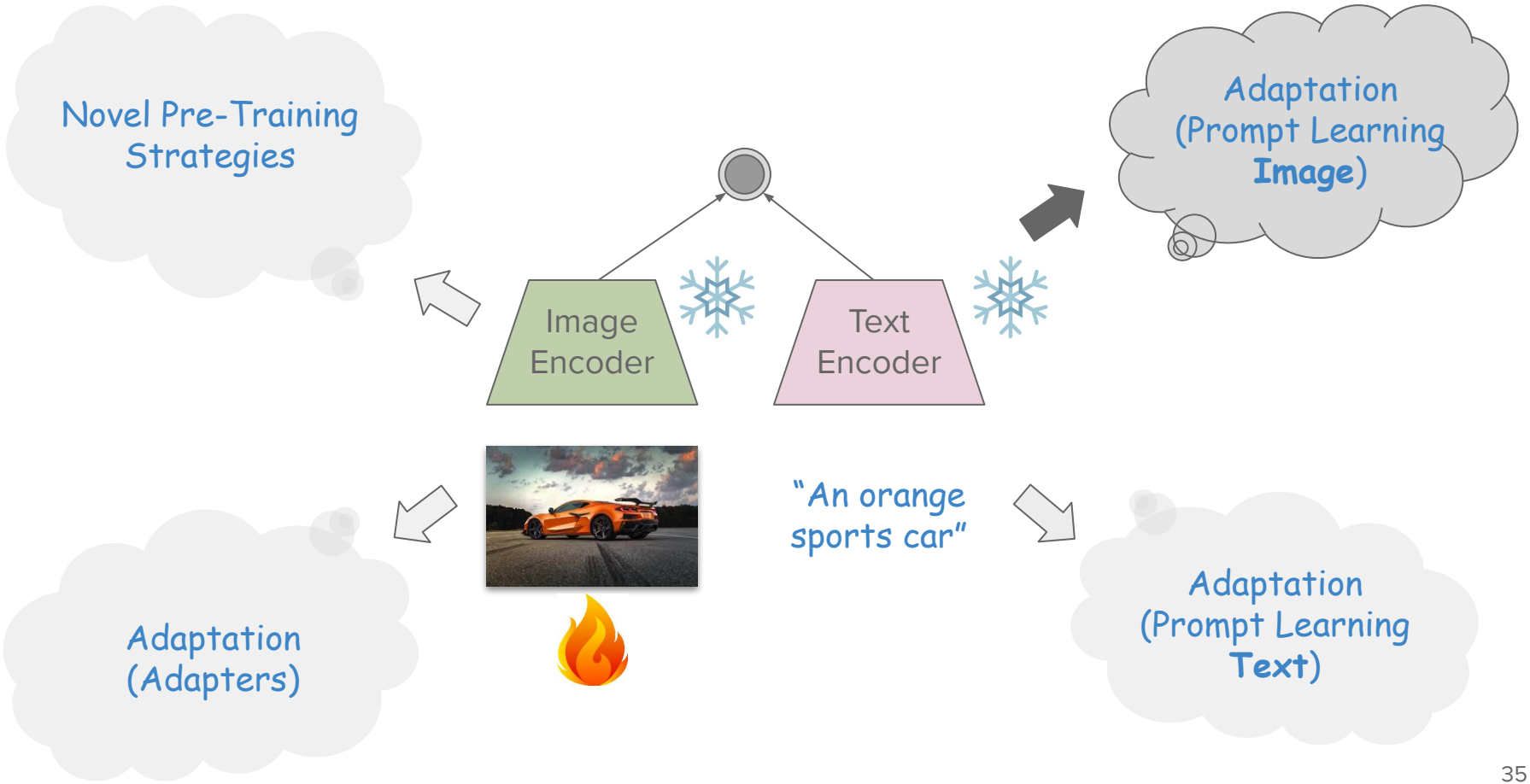
Where we go from CLIP?



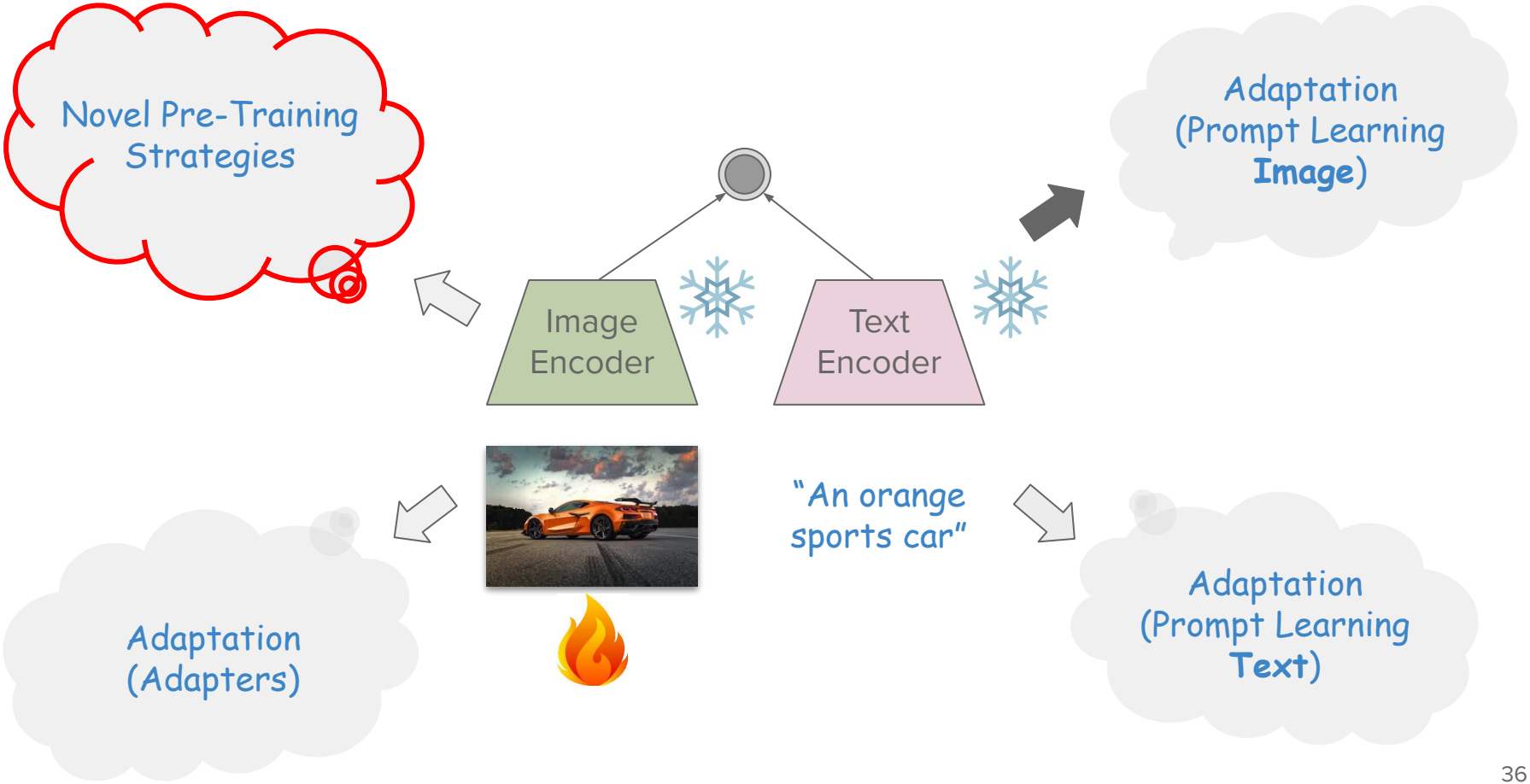
Where we go from CLIP?



Where we go from CLIP?

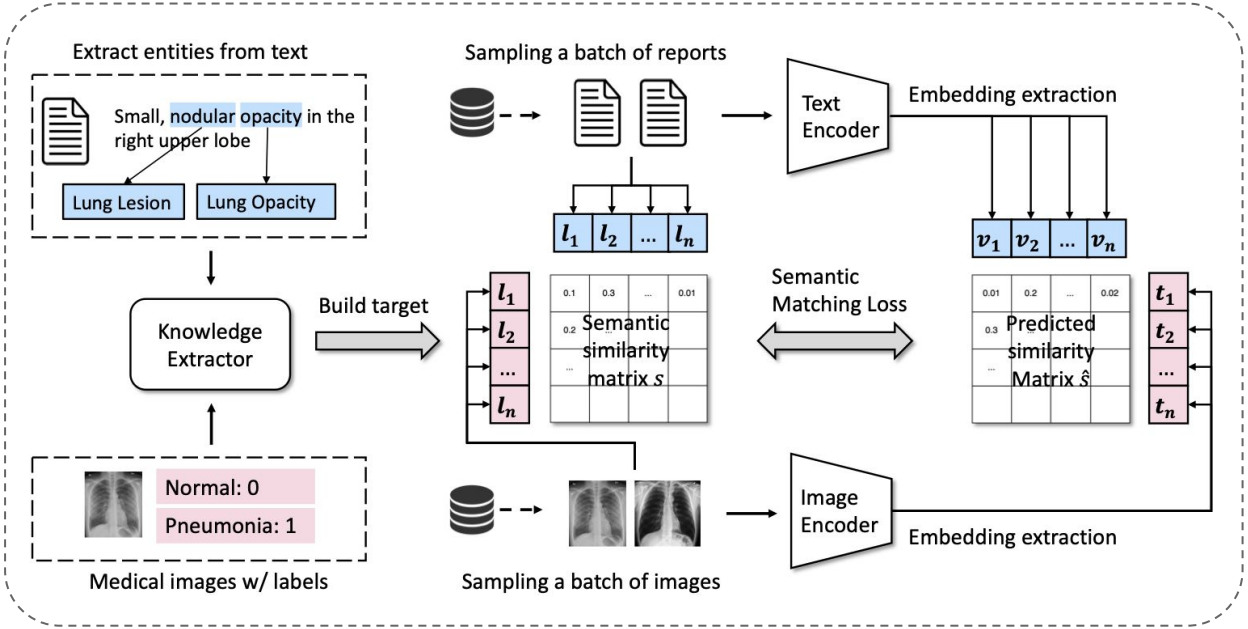


Where we go from CLIP?



Improving the pre-training

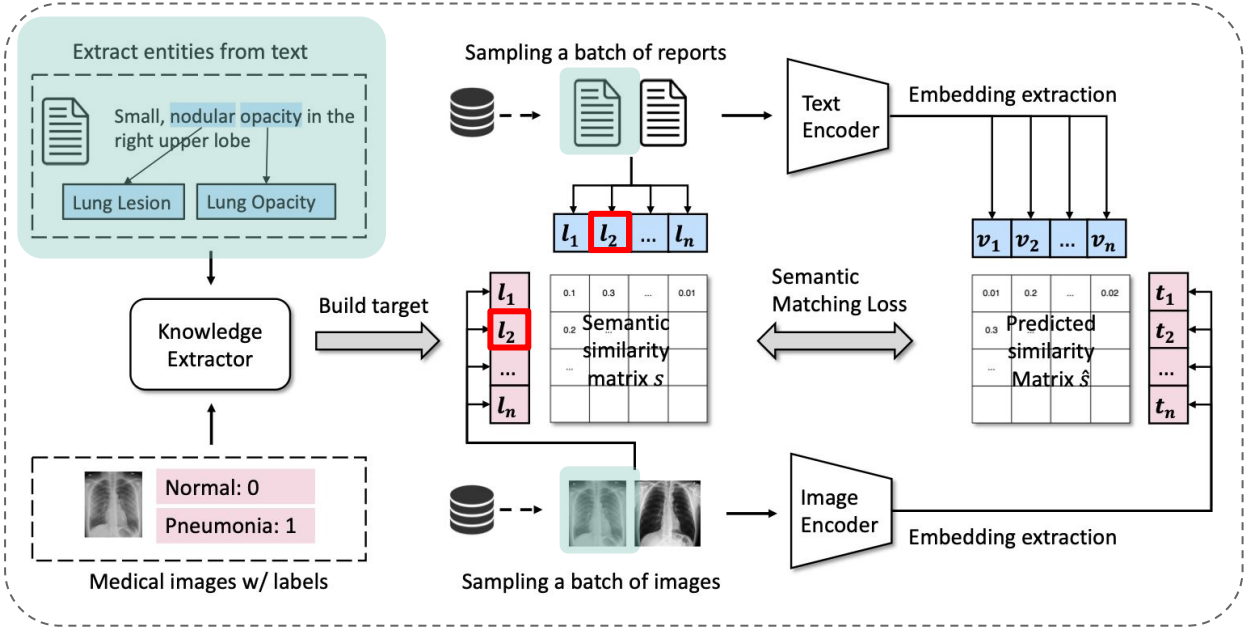
MedCLIP



Wang et *al*. Contrastive Learning from Unpaired Medical Images and Text. EMNLP'22

Improving the pre-training

MedCLIP



For a given image-report pair

Use the rapport to extract entities

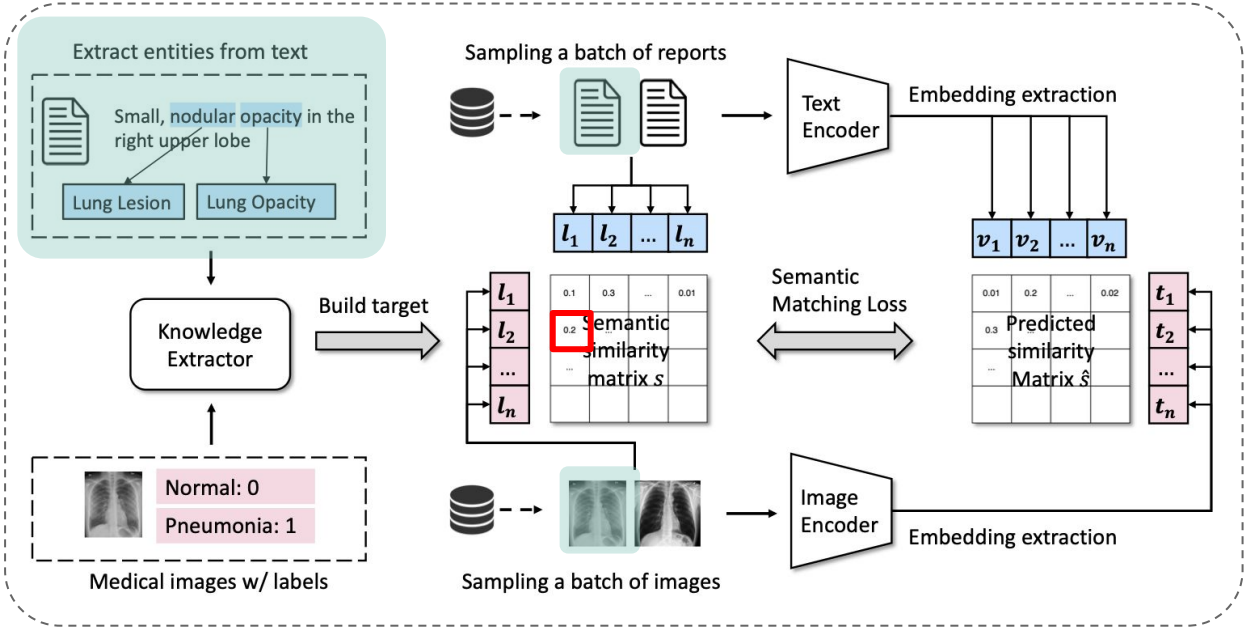
$$\mathbf{l}_i = [0, 1, 0, 0, 1, \ldots, 1]$$

Multi-label one-hot encoding

Wang et *al*. Contrastive Learning from Unpaired Medical Images and Text. EMNLP'22

Improving the pre-training

MedCLIP



Similarity between image and text multi-label encodings

$$s = \frac{\mathbf{l}_{img}^{\top} \cdot \mathbf{l}_{txt}}{\|\mathbf{l}_{img}\| \cdot \|\mathbf{l}_{txt}\|}$$

For a given image-report pair

Use the rapport to extract entities

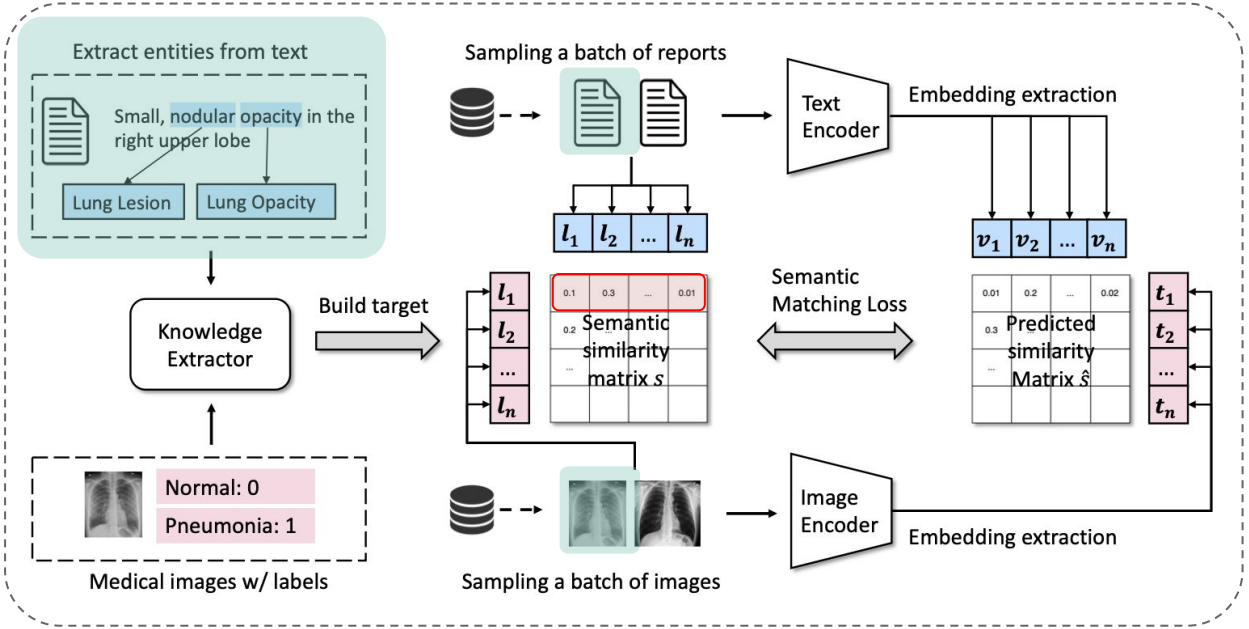
$$\mathbf{l}_i = [0, 1, 0, 0, 1, ..., 1]$$

Multi-label one-hot encoding

Wang et *al*. Contrastive Learning from Unpaired Medical Images and Text. EMNLP'22

Improving the pre-training

MedCLIP



Similarity between image and text multi-label encodings

$$s = \frac{\mathbf{l}_{img}^{\top} \cdot \mathbf{l}_{txt}}{\|\mathbf{l}_{img}\| \cdot \|\mathbf{l}_{txt}\|}$$

Image-to-text soft target

$$y_{i,j}^{v \to t} = \frac{\exp(s_{i,j}/\tau)}{\sum_{j=1}^{N} \exp(s_{i,j}/\tau)}$$

For a given image-report pair

Use the rapport to extract entities

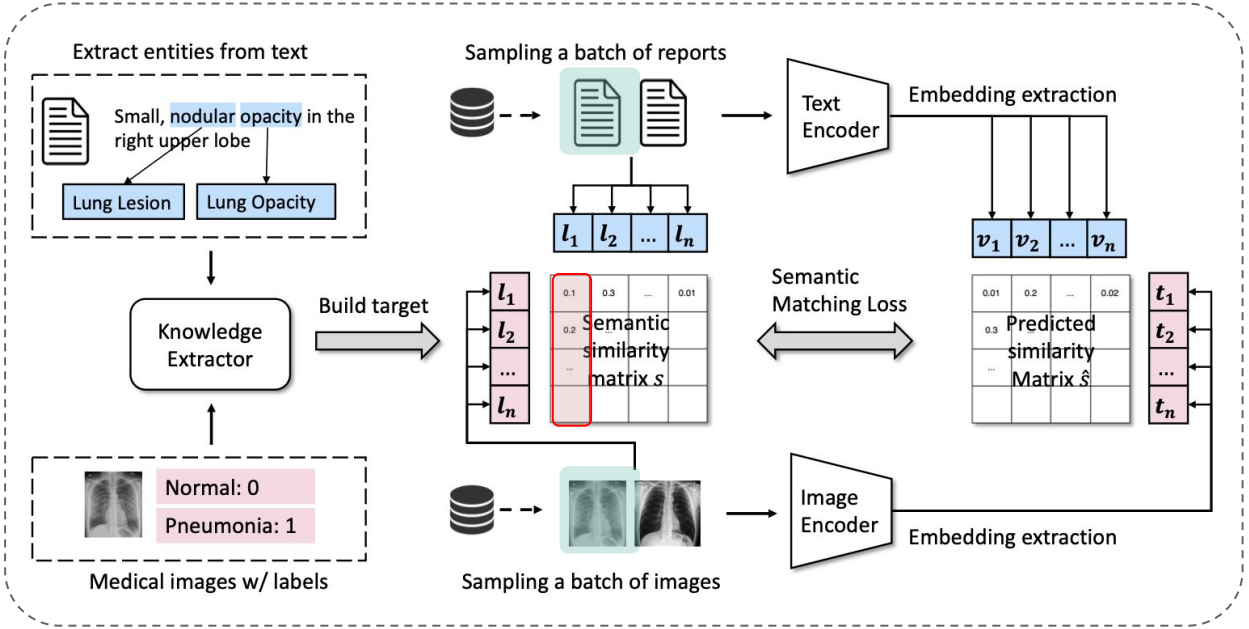
$$\mathbf{l}_i = [0, 1, 0, 0, 1, ..., 1]$$

Multi-label one-hot encoding

Wang et *al*. Contrastive Learning from Unpaired Medical Images and Text. EMNLP'22

Improving the pre-training

MedCLIP



Similarity between image and text multi-label encodings

$$s = \frac{\mathbf{l}_{img}^{\top} \cdot \mathbf{l}_{txt}}{\|\mathbf{l}_{img}\| \cdot \|\mathbf{l}_{txt}\|}$$

Image-to-text soft target

$$y_{i,j}^{v \to t} = \frac{\exp(s_{i,j}/\tau)}{\sum_{j=1}^{N} \exp(s_{i,j}/\tau)}$$

Text-to-image soft target

$$y_{j,i}^{t \to v} = \frac{\exp(s_{j,i}/\tau)}{\sum_{i=1}^{N} \exp(s_{j,i}/\tau)}$$

For a given image-report pair

Use the rapport to extract entities

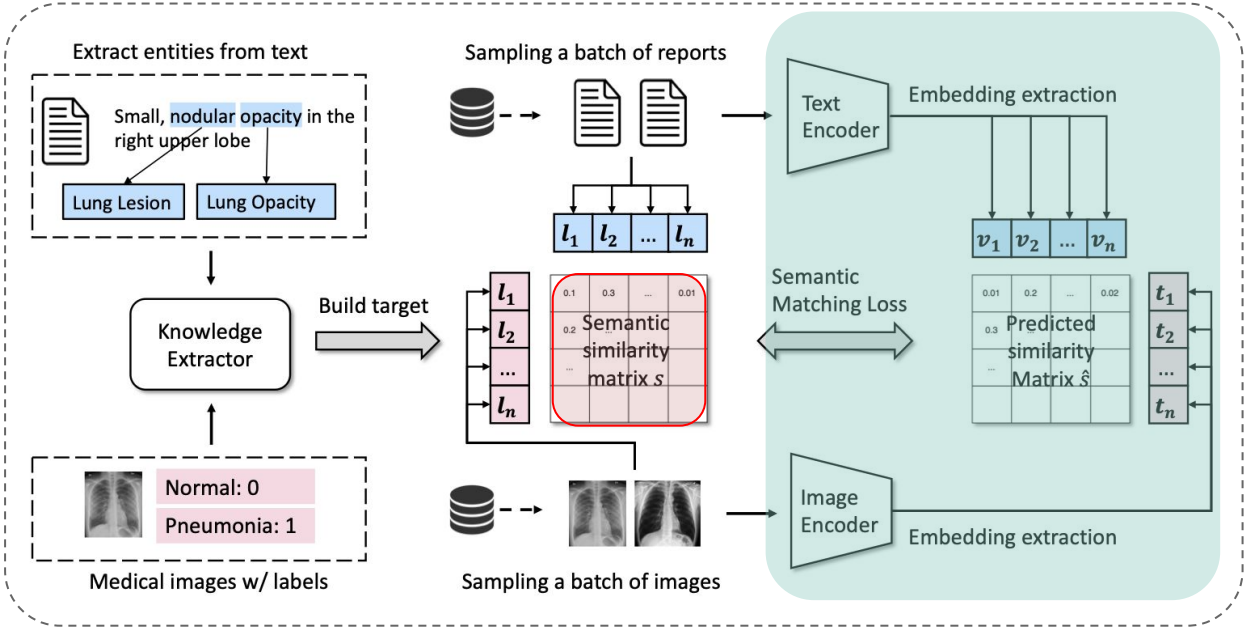
$$\mathbf{l}_i = [0, 1, 0, 0, 1, ..., 1]$$

Multi-label one-hot encoding

Wang et *al*. Contrastive Learning from Unpaired Medical Images and Text. EMNLP'22

Improving the pre-training

MedCLIP



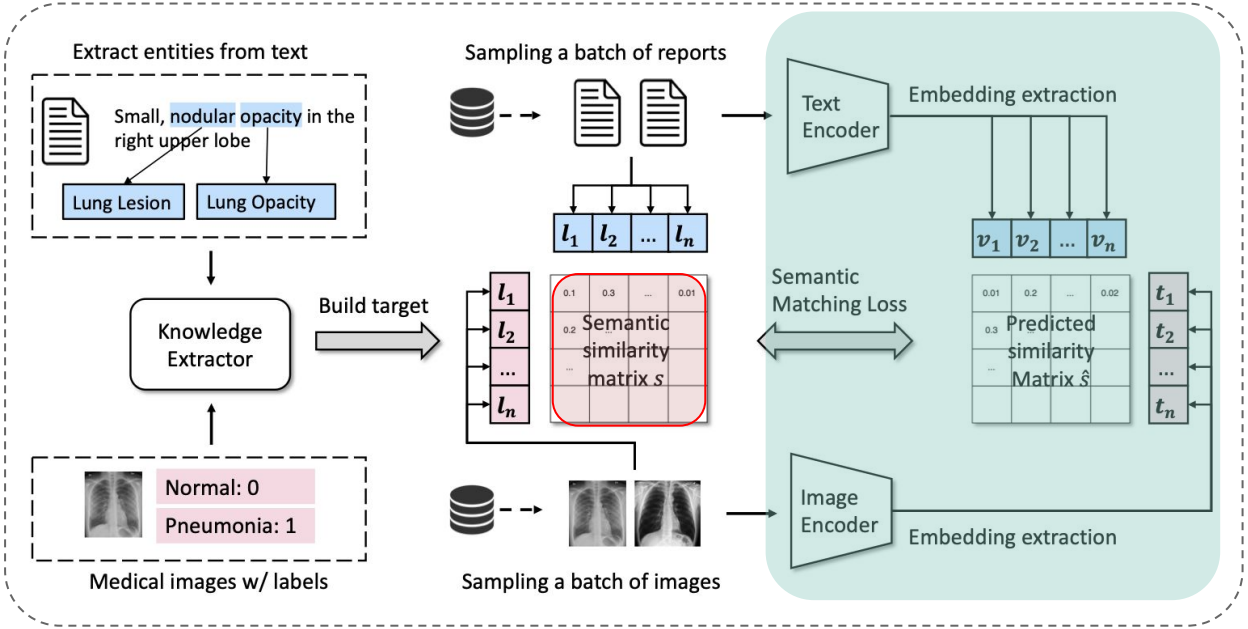
We already have the pseudo (soft) labels

$\mathbf{y}^{v \to t}$ $\mathbf{y}^{t \to v}$

Wang et *al*. Contrastive Learning from Unpaired Medical Images and Text. EMNLP'22

Improving the pre-training

MedCLIP



We already have the pseudo (soft) labels

$$\mathbf{y}^{v\to t} \qquad \mathbf{y}^{t\to v}$$

Compute predictions

$$\hat{y}^{v}_{i,j} = \frac{\exp\left(\mathbf{v}_i \cdot \mathbf{z}_j^\top/\tau\right)}{\sum_{n=1}^{N}\exp\left(\mathbf{v}_i \cdot \mathbf{z}_n^\top/\tau\right)}$$

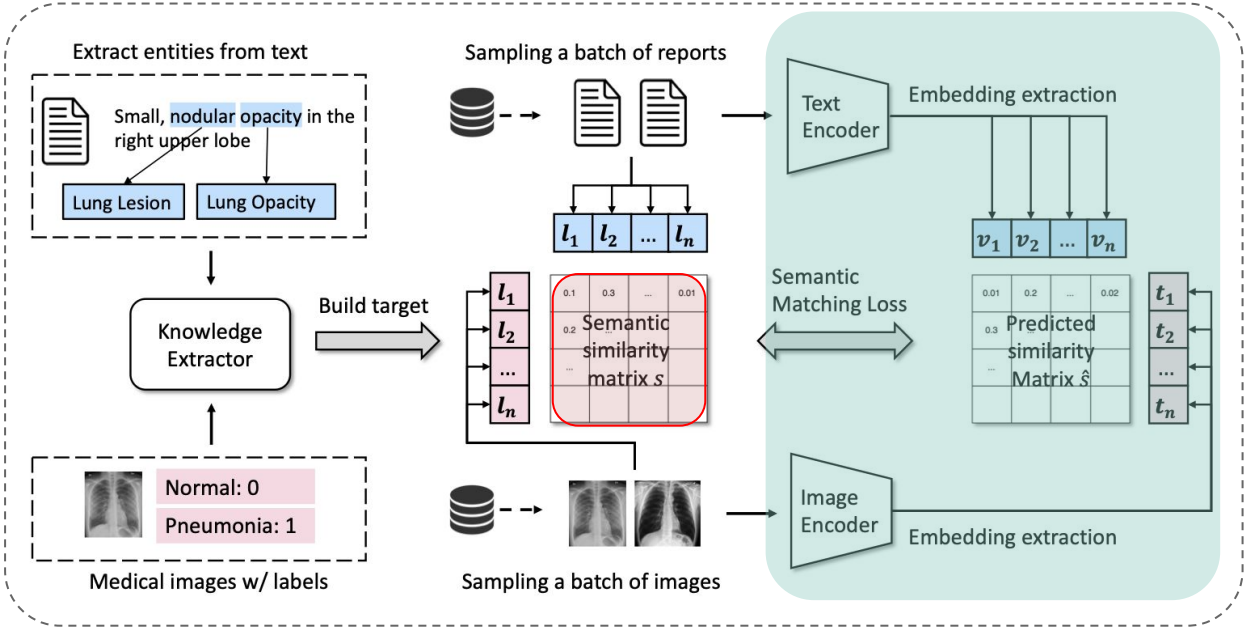
$$\hat{y}^{t}_{j,i} = \frac{\exp\left(\mathbf{v}_j \cdot \mathbf{z}_i^\top/\tau\right)}{\sum_{i=1}^{N}\exp\left(\mathbf{v}_j \cdot \mathbf{z}_i^\top/\tau\right)}$$

Note: this is the same as CLIP

Wang et *al*. Contrastive Learning from Unpaired Medical Images and Text. EMNLP'22

Improving the pre-training

MedCLIP



We already have the pseudo (soft) labels

$$\mathbf{y}^{v\to t} \qquad \mathbf{y}^{t\to v}$$

Compute predictions

$$\hat{y}^v_{i,j} = \frac{\exp\left(\mathbf{v}_i \cdot \mathbf{z}_j^\top/\tau\right)}{\sum_{n=1}^N \exp\left(\mathbf{v}_i \cdot \mathbf{z}_n^\top/\tau\right)}$$

$$\hat{y}^t_{j,i} = \frac{\exp\left(\mathbf{v}_j \cdot \mathbf{z}_i^\top/\tau\right)}{\sum_{i=1}^N \exp\left(\mathbf{v}_j \cdot \mathbf{z}_i^\top/\tau\right)}$$

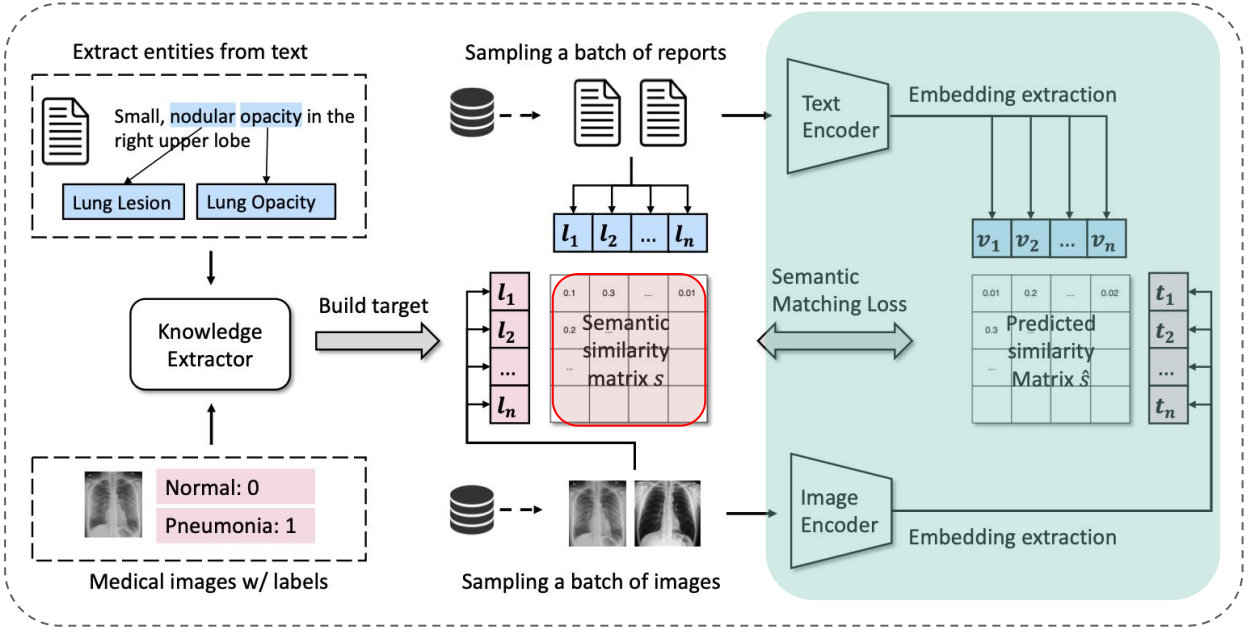
Note: this is the same as CLIP

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{N}(\mathcal{H}(\mathbf{y}_i^{v\to t}, \hat{\mathbf{y}}_i^v) + \mathcal{H}(\mathbf{y}_i^{t\to v}, \hat{\mathbf{y}}_i^t))$$

Wang et *al*. Contrastive Learning from Unpaired Medical Images and Text. EMNLP'22

Improving the pre-training

MedCLIP



We already have the pseudo (soft) labels

$$\mathbf{y}^{v\to t} \qquad \mathbf{y}^{t\to v}$$

Compute predictions

$$\hat{y}^v_{i,j} = \frac{\exp\left(\mathbf{v}_i \cdot \mathbf{z}_j^\top/\tau\right)}{\sum_{n=1}^N \exp\left(\mathbf{v}_i \cdot \mathbf{z}_n^\top/\tau\right)}$$

$$\hat{y}^t_{j,i} = \frac{\exp\left(\mathbf{v}_j \cdot \mathbf{z}_i^\top/\tau\right)}{\sum_{i=1}^N \exp\left(\mathbf{v}_j \cdot \mathbf{z}_i^\top/\tau\right)}$$

Note: this is the same as CLIP

Inference

For a given class c

$$\hat{y}_c = \frac{\exp(\mathbf{v} \cdot \mathbf{z}_c^\top/\tau)}{\sum_{k=1}^K \exp(\mathbf{v} \cdot \mathbf{z}_k^\top/\tau)}$$

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^N (\mathcal{H}(\mathbf{y}_i^{v\to t}, \hat{\mathbf{y}}_i^v) + \mathcal{H}(\mathbf{y}_i^{t\to v}, \hat{\mathbf{y}}_i^t))$$

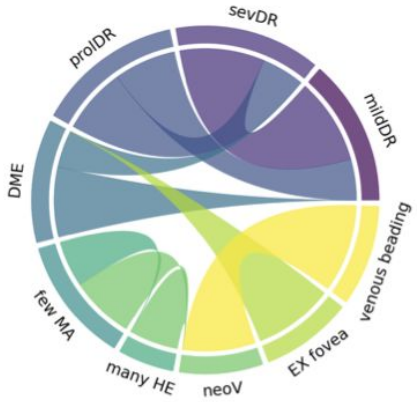
Wang et *al*. Contrastive Learning from Unpaired Medical Images and Text. EMNLP'22

Improving the pre-training

FLAIR

Incorporating expert-domain knowledge

CLIP



Limitations on medical domains (CLIP mostly focuses on general medical relations)

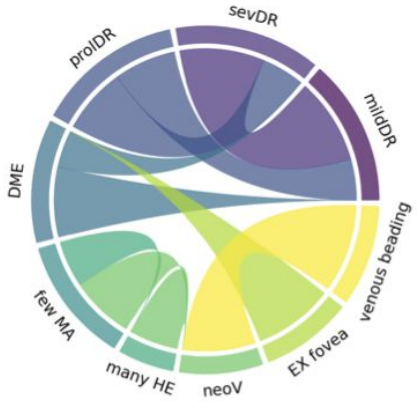
Silva-Rodriguez et *al*. A Foundation LAnguage-Image model of the Retina (FLAIR): Encoding expert knowledge in text supervision. MedIA'24 (Under review)

Improving the pre-training

FLAIR

Incorporating expert-domain knowledge

CLIP



Limitations on medical domains (CLIP mostly focuses on general medical relations)

How to capture hierarchical dependencies between concepts?

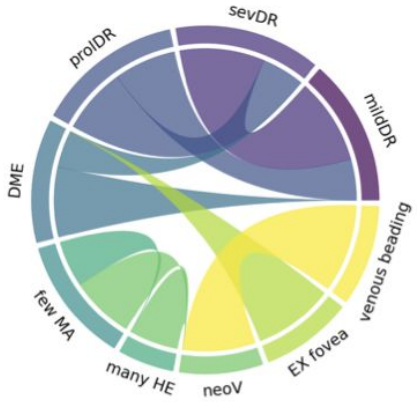
Silva-Rodriguez et *al*. A Foundation LAnguage-Image model of the Retina (FLAIR): Encoding expert knowledge in text supervision. MedIA'24 (Under review)

Improving the pre-training

FLAIR

Incorporating expert-domain knowledge

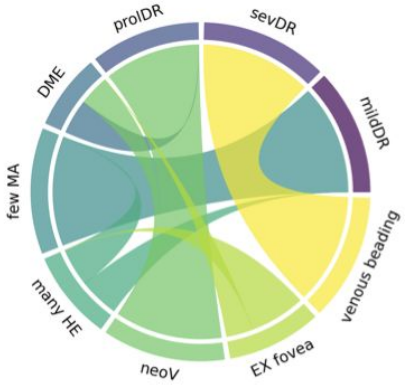
CLIP



Limitations on medical domains (CLIP mostly focuses on general medical relations)

How to capture hierarchical dependencies between concepts?

Adding domain specific-knowledge



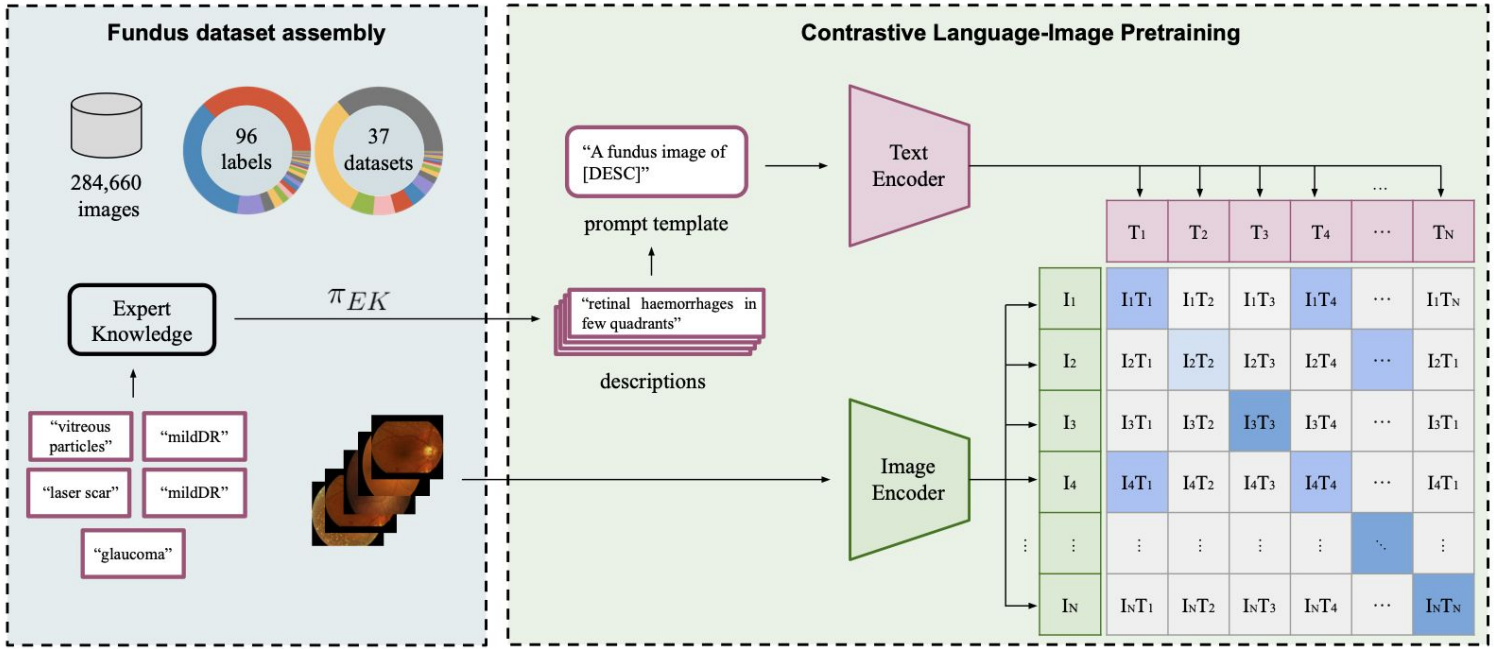
E.g., ***"neovascularization"*** is the differential sign for ***proIDR***.

Silva-Rodriguez et *al*. A Foundation LAnguage-Image model of the Retina (FLAIR): Encoding expert knowledge in text supervision. MedIA'24 (Under review)

Improving the pre-training

FLAIR

Incorporating expert-domain knowledge

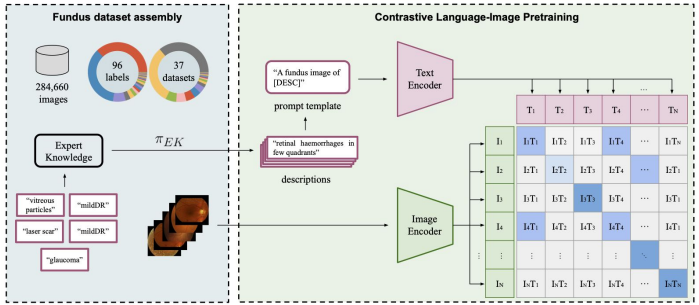


Silva-Rodriguez et *al*. A Foundation LAnguage-Image model of the Retina (FLAIR): Encoding expert knowledge in text supervision. MedIA'24 (Under review)

Improving the pre-training

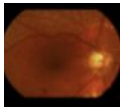
FLAIR

Incorporating expert-domain knowledge

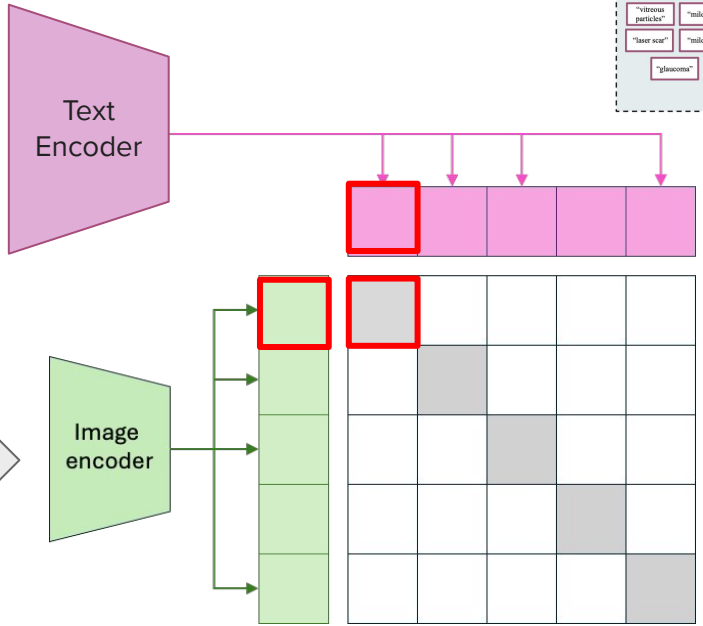


"A fundus image of possible signs of hemorrhage with blot, dot, or flame-shaped"

Expert Knowledge



Severe hypertensive retinopathy



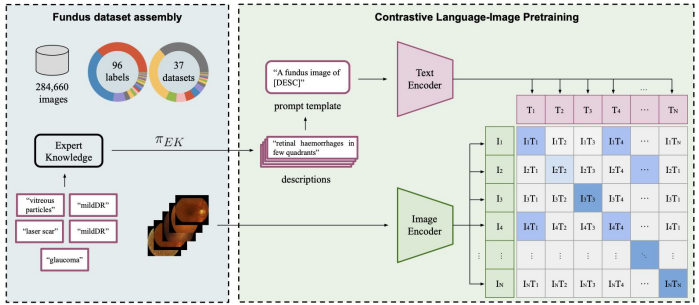
This would be standard CLIP

Silva-Rodriguez et *al*. A Foundation LAnguage-Image model of the Retina (FLAIR): Encoding expert knowledge in text supervision. MedIA'24 (Under review)

Improving the pre-training

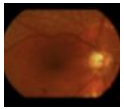
FLAIR

Incorporating expert-domain knowledge

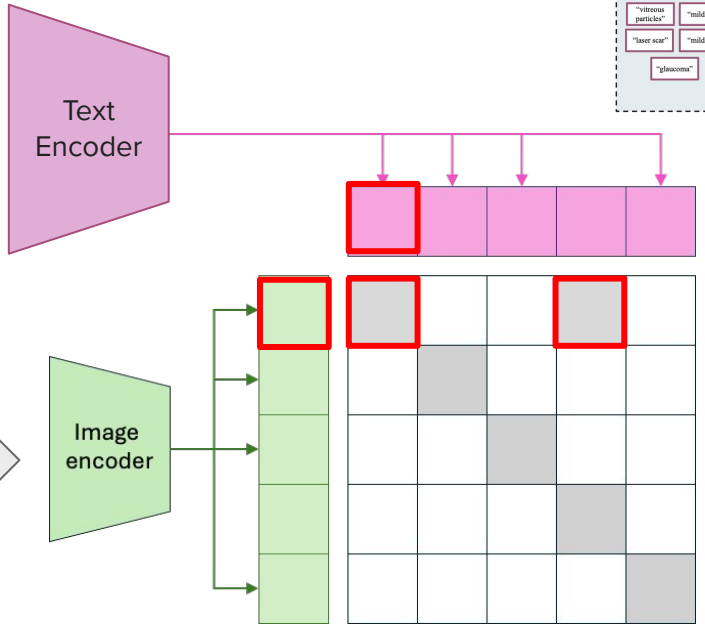


"A fundus image of possible signs of hemorrhage with blot, dot, or flame-shaped"

Expert Knowledge



Severe hypertensive retinopathy



What we have now:

$$\mathcal{L}^v = -\sum_{i\in\mathcal{B}} \frac{1}{|P_{\mathcal{T}_\mathcal{B}}(i)|} \sum_{i'\in P_{\mathcal{T}_\mathcal{B}}(i)} \log \frac{\exp\left(\mathbf{v}_i \cdot \mathbf{z}_{i'}^\top/\tau\right)}{\sum_{j\in\mathcal{T}_\mathcal{B}} \exp\left(\mathbf{v}_i \cdot \mathbf{z}_j^\top/\tau\right)}$$

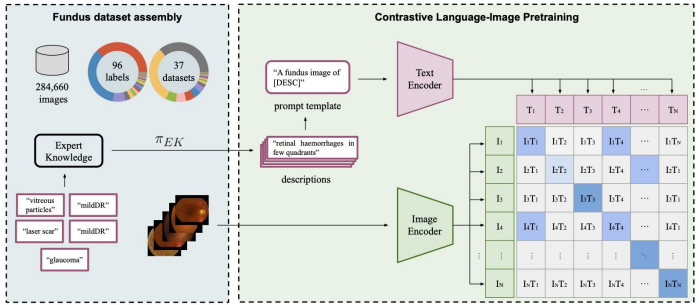
For each image in the batch

Silva-Rodriguez et *al*. A Foundation LAnguage-Image model of the Retina (FLAIR): Encoding expert knowledge in text supervision. MedIA'24 (Under review)

Improving the pre-training

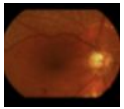
FLAIR

Incorporating expert-domain knowledge

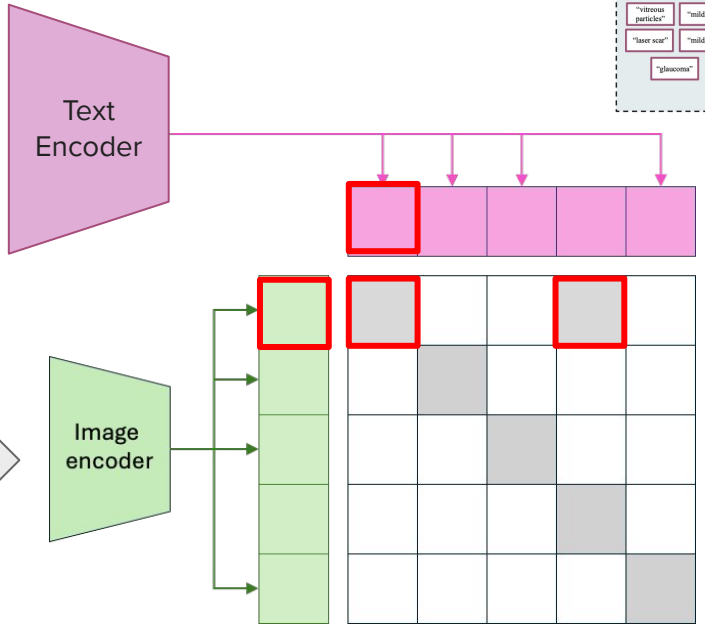


"A fundus image of possible signs of hemorrhage with blot, dot, or flame-shaped"

Expert Knowledge



Severe hypertensive retinopathy



What we have now:

$$\mathcal{L}^v = -\sum_{i\in\mathcal{B}} \frac{1}{|P_{\mathcal{T}_\mathcal{B}}(i)|} \sum_{i'\in P_{\mathcal{T}_\mathcal{B}}(i)} \log \frac{\exp\left(\mathbf{v}_i \cdot \mathbf{z}_{i'}^\top/\tau\right)}{\sum_{j\in\mathcal{T}_\mathcal{B}} \exp\left(\mathbf{v}_i \cdot \mathbf{z}_j^\top/\tau\right)}$$

For each image in the batch

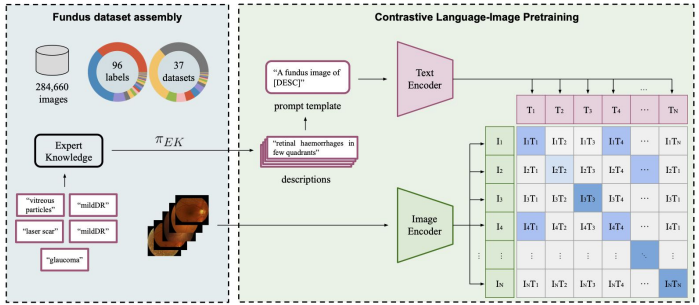
For each text prompt whose class is same as class of image i

Silva-Rodriguez et *al*. A Foundation LAnguage-Image model of the Retina (FLAIR): Encoding expert knowledge in text supervision. MedIA'24 (Under review)

Improving the pre-training

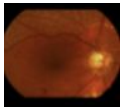
FLAIR

Incorporating expert-domain knowledge

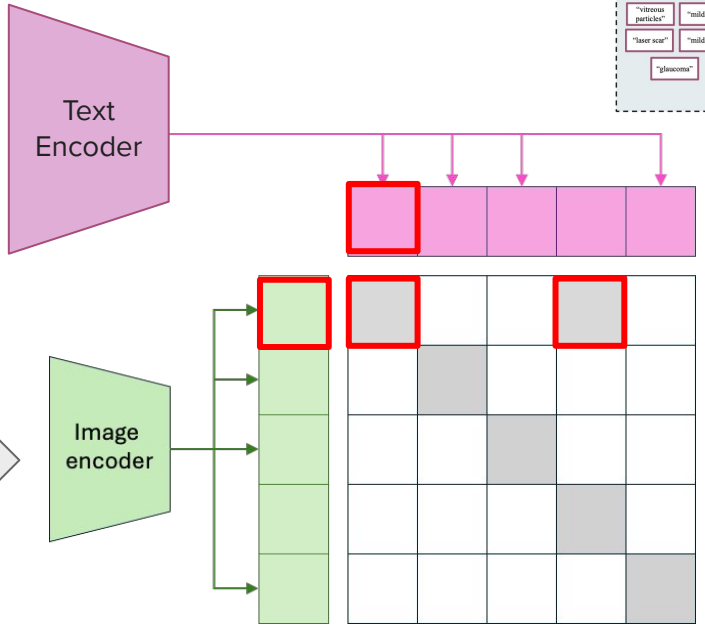


"A fundus image of possible signs of hemorrhage with blot, dot, or flame-shaped"

Expert Knowledge



Severe hypertensive retinopathy



What we have now:

$$\mathcal{L}^v = -\sum_{i \in \mathcal{B}} \frac{1}{|P_{\mathcal{T}_\mathcal{B}}(i)|} \sum_{i' \in P_{\mathcal{T}_\mathcal{B}}(i)} \log \frac{\exp\left(\mathbf{v}_i \cdot \mathbf{z}_{i'}^\top / \tau\right)}{\sum_{j \in \mathcal{T}_\mathcal{B}} \exp\left(\mathbf{v}_i \cdot \mathbf{z}_j^\top / \tau\right)}$$

For each image in the batch

Same class: Maximize similarity

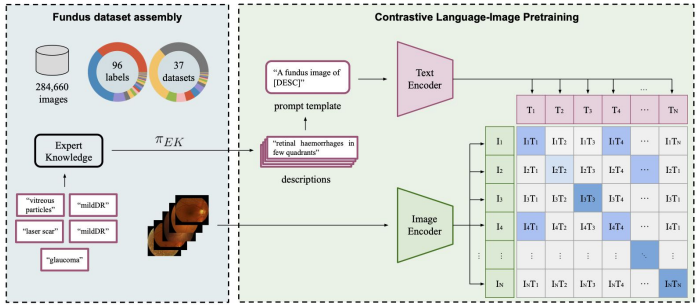
For each text prompt whose class is same as class of image i

Silva-Rodriguez et *al*. A Foundation LAnguage-Image model of the Retina (FLAIR): Encoding expert knowledge in text supervision. MedIA'24 (Under review)

Improving the pre-training

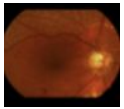
FLAIR

Incorporating expert-domain knowledge

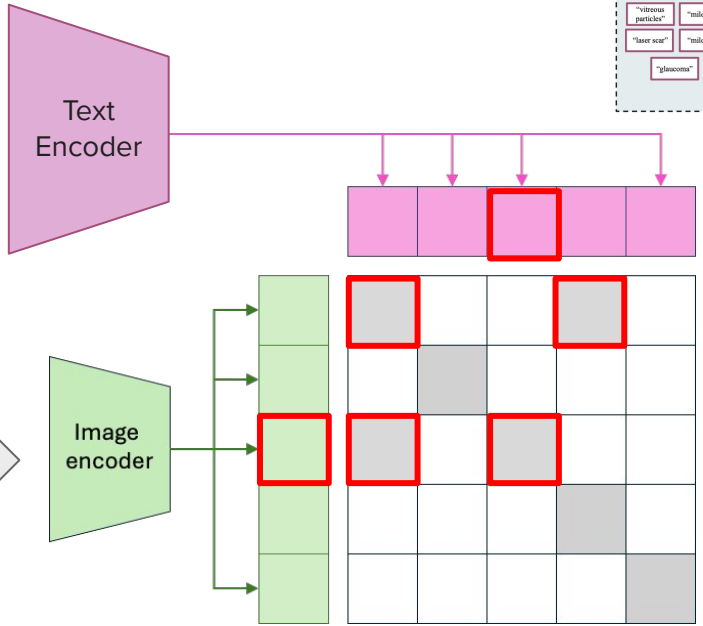


"A fundus image of arterio-venous crossing changes, macular star.."

Expert Knowledge



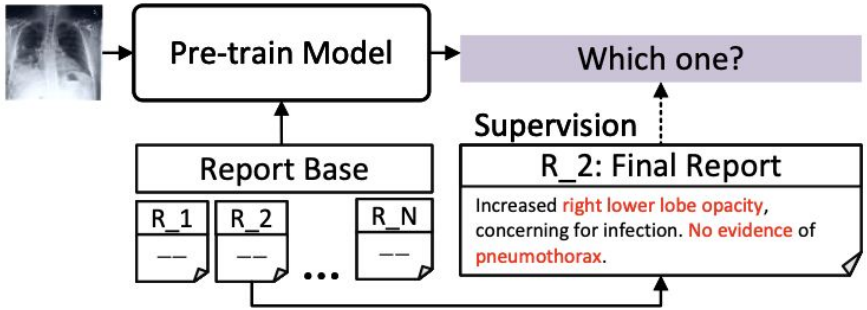
Severe hypertensive retinopathy



Silva-Rodriguez et *al*. A Foundation LAnguage-Image model of the Retina (FLAIR): Encoding expert knowledge in text supervision. MedIA'24 (Under review)

Improving the pre-training

MedKLIP

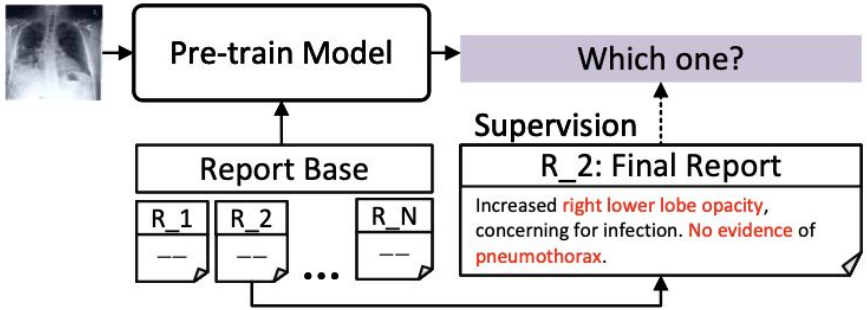


Standard VLP flowchart

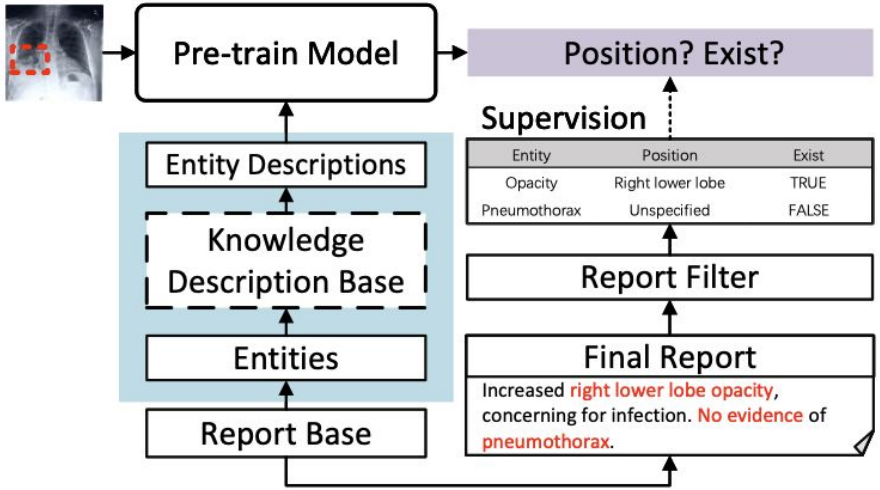
Wu et *al*. MedKLIP: Medical knowledge enhanced Language-Image Pre-Training.

Improving the pre-training

MedKLIP



Standard VLP flowchart



MedKLIP
(Inclusion of entities)

Motivations/challenges:

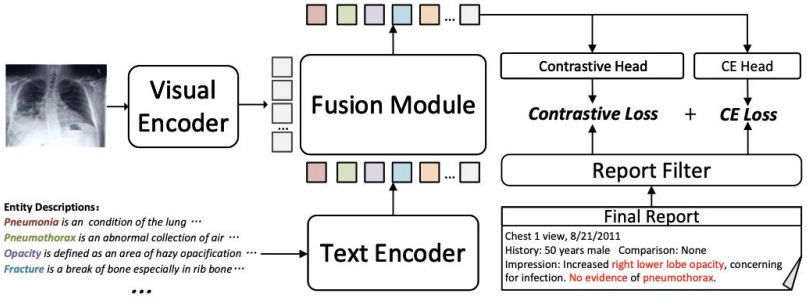
1 - Medical reports contain language that is esoteric to the common person
2 - Multiple ways to phrase a disease.

Wu et *al*. MedKLIP: Medical knowledge enhanced Language-Image Pre-Training.

Improving the pre-training

MedKLIP

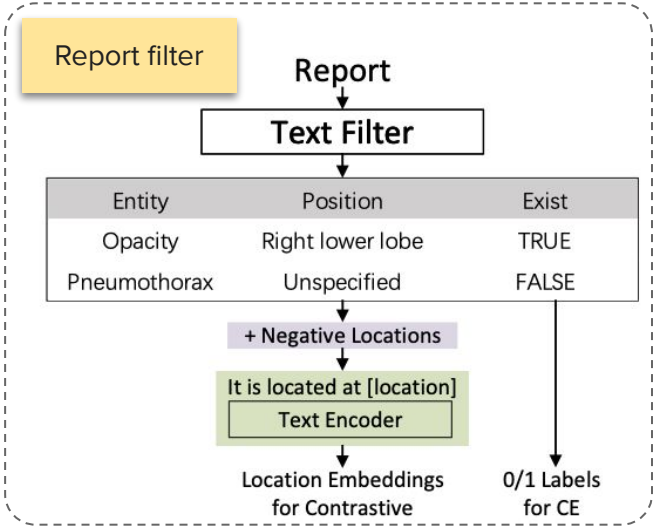
Entity: *'the disease we are looking for'*

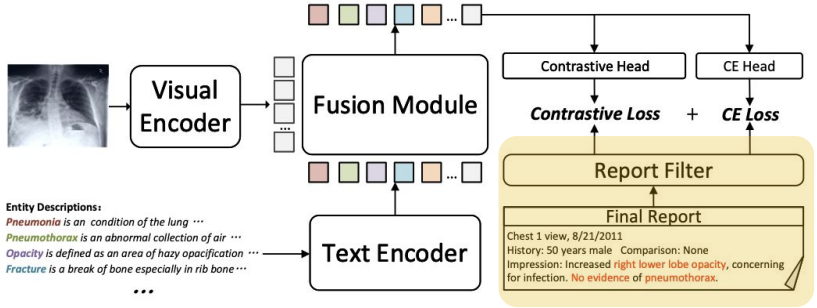


Wu et *al*. MedKLIP: Medical knowledge enhanced Language-Image Pre-Training.

Improving the pre-training

MedKLIP



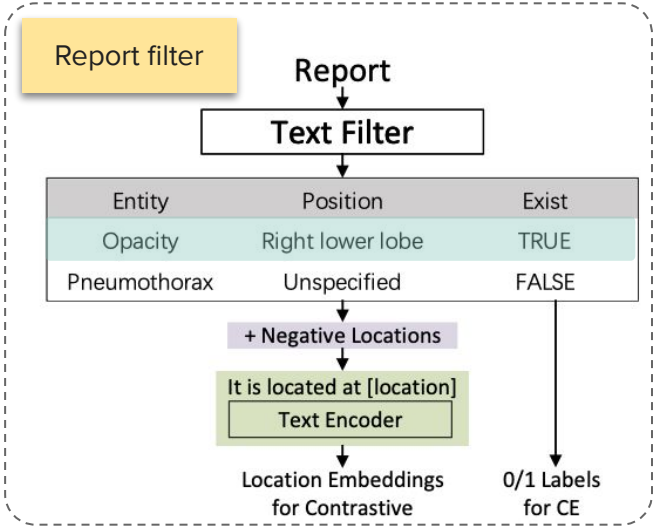


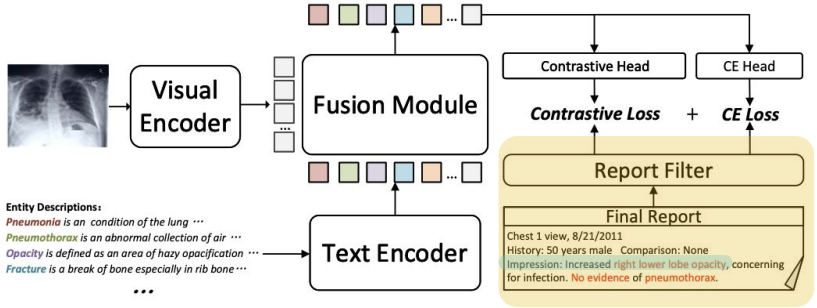
"Simplify" data in medical reports: *triplets* $\{e_1, e_2, e_3\}$

Wu et *al*. MedKLIP: Medical knowledge enhanced Language-Image Pre-Training.

Improving the pre-training

MedKLIP





"Simplify" data in medical reports: *triplets* $\{e_1, e_2, e_3\}$

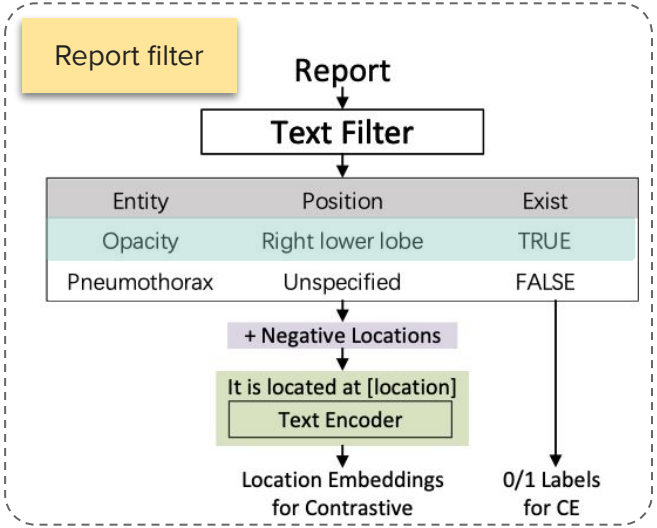
Pre-trained model

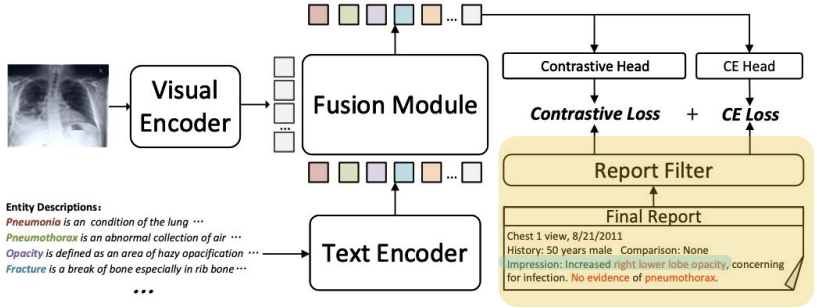
From text input (e.g., medical report), we can have *j* sentences, with *n* entities. $\Phi(S_j) = \{entity_n, position_n, exist_n\}$

Wu et *al*. MedKLIP: Medical knowledge enhanced Language-Image Pre-Training.

Improving the pre-training

MedKLIP





"Simplify" data in medical reports: *triplets* $\{e_1, e_2, e_3\}$

Pre-trained model

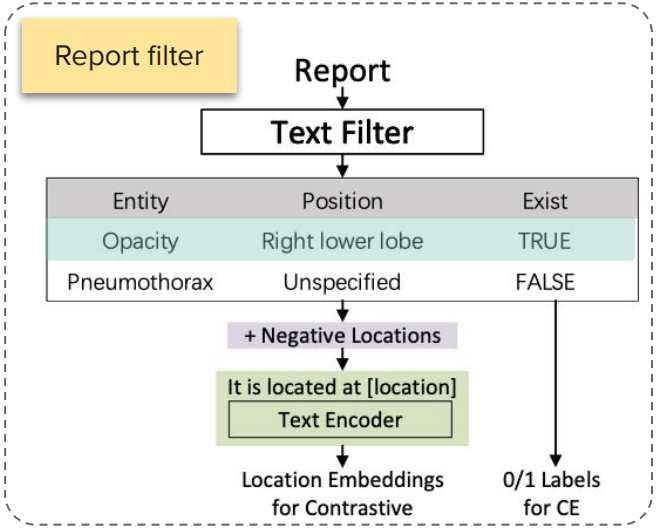
'the disease'

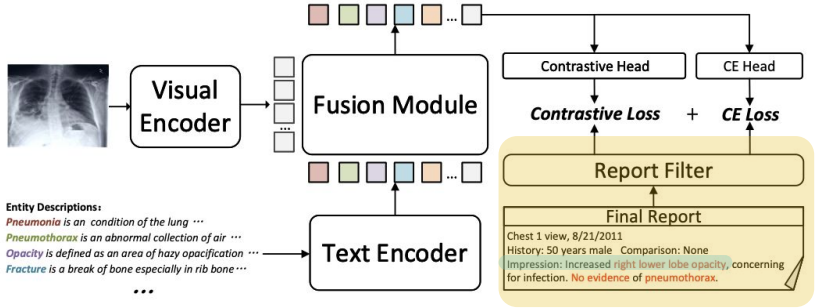
From text input (e.g., medical report), we can have *j* sentences, with *n* entities. $\Phi(S_j) = \{entity_n, position_n, exist_n\}$

Wu et *al*. MedKLIP: Medical knowledge enhanced Language-Image Pre-Training.

Improving the pre-training

MedKLIP





"Simplify" data in medical reports: *triplets* $\{e_1, e_2, e_3\}$

Pre-trained model

'the disease'

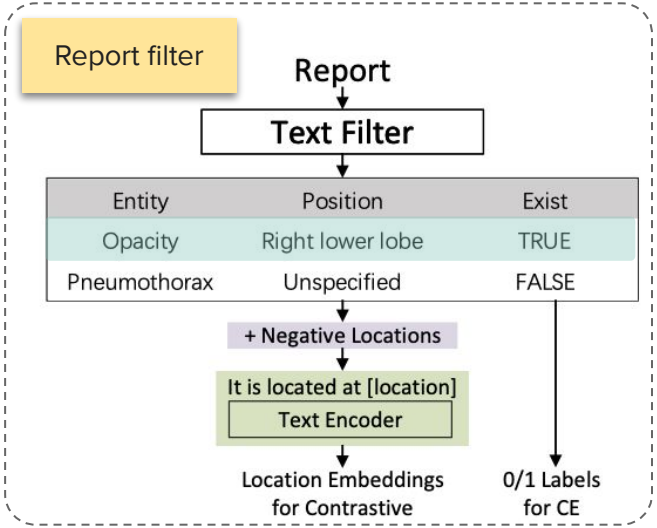
From text input (e.g., medical report), we can have *j* sentences, with *n* entities. $\Phi(S_j) = \{entity_n, position_n, exist_n\}$

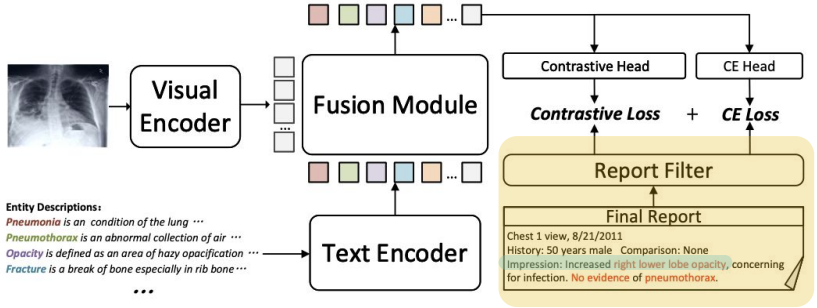
'position (from text)'

Wu et *al*. MedKLIP: Medical knowledge enhanced Language-Image Pre-Training.

Improving the pre-training

MedKLIP





"Simplify" data in medical reports: *triplets* $\{e_1, e_2, e_3\}$

From text input (e.g., medical report), we can have *j* sentences, with *n* entities.

Pre-trained model

'the disease'

'whether the disease exists'

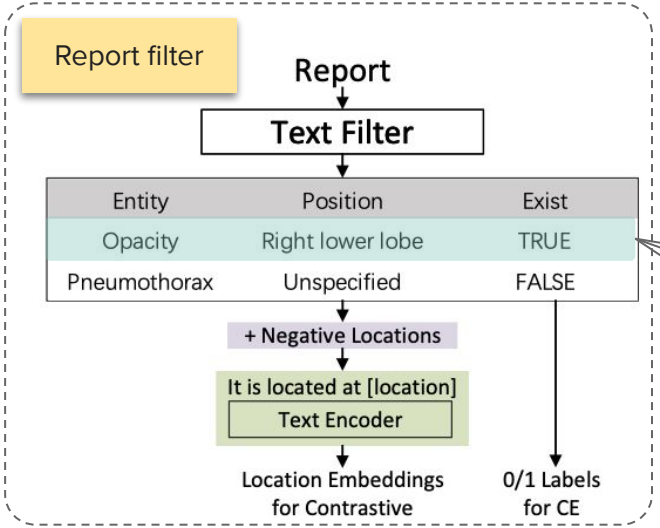
$$\Phi(S_j) = \{entity_n, position_n, exist_n\}$$

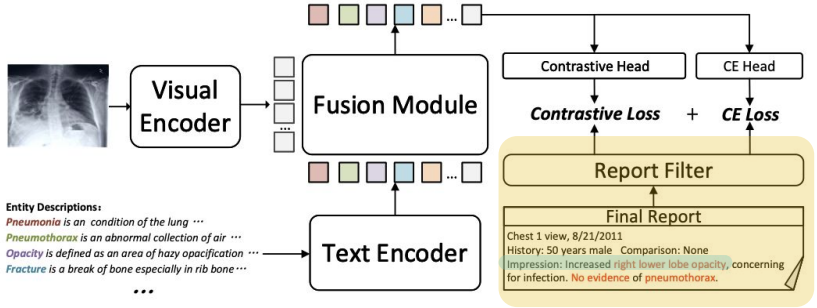
'position (from text)'

Wu et *al*. MedKLIP: Medical knowledge enhanced Language-Image Pre-Training.

Improving the pre-training

MedKLIP





'Located in the right lower lobe'

'Exists from the image'

'Opacity'

"Simplify" data in medical reports: *triplets* $\{e_1, e_2, e_3\}$

From text input (e.g., medical report), we can have *j* sentences, with *n* entities. $\Phi(S_j) = \{entity_n, position_n, exist_n\}$

Wu et *al*. MedKLIP: Medical knowledge enhanced Language-Image Pre-Training.

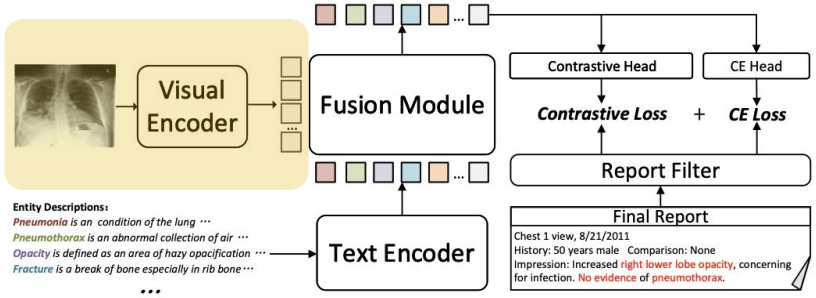
Improving the pre-training

MedKLIP

Visual encoder

Extract visual features

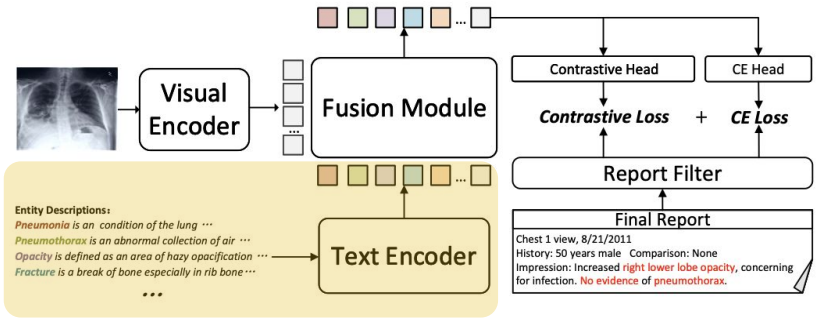
$$\mathbf{v}_i = f_\theta^v(\mathbf{x}_i)$$



Wu et *al*. MedKLIP: Medical knowledge enhanced Language-Image Pre-Training.

Improving the pre-training

MedKLIP



Visual encoder

Extract visual features

$$\mathbf{v}_i = f_\theta^v(\mathbf{x}_i)$$

Text encoder
(and entity expansion!)

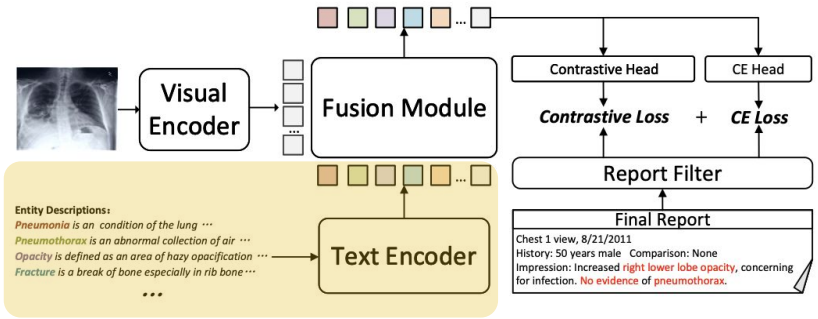
ClinicalBERT

Entity = 'Pneumonia' → Description: *'Pneumonia is a condition of the lung that….'*

Wu et *al*. MedKLIP: Medical knowledge enhanced Language-Image Pre-Training.

Improving the pre-training

MedKLIP



Visual encoder

Extract visual features

$$\mathbf{v}_i = f_\theta^v(\mathbf{x}_i)$$

Text encoder
(and entity expansion!)

ClinicalBERT

Entity = 'Pneumonia'

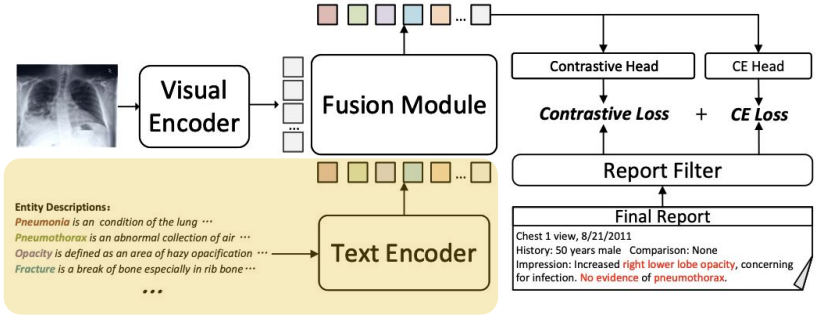
Description: *'Pneumonia is a condition of the lung that….'*

$$\mathbf{z}_i = f_\theta^t(\mathbf{t}_i)$$

Wu et *al*. MedKLIP: Medical knowledge enhanced Language-Image Pre-Training.

Improving the pre-training

MedKLIP



Visual encoder

Extract visual features

$$\mathbf{v}_i = f_\theta^v(\mathbf{x}_i)$$

Text encoder
(and entity expansion!)

ClinicalBERT

Entity = 'Pneumonia'

Description: *'Pneumonia is a condition of the lung that….'*

$$\mathbf{z}_i = f_\theta^t(\mathbf{t}_i)$$

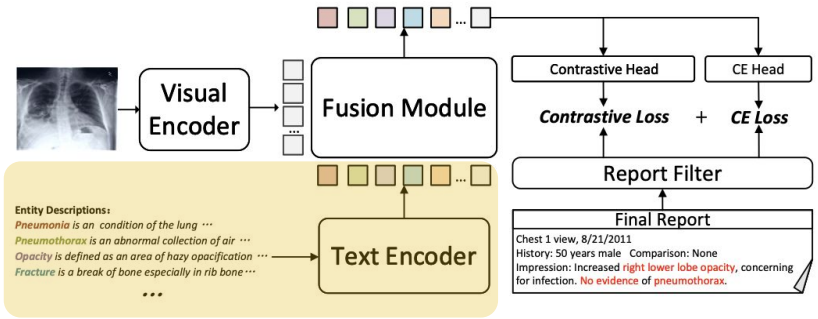
Extract the position encoded features

$$\mathbf{p}_i = f_\theta^t(\mathbf{pos}_i)$$

Wu et *al*. MedKLIP: Medical knowledge enhanced Language-Image Pre-Training.

Improving the pre-training

MedKLIP



Visual encoder

Extract visual features

$$\mathbf{v}_i = f_\theta^v(\mathbf{x}_i)$$

Text encoder
(and entity expansion!)

ClinicalBERT

Entity = 'Pneumonia'

Description: *'Pneumonia is a condition of the lung that….'*

$$\mathbf{z}_i = f_\theta^t(\mathbf{t}_i)$$

Extract the position encoded features

$$\mathbf{p}_i = f_\theta^t(\mathbf{pos}_i)$$

$$\Phi(S_j) = \{entity_n, position_n, exist_n\}$$

$$\{\mathbf{z_n}, \mathbf{pos}_n, l_n\}$$

Exist or not ('0' or '1')

Wu et *al*. MedKLIP: Medical knowledge enhanced Language-Image Pre-Training.

Improving the pre-training

MedKLIP

Visual encoder

Extract visual features

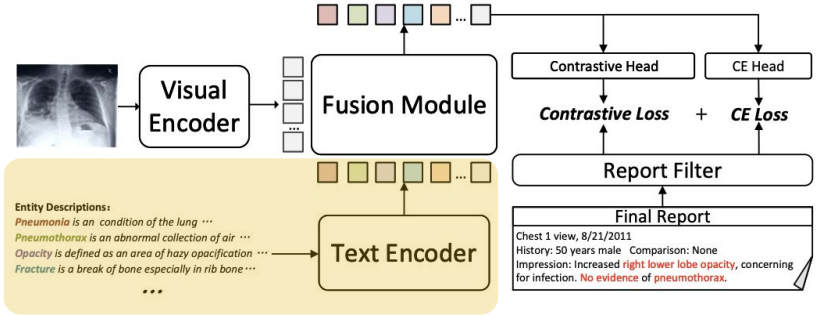
$$\mathbf{v}_i = f_\theta^v(\mathbf{x}_i)$$

Text encoder
(and entity expansion!)

$$S_1 = \{\mathbf{z_n}, \mathbf{pos}_n, l_n\}$$

$$\vdots$$

$$S_M = \{\mathbf{z_n}, \mathbf{pos}_n, l_n\}$$



Wu et *al*. MedKLIP: Medical knowledge enhanced Language-Image Pre-Training.

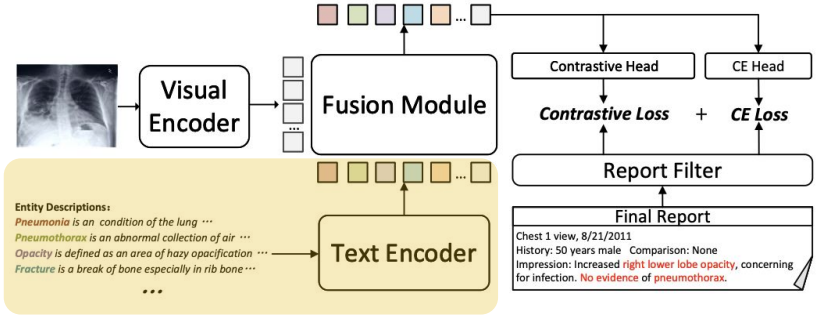
Improving the pre-training

MedKLIP

Visual encoder

Extract visual features

$$\mathbf{v}_i = f_\theta^v(\mathbf{x}_i)$$



Text encoder
(and entity expansion!)

$$S_1 = \{\mathbf{z_n}, \mathbf{pos}_n, l_n\}$$

$$\vdots$$

$$S_M = \{\mathbf{z_n}, \mathbf{pos}_n, l_n\}$$

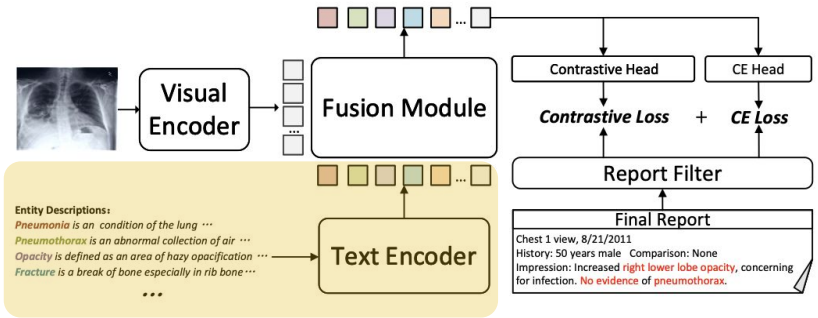
Most common entities (Query)

$$\mathcal{Q} = \{\mathbf{z}_4, \mathbf{z}_{36}, \ldots, \mathbf{z}_{89}\}$$

Wu et *al*. MedKLIP: Medical knowledge enhanced Language-Image Pre-Training.

Improving the pre-training

MedKLIP



Visual encoder

Extract visual features

$$\mathbf{v}_i = f_\theta^v(\mathbf{x}_i)$$

Text encoder
(and entity expansion!)

$$S_1 = \{\mathbf{z_n}, \mathbf{pos}_n, l_n\}$$

$$\vdots$$

$$S_M = \{\mathbf{z_n}, \mathbf{pos}_n, l_n\}$$

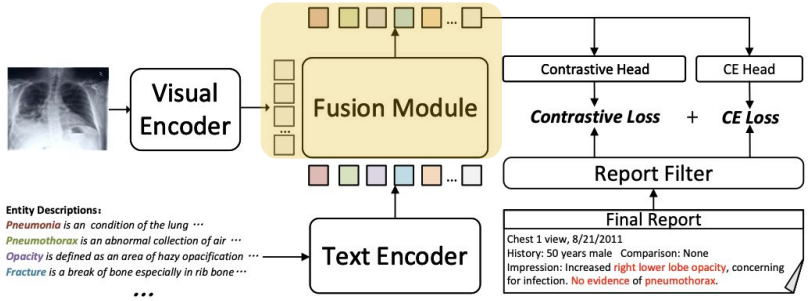
Most common entities (Query) $\mathcal{Q} = \{\mathbf{z}_4, \mathbf{z}_{36}, ..., \mathbf{z}_{89}\}$

Most common positions (P) $\mathcal{P} = \{\mathbf{pos}_9, \mathbf{pos}_{45}, ..., \mathbf{pos}_{57}\}$

Wu et *al*. MedKLIP: Medical knowledge enhanced Language-Image Pre-Training.

Improving the pre-training

MedKLIP



Fusion module

Query $\mathcal{Q} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{|Q|}\}$

Key and Value (for a given image) $\mathbf{v}_i$

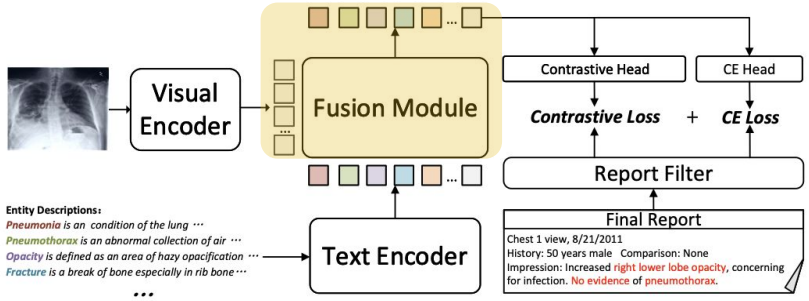
Fusion module

$$\{\hat{\mathbf{s}}_i, \hat{\mathbf{p}}_i, \hat{\mathbf{m}}_i\} = f_\psi(\mathbf{v}_i, \mathcal{Q})$$

Wu et *al*. MedKLIP: Medical knowledge enhanced Language-Image Pre-Training.

Improving the pre-training

MedKLIP



Fusion module

Query $\mathcal{Q} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{|Q|}\}$

Key and Value (for a given image) $\mathbf{v}_i$

Fusion module

$$\{\hat{\mathbf{s}}_i, \hat{\mathbf{p}}_i, \hat{\mathbf{m}}_i\} = f_\psi(\mathbf{v}_i, \mathcal{Q})$$

Prediction indicating the presence of entities

$\hat{\mathbf{s}} \in \mathbb{R}^{|\mathcal{Q}|}$

Prediction of spatial positioning of entities

$\hat{\mathbf{p}} \in \mathbb{R}^{|\mathcal{Q}| \times d}$

Cross-attention maps (from the transformer)

Wu et *al*. MedKLIP: Medical knowledge enhanced Language-Image Pre-Training.

Improving the pre-training

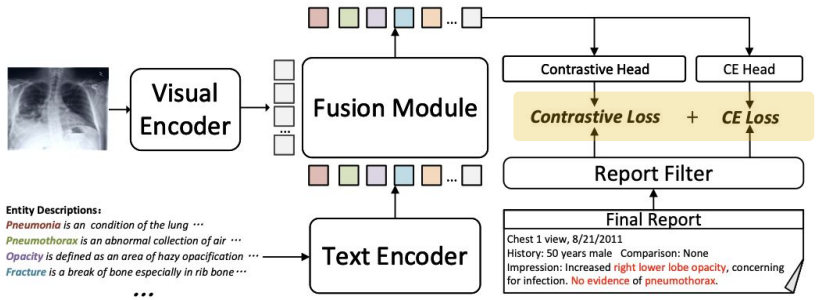
MedKLIP

Loss function

$$\{\hat{\mathbf{s}}_i, \hat{\mathbf{p}}_i, \hat{\mathbf{m}}_i\} = f_\psi(\mathbf{v}_i, \mathcal{Q})$$

$$\mathcal{L}_{\text{loc}} = -\frac{1}{|Q|}\sum_{k=1}^{|Q|} \frac{e^{\langle \hat{p}_k, p_k \rangle}}{e^{\langle \hat{p}_k, p_k \rangle} + \sum_{u=1}^{M} e^{\langle \hat{p}_k, P_{\mathcal{I}(k,u)} \rangle}}$$

Contrastive learning on the positions across the same class *k*



Wu et *al*. MedKLIP: Medical knowledge enhanced Language-Image Pre-Training.

Improving the pre-training

MedKLIP

Loss function

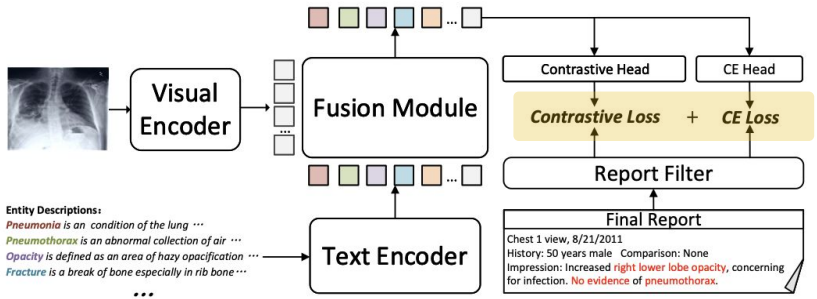
$$\{\hat{\mathbf{s}}_i, \hat{\mathbf{p}}_i, \hat{\mathbf{m}}_i\} = f_\psi(\mathbf{v}_i, \mathcal{Q})$$

$$\mathcal{L}_{\text{loc}} = -\frac{1}{|Q|}\sum_{k=1}^{|Q|} \frac{e^{\langle \hat{p}_k, p_k \rangle}}{e^{\langle \hat{p}_k, p_k \rangle} + \sum_{u=1}^{M} e^{\langle \hat{p}_k, P_{\mathcal{I}(k,u)} \rangle}}$$

Contrastive learning on the positions across the same class *k*

$$\mathcal{L}_{\text{cls}} = \mathcal{H}(\mathbf{l}_i, \hat{\mathbf{s}}_i)$$

Classification loss



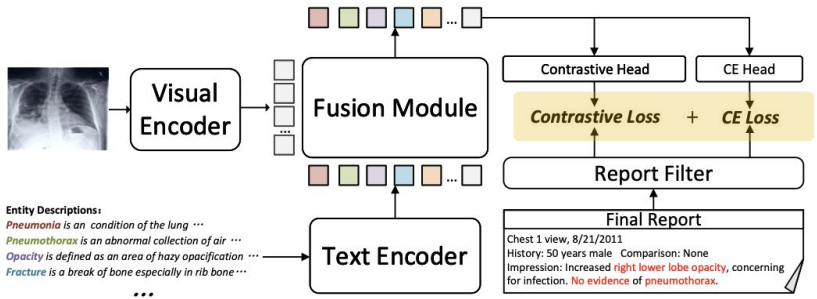
Wu et *al*. MedKLIP: Medical knowledge enhanced Language-Image Pre-Training.

Improving the pre-training

MedKLIP

Loss function

$$\{\hat{\mathbf{s}}_i, \hat{\mathbf{p}}_i, \hat{\mathbf{m}}_i\} = f_\psi(\mathbf{v}_i, \mathcal{Q})$$



$$\mathcal{L}_{\text{loc}} = -\frac{1}{|Q|}\sum_{k=1}^{|Q|} \frac{e^{\langle \hat{p}_k, p_k \rangle}}{e^{\langle \hat{p}_k, p_k \rangle} + \sum_{u=1}^{M} e^{\langle \hat{p}_k, P_{\mathcal{I}(k,u)} \rangle}}$$

Contrastive learning on the positions across the same class *k*

$$\mathcal{L}_{\text{cls}} = \mathcal{H}(\mathbf{l}_i, \hat{\mathbf{s}}_i)$$

Classification loss

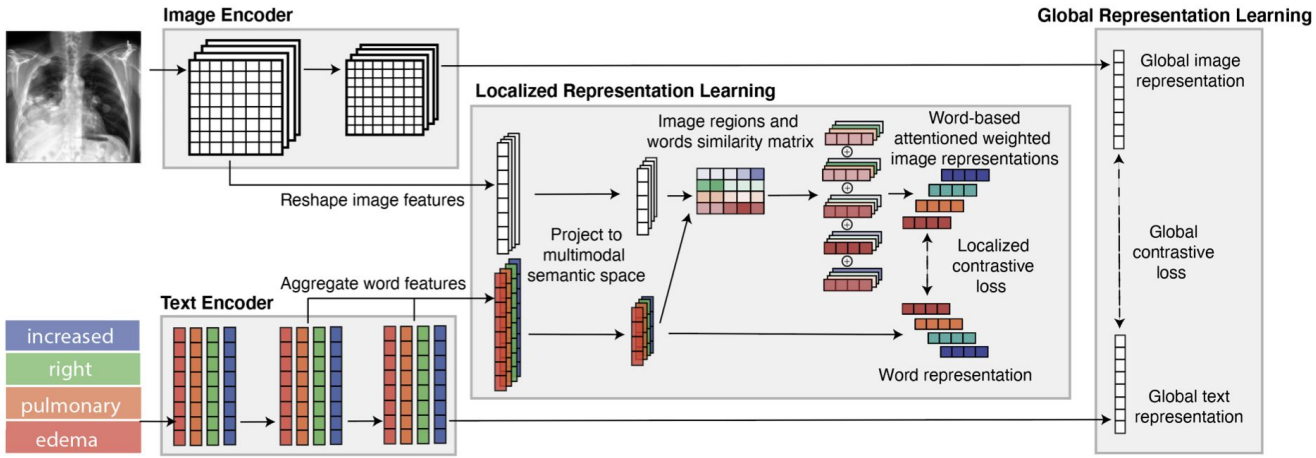
$$\mathcal{L}_{\text{total}} = \alpha_1 \mathcal{L}_{\text{loc}} + \alpha_2 \mathcal{L}_{\text{cls}}$$

Wu et *al*. MedKLIP: Medical knowledge enhanced Language-Image Pre-Training.

Improving the pre-training

GLoRIA

Integrate global and local representations

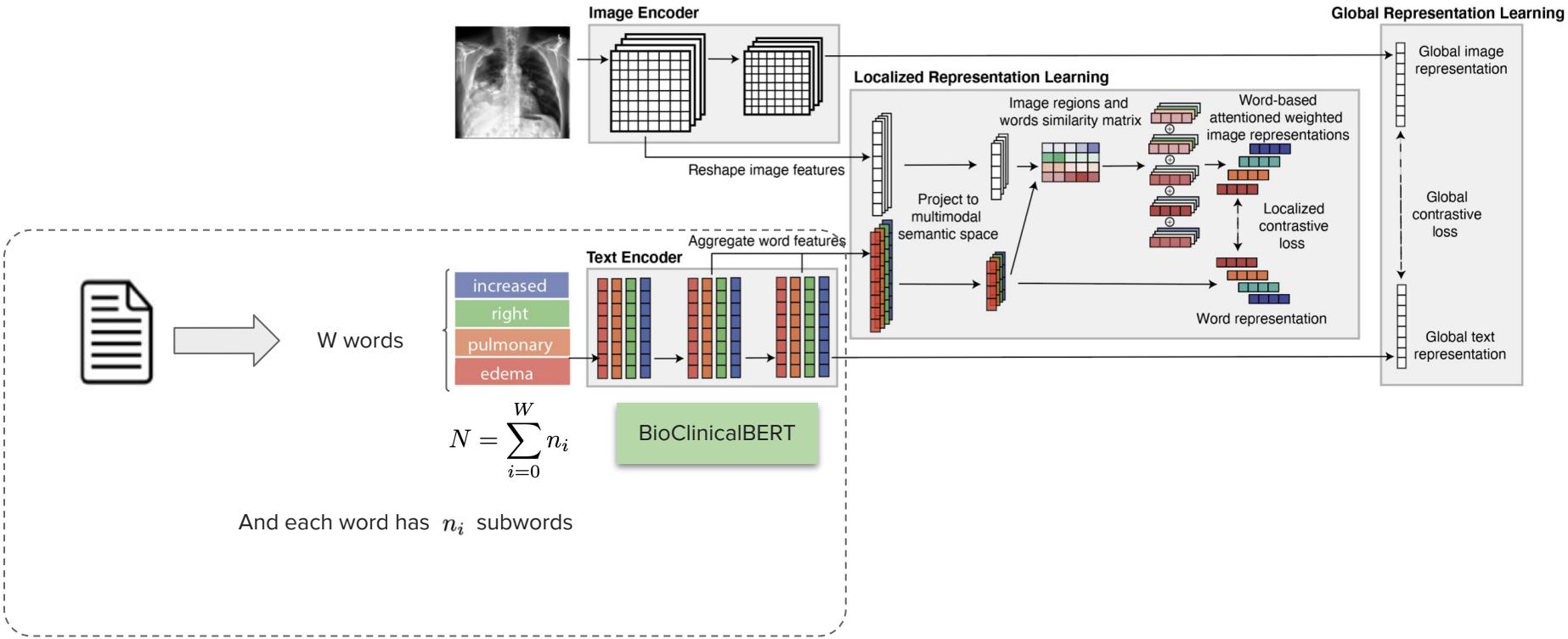


Huang et *al*. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. ICCV'21

Improving the pre-training

GLoRIA

Integrate global and local representations



$$N = \sum_{i=0}^{W} n_i$$

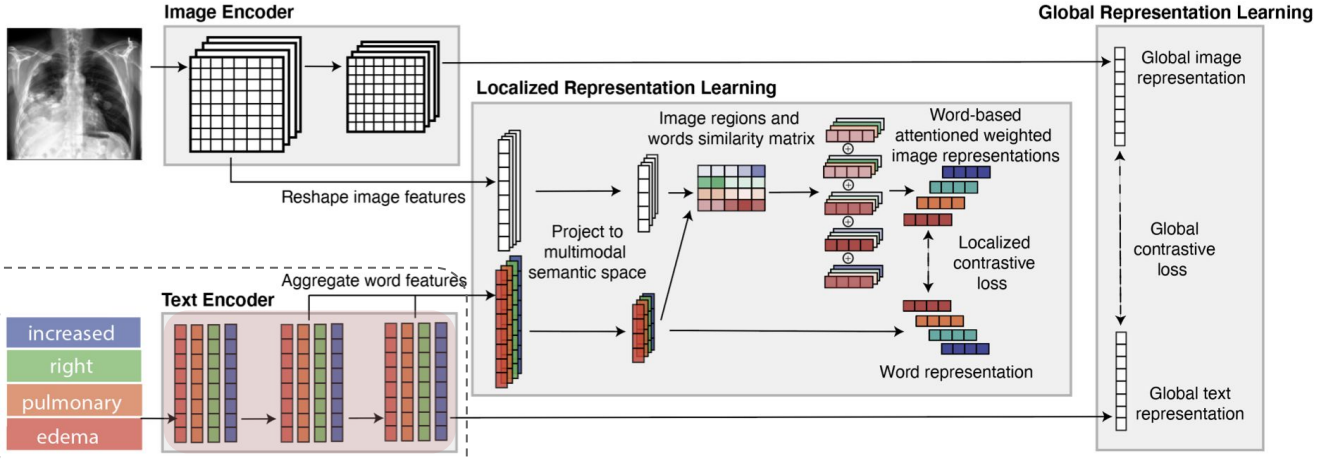
And each word has n_i subwords

Huang et *al*. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. ICCV'21

Improving the pre-training

GLoRIA

Integrate global and local representations



W words

$$N = \sum_{i=0}^{W} n_i$$

BioClinicalBERT

And each word has n_i subwords

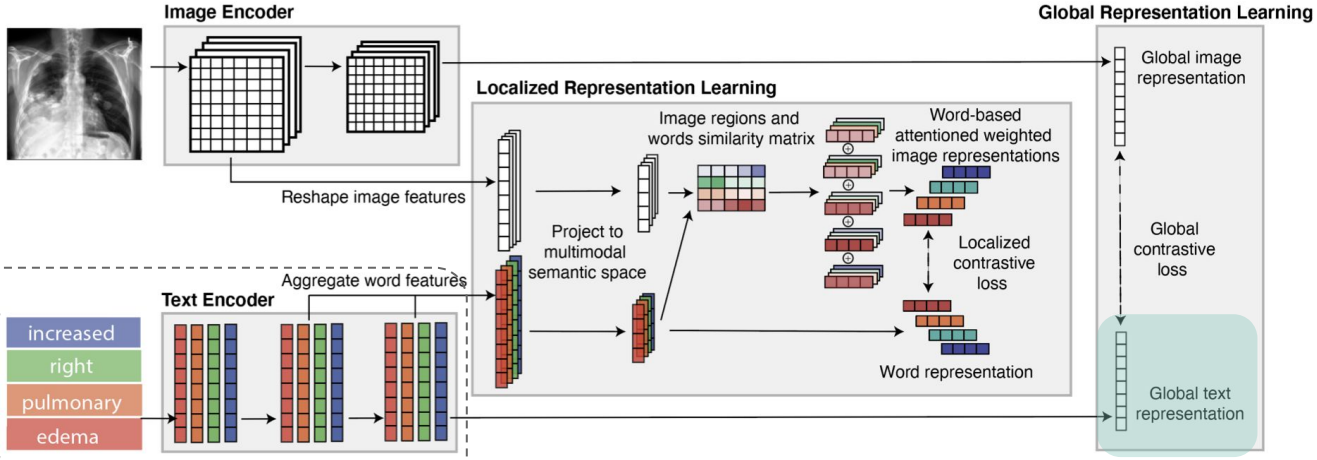
Local features $g_l \in \mathbb{R}^{K \times N}$

Huang et *al*. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. ICCV'21

Improving the pre-training

GLoRIA

Integrate global and local representations



W words

$$N = \sum_{i=0}^{W} n_i$$

BioClinicalBERT

And each word has n_i subwords

Local features $g_l \in \mathbb{R}^{K \times N}$

Global features $g_g = \sum_{i=1}^{N} g_{l_i} \in \mathbb{R}^K$

Huang et *al*. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. ICCV'21

Improving the pre-training

GLoRIA

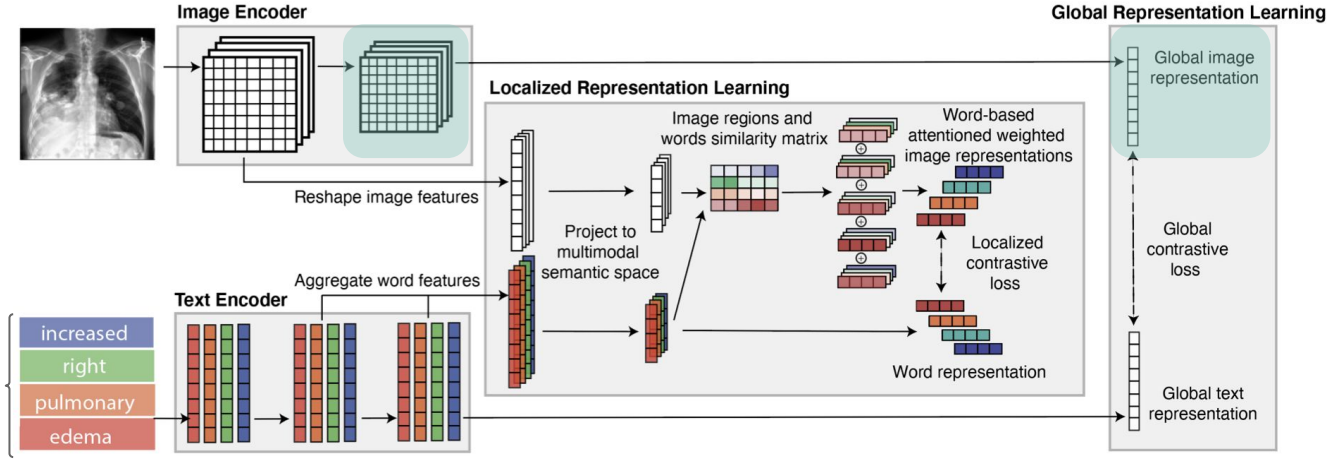
Integrate global and local representations

Global contrastive

Global visual features

$$\mathbf{v}_g = R_{vg}(f_\theta^v(\mathbf{x})) \in \mathbb{R}^D$$

Projection function



Huang et *al*. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. ICCV'21

Improving the pre-training

GLoRIA

Integrate global and local representations

Global contrastive

Global visual features

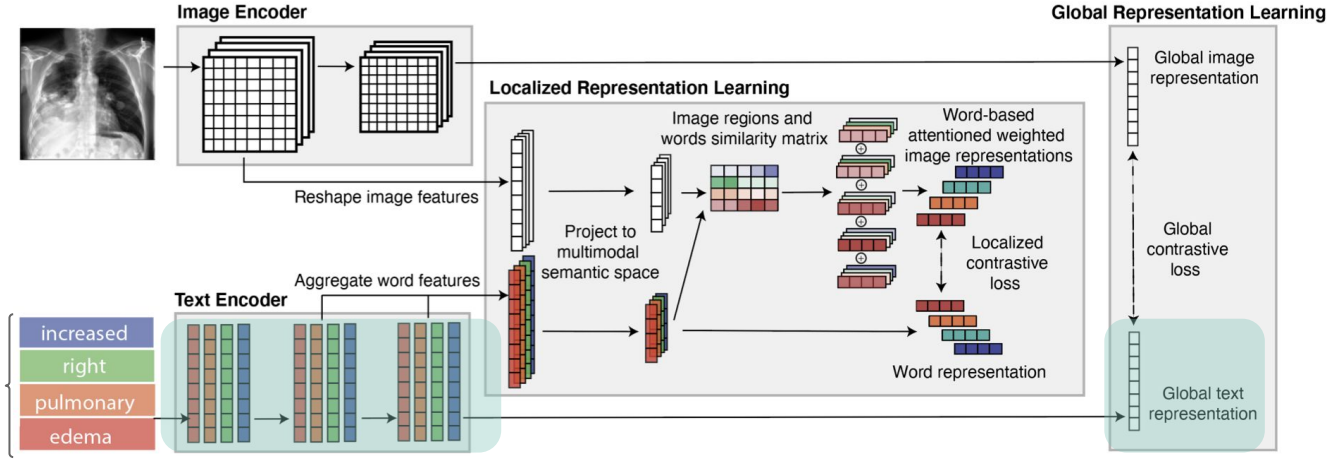
$$\mathbf{v}_g = R_{vg}(f_\theta^v(\mathbf{x})) \in \mathbb{R}^D$$

Projection function

Global text features

$$\mathbf{z}_g = R_{tg}(f_\theta^t(\mathbf{t})) \in \mathbb{R}^D$$

Projection function



Huang et *al*. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. ICCV'21

Improving the pre-training

GLoRIA

Integrate global and local representations

Global contrastive

Global visual features

$$\mathbf{v}_g = R_{vg}(f_\theta^v(\mathbf{x})) \in \mathbb{R}^D$$

Projection function

Global text features

$$\mathbf{z}_g = R_{tg}(f_\theta^t(\mathbf{t})) \in \mathbb{R}^D$$

Projection function

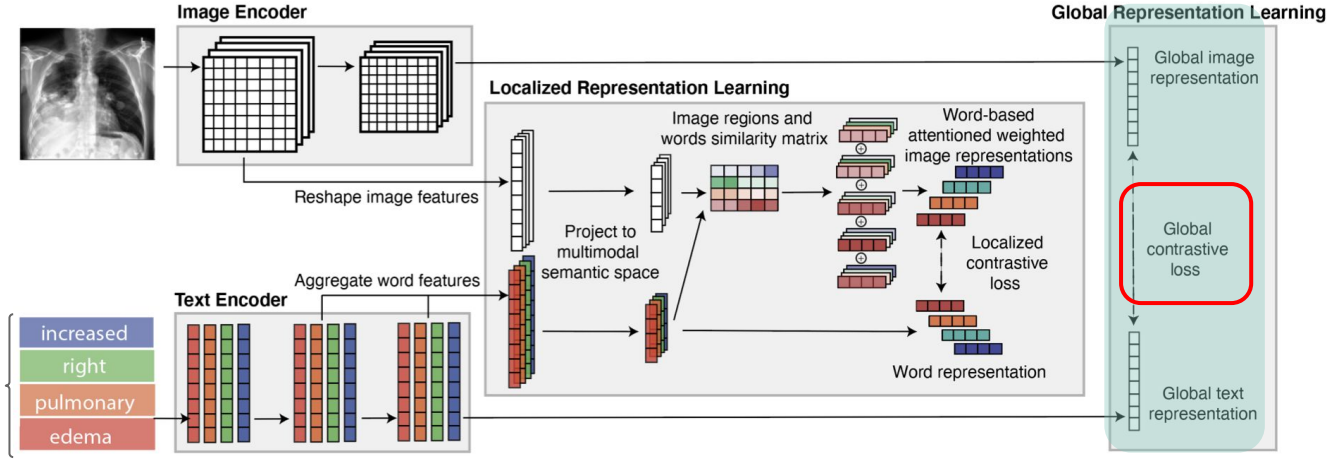


Image-to-text
Global contrastive loss

$$\mathcal{L}_g^{v\to t} = \sum_{i\in\mathcal{B}} -\log \frac{\exp(\langle \mathbf{v}_{gi}, \mathbf{z}_{gi}\rangle/\tau)}{\sum_{j=1}^{N}\exp(\langle \mathbf{v}_{gi}, \mathbf{z}_{gj}\rangle/\tau)}$$

Huang et *al*. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. ICCV'21

Improving the pre-training

GLoRIA

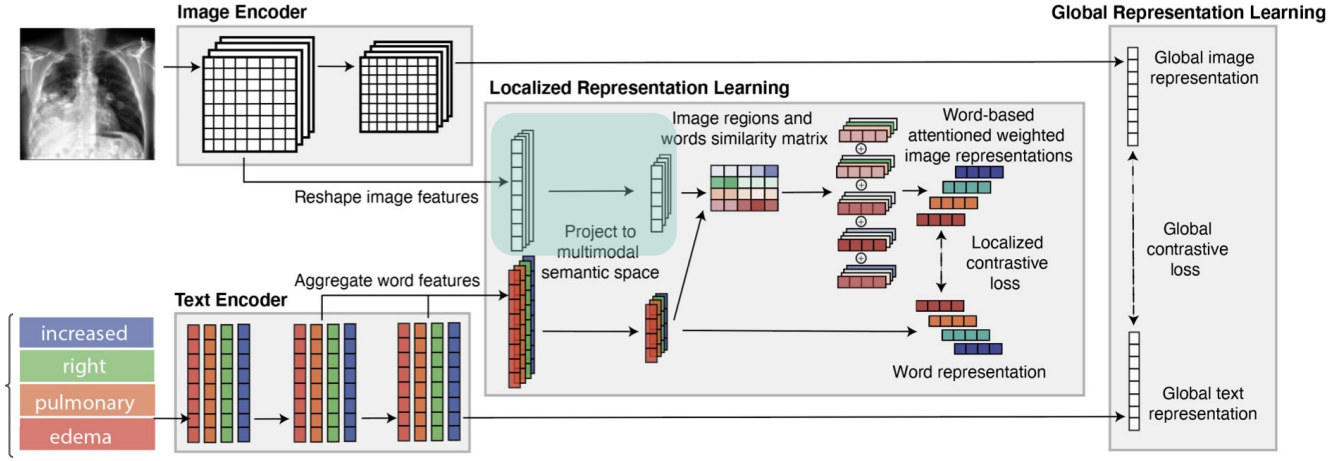
Integrate global and local representations

Local contrastive

Local visual features

$$\mathbf{v}_l = R_{vl}(f_\theta^v(\mathbf{x})) \in \mathbb{R}^{D \times M}$$

Projection (local) function



Huang et *al*. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. ICCV'21

Improving the pre-training

GLoRIA

Integrate global and local representations

Local contrastive

Local visual features

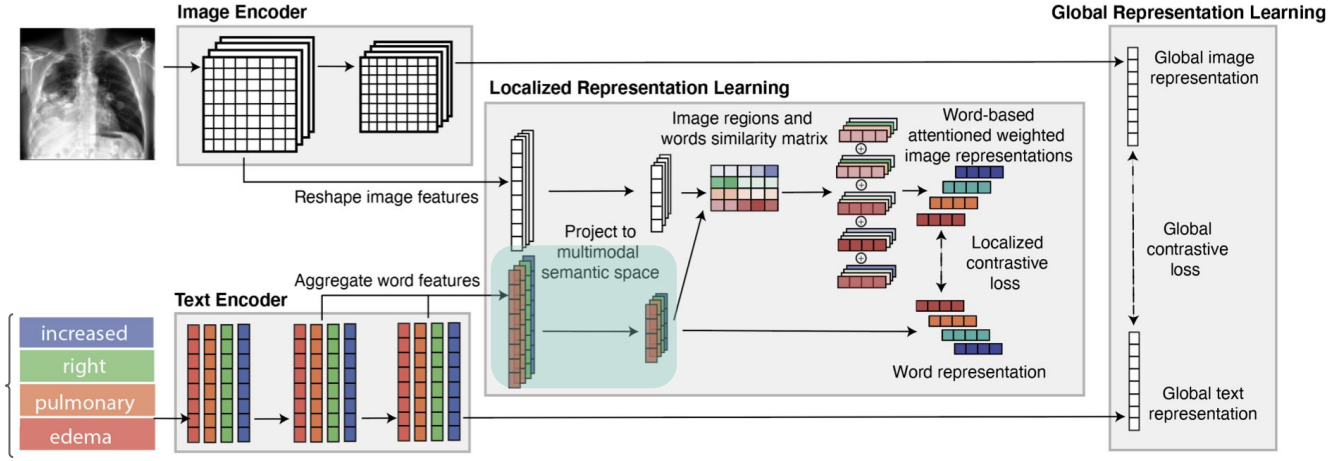
$$\mathbf{v}_l = R_{vl}(f_\theta^v(\mathbf{x})) \in \mathbb{R}^{D \times M}$$

Projection (local) function

Local text features

$$\mathbf{z}_l = R_{zl}(f_\theta^t(\mathbf{t})) \in \mathbb{R}^{D \times W}$$

Projection (local) function



Huang et *al*. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. ICCV'21

Improving the pre-training

GLoRIA

Integrate global and local representations

Local contrastive

Local visual features

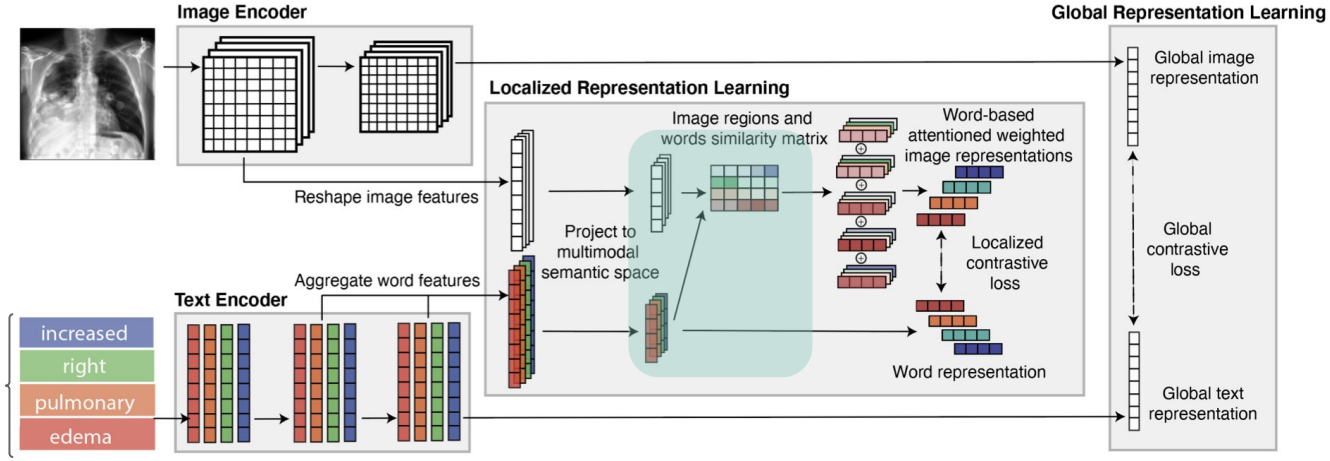
$$\mathbf{v}_l = R_{vl}(f_\theta^v(\mathbf{x})) \in \mathbb{R}^{D\times M}$$

Projection (local) function

Local text features

$$\mathbf{z}_l = R_{zl}(f_\theta^t(\mathbf{t})) \in \mathbb{R}^{D\times W}$$

Projection (local) function



Similarity matrix

$$\mathbf{s} = \mathbf{v}_l^\top \mathbf{z}_l \in \mathbb{R}^{M\times W}$$

Huang et *al*. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. ICCV'21

Improving the pre-training

GLoRIA

Integrate global and local representations

Local contrastive

Local visual features

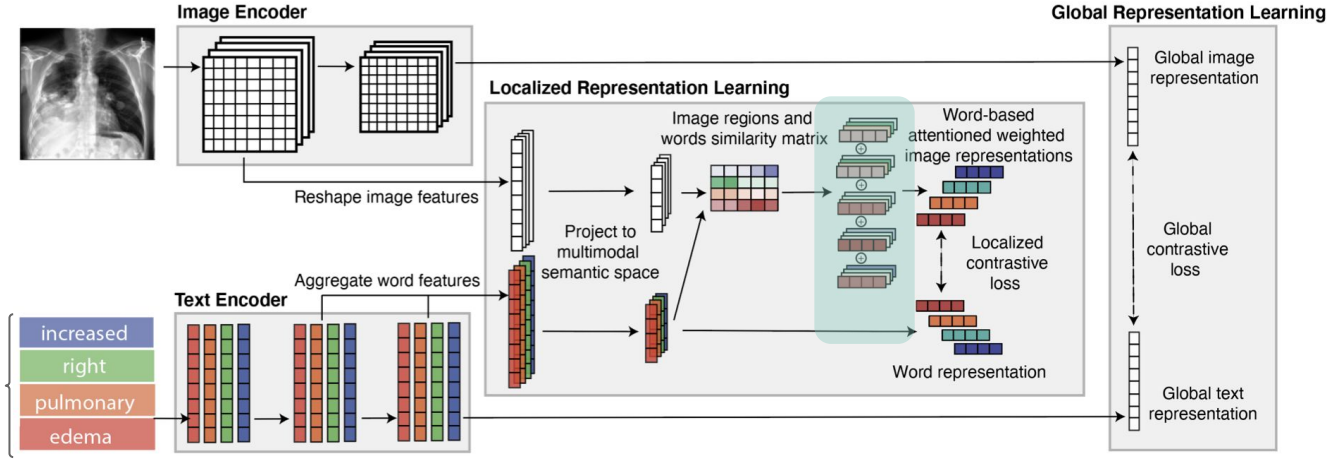
$$\mathbf{v}_l = R_{vl}(f_\theta^v(\mathbf{x})) \in \mathbb{R}^{D\times M}$$

Projection (local) function

Local text features

$$\mathbf{z}_l = R_{zl}(f_\theta^t(\mathbf{t})) \in \mathbb{R}^{D\times W}$$

Projection (local) function



Similarity matrix

$$\mathbf{s} = \mathbf{v}_l^\top \mathbf{z}_l \in \mathbb{R}^{M\times W}$$

Attention matrix
(normalized similarity)

$$\mathbf{a}_{ij} = \frac{\exp(\mathbf{s}_{ij}/\tau')}{\sum_{m=1}^{M}\exp(\mathbf{s}_{im}/\tau')}$$

Huang et *al*. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. ICCV'21

Improving the pre-training

GLoRIA

Integrate global and local representations

Local contrastive

Local visual features

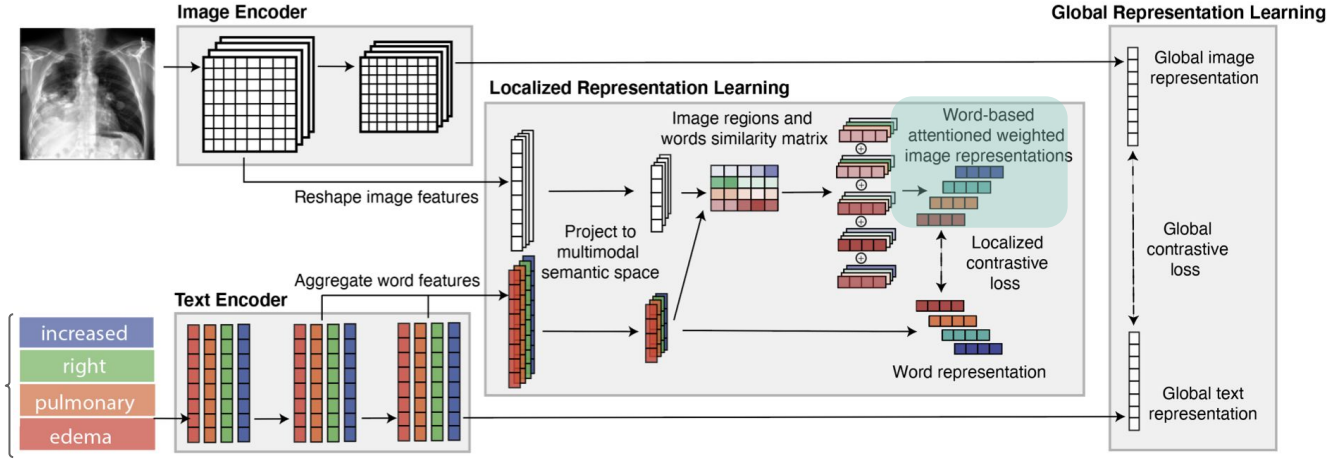
$$\mathbf{v}_l = R_{vl}(f_\theta^v(\mathbf{x})) \in \mathbb{R}^{D\times M}$$

Projection (local) function

Local text features

$$\mathbf{z}_l = R_{zl}(f_\theta^t(\mathbf{t})) \in \mathbb{R}^{D\times W}$$

Projection (local) function



Similarity matrix

$$\mathbf{s} = \mathbf{v}_l^\top \mathbf{z}_l \in \mathbb{R}^{M\times W}$$

Attention matrix
(normalized similarity)

$$\mathbf{a}_{ij} = \frac{\exp(\mathbf{s}_{ij}/\tau')}{\sum_{m=1}^{M}\exp(\mathbf{s}_{im}/\tau')}$$

Context-aware image representation
(weighted sum of all image sub-regions to a given word)

$$\mathbf{c}_i = \sum_{j=1}^{M} \mathbf{a}_{ij}\mathbf{v}_{l_j}$$

Huang et *al*. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. ICCV'21

Improving the pre-training

GLoRIA

Integrate global and local representations

Local contrastive

Local visual features

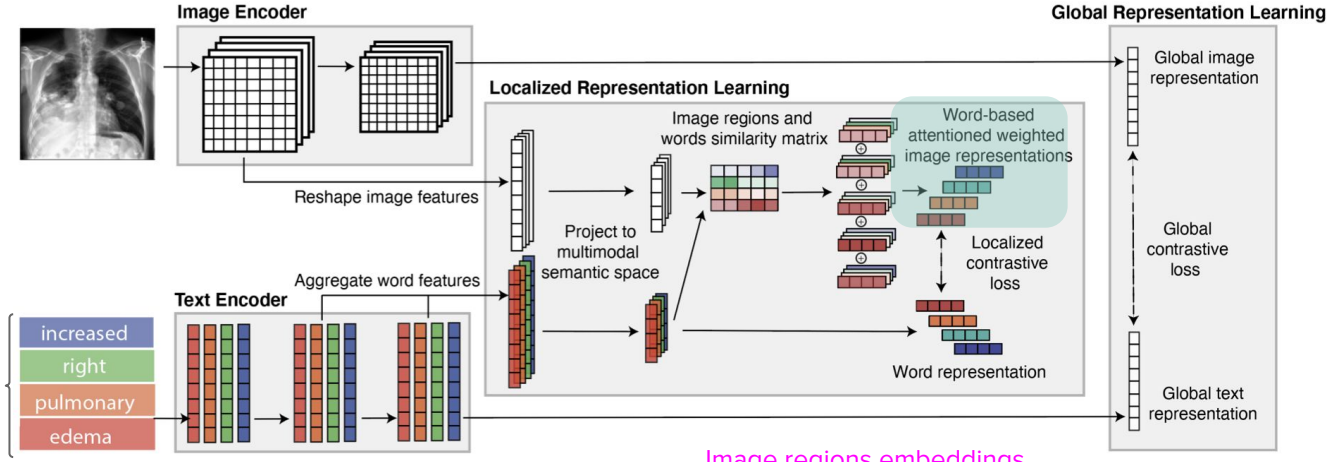
$$\mathbf{v}_l = R_{vl}(f_\theta^v(\mathbf{x})) \in \mathbb{R}^{D\times M}$$

Projection (local) function

Local text features

$$\mathbf{z}_l = R_{zl}(f_\theta^t(\mathbf{t})) \in \mathbb{R}^{D\times W}$$

Projection (local) function

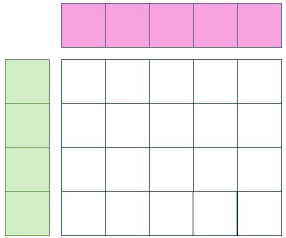


What does **Context-aware image representation** mean?

$$\mathbf{c}_i = \sum_{j=1}^{M} \mathbf{a}_{ij}\mathbf{v}_{l_j}$$

Text embeddings

Image regions embeddings



Similarity (normalized) matrix

Huang et *al*. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. ICCV'21

Improving the pre-training

GLoRIA

Integrate global and local representations

Local contrastive

Local visual features

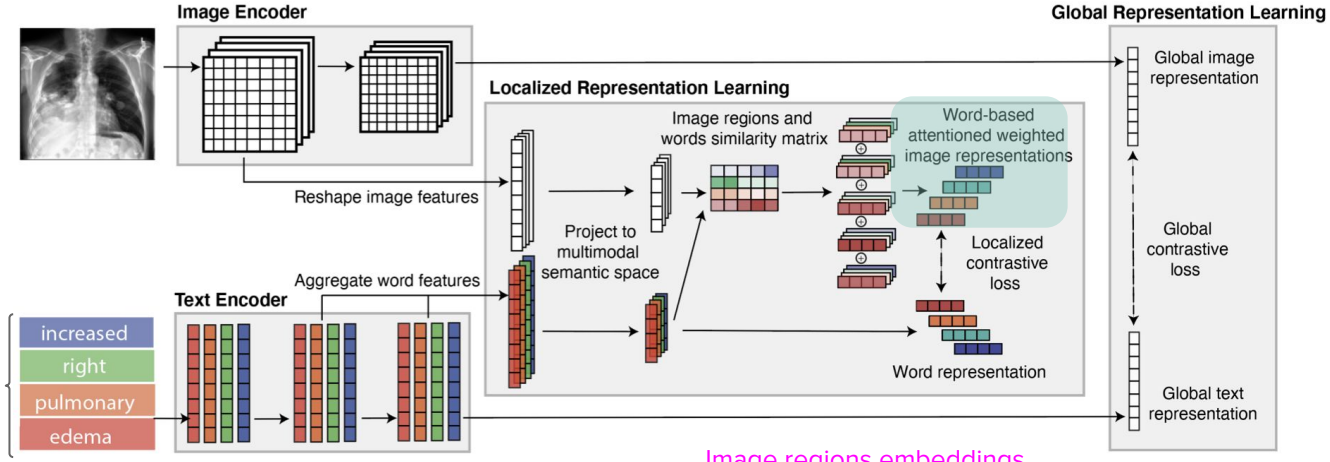
$$\mathbf{v}_l = R_{vl}(f_\theta^v(\mathbf{x})) \in \mathbb{R}^{D\times M}$$

Projection (local) function

Local text features

$$\mathbf{z}_l = R_{zl}(f_\theta^t(\mathbf{t})) \in \mathbb{R}^{D\times W}$$

Projection (local) function

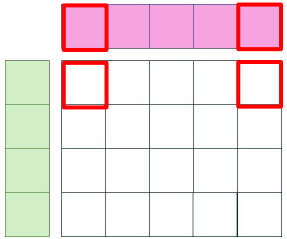


What does **Context-aware image representation** mean?

$$\mathbf{c}_i = \sum_{j=1}^{M} \mathbf{a}_{ij}\mathbf{v}_{l_j}$$

Text embeddings

Image regions embeddings



$c_1=a_{11}z_1+\ldots+a_{1n}z_n$

Huang et *al*. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. ICCV'21

Improving the pre-training

GLoRIA

Integrate global and local representations

Local contrastive

Local visual features

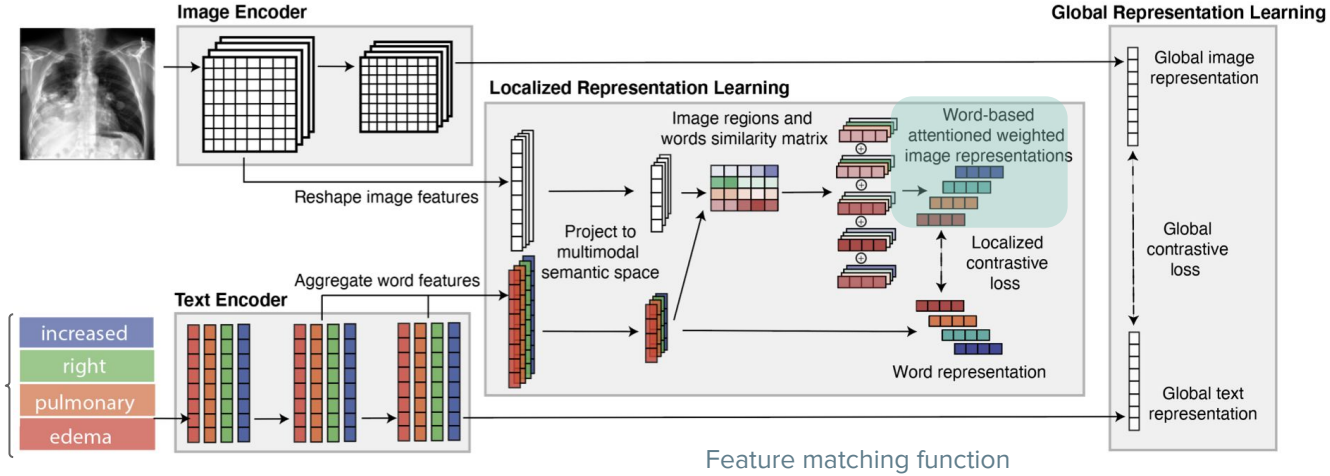
$$\mathbf{v}_l = R_{vl}(f_\theta^v(\mathbf{x})) \in \mathbb{R}^{D\times M}$$

Projection (local) function

Local text features

$$\mathbf{z}_l = R_{zl}(f_\theta^t(\mathbf{t})) \in \mathbb{R}^{D\times W}$$

Projection (local) function



What does **Context-aware image representation** mean?

$$\mathbf{c}_i = \sum_{j=1}^{M} \mathbf{a}_{ij}\mathbf{v}_{l_j}$$

$$Z(\mathbf{x},\mathbf{t}) = \log(\sum_{j=1}^{W} \exp(\langle \mathbf{c}_j, \rangle/\tau_2))^{\tau_2}$$

Huang et *al*. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. ICCV'21

Improving the pre-training

GLoRIA

Integrate global and local representations

Local contrastive

Local visual features

$$\mathbf{v}_l = R_{vl}(f_\theta^v(\mathbf{x})) \in \mathbb{R}^{D\times M}$$

Projection (local) function

Local text features

$$\mathbf{z}_l = R_{zl}(f_\theta^t(\mathbf{t})) \in \mathbb{R}^{D\times W}$$

Projection (local) function

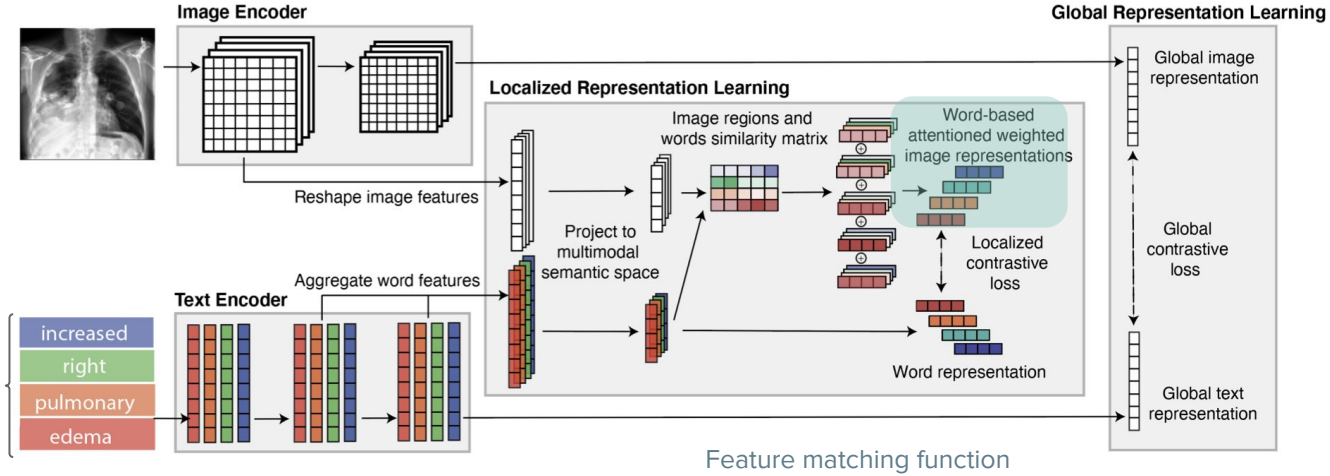


Image-to-text
Local contrastive loss

$$\mathcal{L}_l^{v\to t} = \sum_{i\in\mathcal{B}} -\log \frac{\exp(Z(\mathbf{x}_i,\mathbf{t}_i)/\tau')}{\sum_{j=1}^{N}\exp(Z(\mathbf{x}_i,\mathbf{t}_j)/\tau')}$$

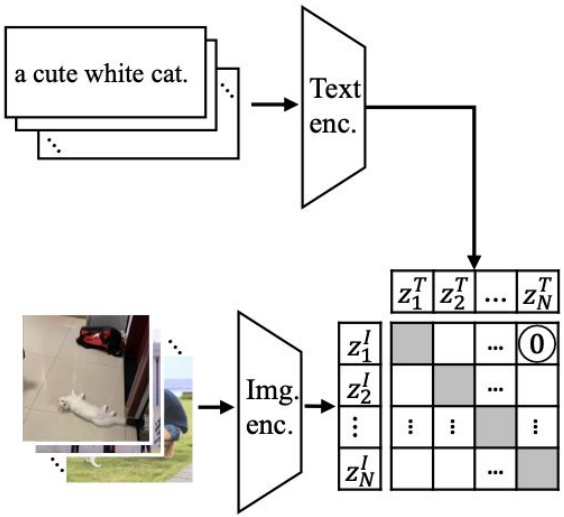
Huang et *al*. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. ICCV'21

Improving the pre-training

DeCLIP

Expanding image-text pairs

Original CLIP Alignment



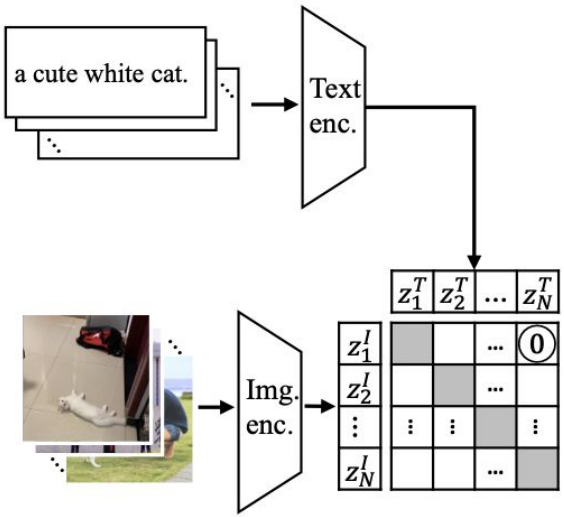
Li et *al*. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm, ICLR'22

Improving the pre-training

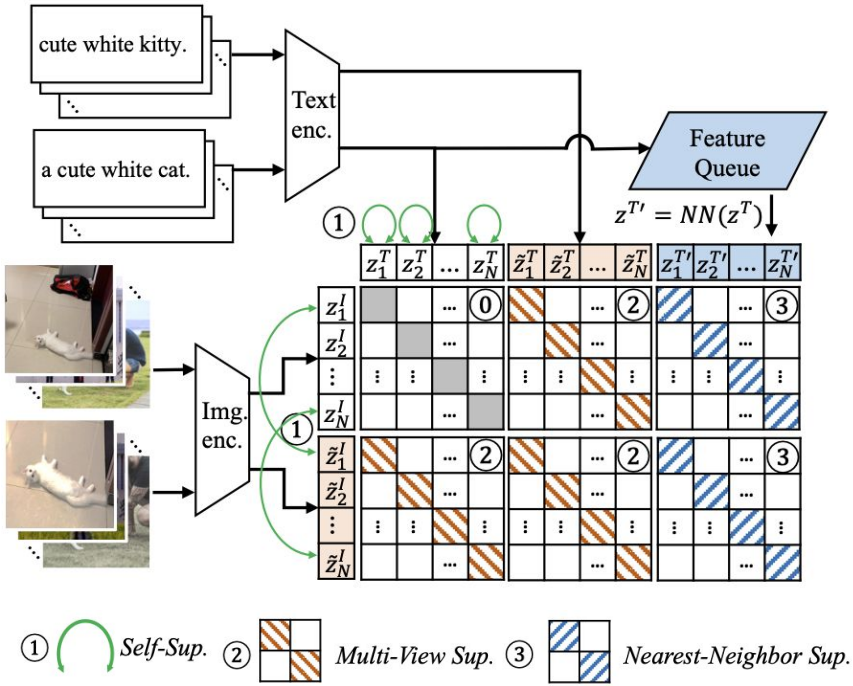
DeCLIP

Expanding image-text pairs

Original CLIP Alignment



DeCLIP



Li et *al*. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm, ICLR'22

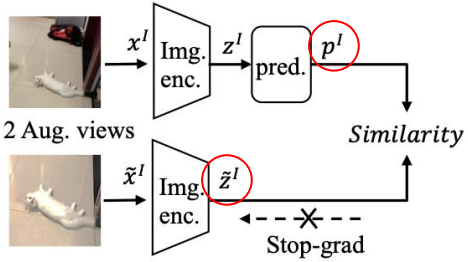
Improving the pre-training

DeCLIP

Expanding image-text pairs

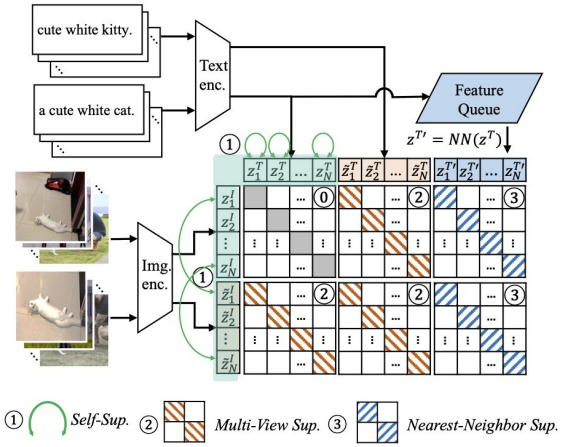
1- Self-supervision with each modality

Image



Maximize similarity

[SimSiam, CVPR'21]



Li et *al*. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm, ICLR'22

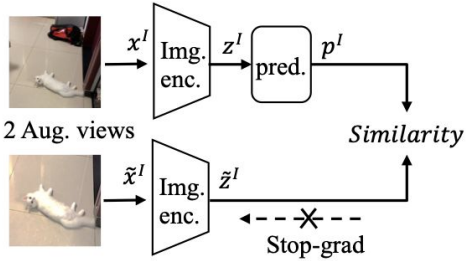
Improving the pre-training

DeCLIP

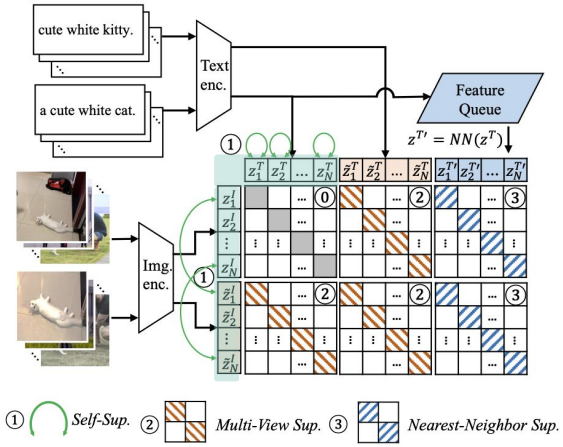
Expanding image-text pairs

1- Self-supervision with each modality

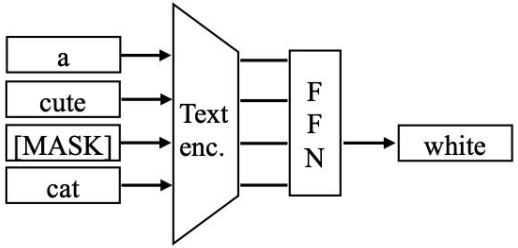
Image



[SimSiam, CVPR'21]



Text



BERT

Objective: predict the original token.

a cute white cat

80% of the times replaced by [mask]

10% of the times replaced by a random token

10% of the times replaced by the correct token

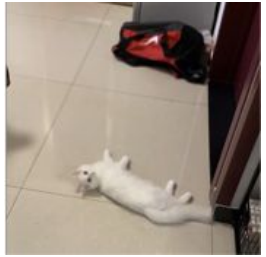
Li et *al*. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm, ICLR'22

Improving the pre-training

DeCLIP

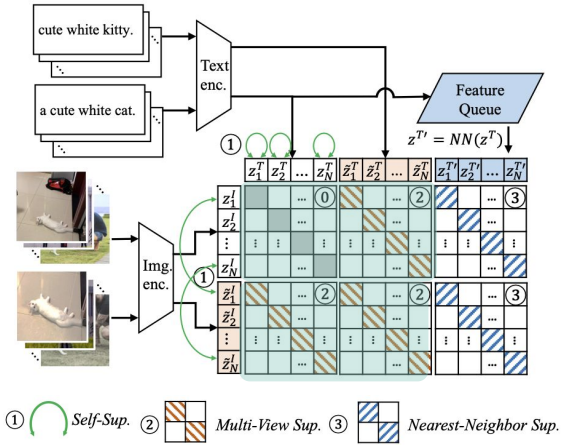
Expanding image-text pairs

2- Multi-view supervision



"A cute white cat"

The target object only occupies a small part of the picture



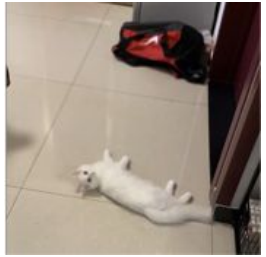
Li et *al*. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm, ICLR'22

Improving the pre-training

DeCLIP

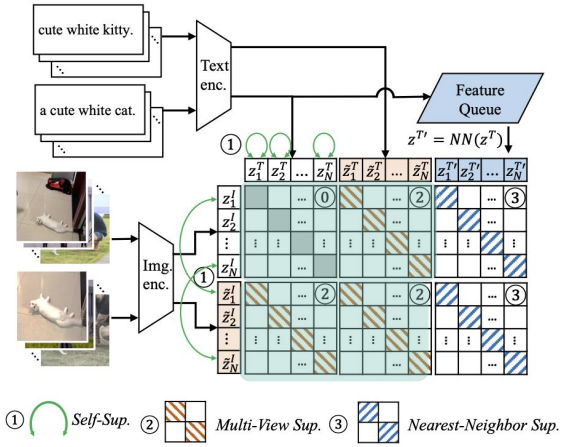
Expanding image-text pairs

2- Multi-view supervision

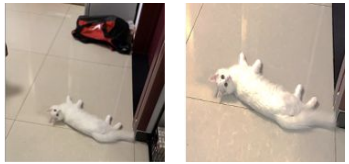


"A cute white cat"

The target object only occupies a small part of the picture



Generate an augmented view (image)



$$\mathbf{x}, \mathbf{x}' \rightarrow \mathbf{v}, \mathbf{v}'$$

Generate an augmented view (of the text too)

$$\mathbf{t}, \mathbf{t}' \rightarrow \mathbf{z}, \mathbf{z}'$$

[EDA, EMNLP'19]

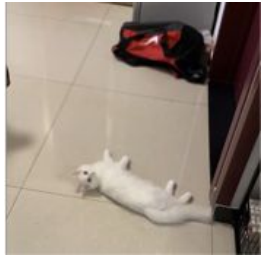
Li et *al*. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm, ICLR'22

Improving the pre-training

DeCLIP

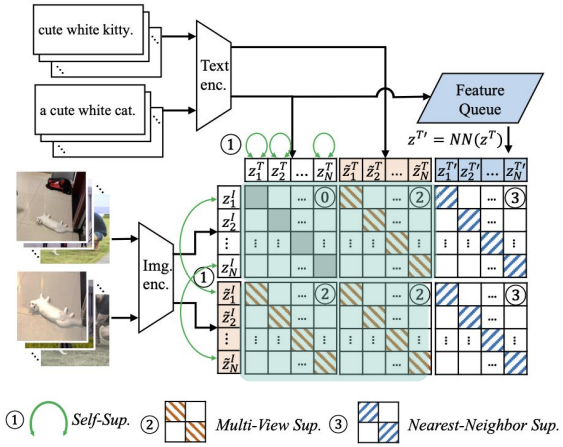
Expanding image-text pairs

2- Multi-view supervision



"A cute white cat"

The target object only occupies a small part of the picture



$$\mathcal{L}_{MVS} = \mathcal{L}_{CLIP}(\mathbf{v}, \mathbf{z}) + \mathcal{L}_{CLIP}(\mathbf{v}, \mathbf{z}') + \mathcal{L}_{CLIP}(\mathbf{v}', \mathbf{z}) + \mathcal{L}_{CLIP}(\mathbf{v}', \mathbf{z}')$$

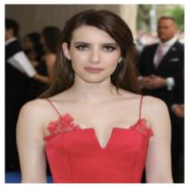
Li et *al*. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm, ICLR'22

Improving the pre-training

DeCLIP

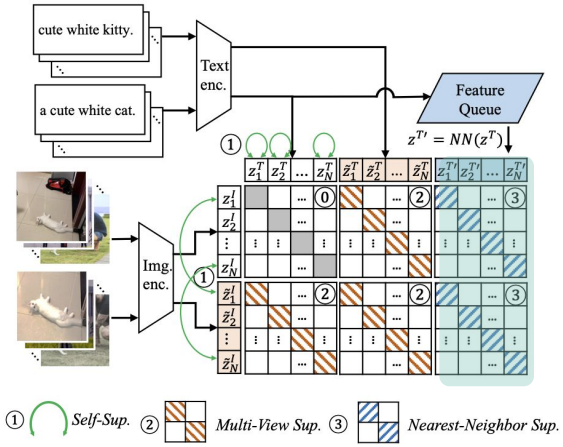
Expanding image-text pairs

3- Nearest-neighbor Supervision



Origin	actor opted for a loose side - parted hairstyle when she attended event .	going to see a lot of vintage tractors this week
NN	actor wore her hair in face - framing layers during the show .	vintage at tractors a gathering

Examples of Nearest Neighbor from Conceptual Captions dataset



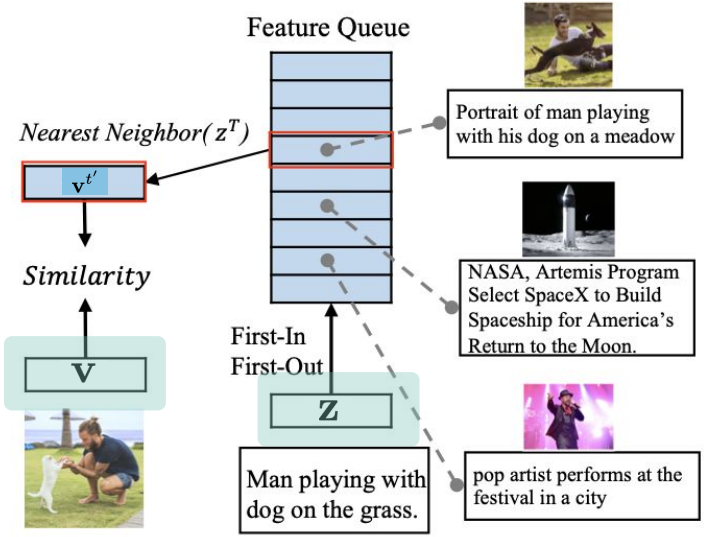
Li et *al*. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm, ICLR'22

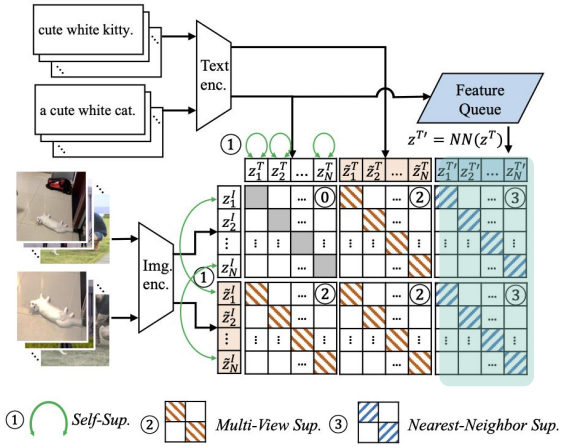
Improving the pre-training

DeCLIP

Expanding image-text pairs

3- Nearest-neighbor Supervision





For a given image-text pair, image and text features are extracted $\mathbf{V}, \mathbf{Z}$

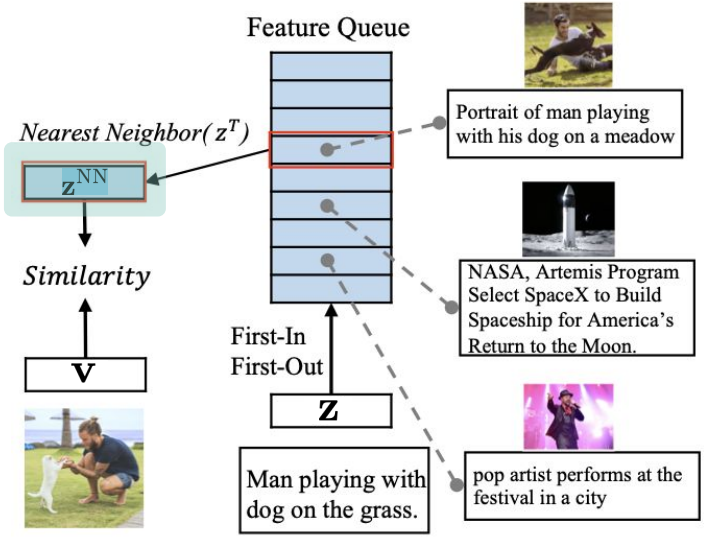
Li et *al*. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm, ICLR'22

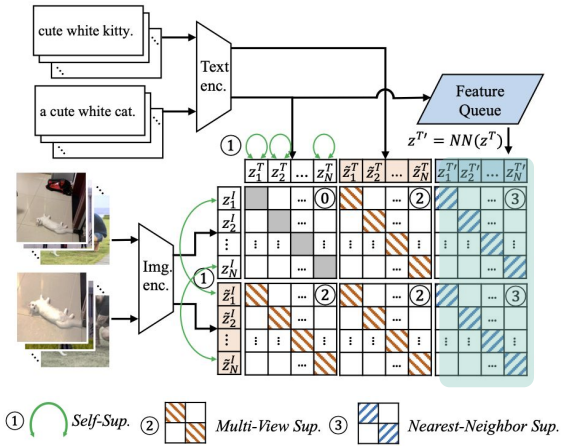
Improving the pre-training

DeCLIP

Expanding image-text pairs

3- Nearest-neighbor Supervision





For a given image-text pair, image and text features are extracted $\mathbf{v}, \mathbf{z}$

Find the closest text representation from K neighbours $\mathbf{z}^{NN}$

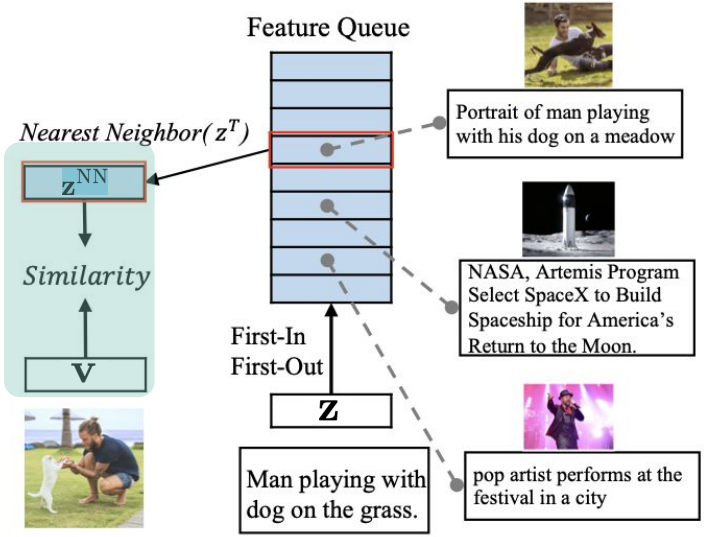
Li et *al*. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm, ICLR'22

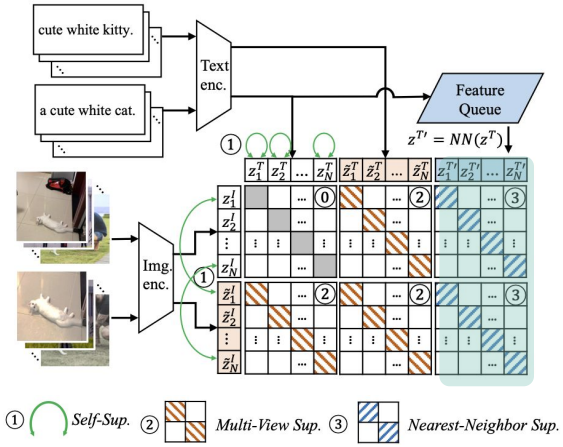
Improving the pre-training

DeCLIP

Expanding image-text pairs

3- Nearest-neighbor Supervision





For a given image-text pair, image and text features are extracted $\mathbf{v}, \mathbf{z}$

Find the closest text representation from K neighbours $\mathbf{z}^{NN}$

Apply a contrastive loss between these representations

Li et *al*. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm, ICLR'22

Improving the pre-training

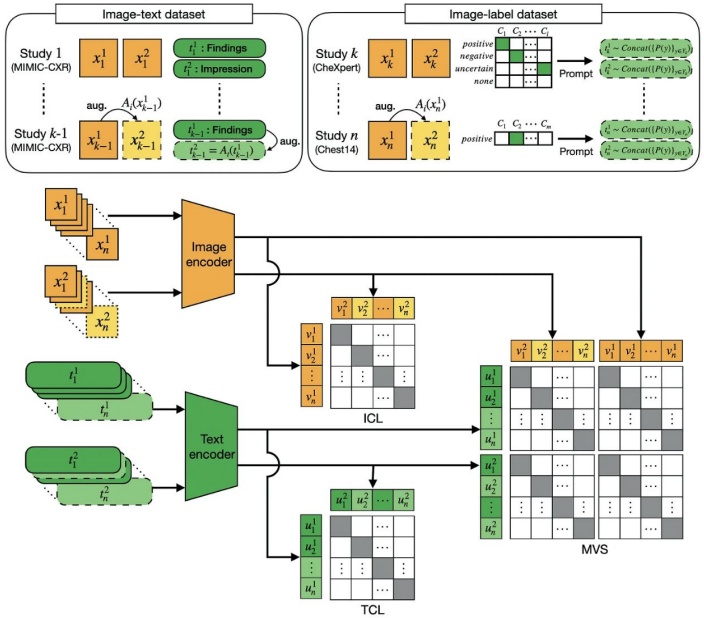
CXR-CLIP

Use-case of DeCLIP
Expanding image-label pairs

$$\text{Study} = \{(\mathbf{x}_n, \mathbf{t}_n)\}_{n=1}^{N}$$

$$\text{Study (image-label)} = \{(\mathbf{x}_m, \mathbf{y}_m)\}_{n=1}^{M}$$

- *For a given study, there may be multiple views from the same patient: x_n^1 and x_n^2*
- *If there is only one image, it is augmented via A()*



You et *al*. CXR-CLIP: Toward large scale chest X-ray language-image pre-training, MICCAI'23

Improving the pre-training

CXR-CLIP

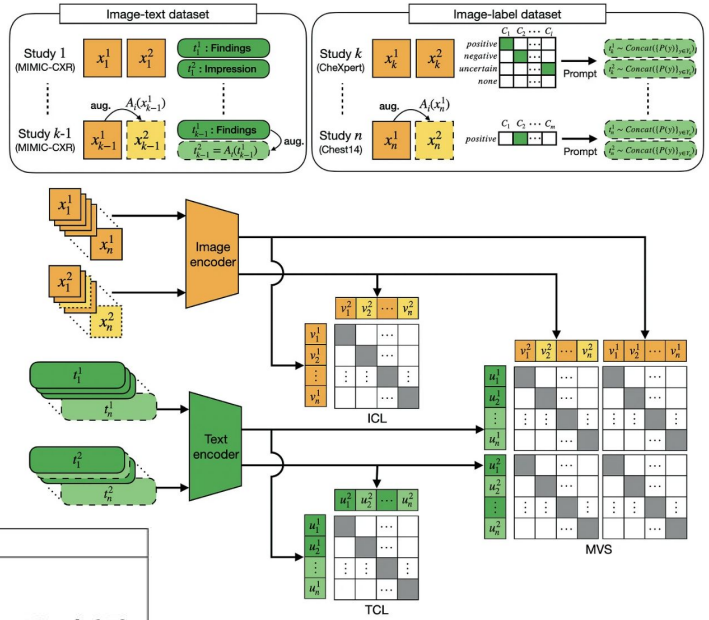
Expanding image-label pairs

$$\text{Study} = \{(\mathbf{x}_n, \mathbf{t}_n)\}_{n=1}^{N}$$

$$\text{Study (image-label)} = \{(\mathbf{x}_m, \mathbf{y}_m)\}_{n=1}^{M}$$

$$\mathbf{t} = \text{Concat}(\{p \sim P(y)\}_{y \in \mathcal{Y}})$$

p is a sampled prompt from *P(y)*,
which is a set of prompts given *y*



	Positive templates	Negative templates
Default	[{E}., There is {E}., {E} is [present, seen, noted]., the presence of {E} is [seen, noted].]	[[There is, ( )] + [no {E}., no radiographic evidence for {E}., no [visible, definite, obvious, appreciable, evident] {E}., no [convincing, definite, ( )] evidence of {E}., no convincing signs of {E}.], No {E} is [visible, present, noted].]
Edema Pneumonia	[Default Positive templates, Findings are + [suggesting, compatible with, suggestive of, representing] + {E}.]	
Cardiomegaly	[heart size, cardiac size, cardiac silhouette, cardiac shadow, cardiac contour] + [is, appears] + [enlarged, increased].	[heart size, cardiac size, cardiac silhouette, cardiac shadow, cardiac contour] + [is, appears] + [normal, within normal limits, unremarkable].

You et *al*. CXR-CLIP: Toward large scale chest X-ray language-image pre-training, MICCAI'23

Improving the pre-training

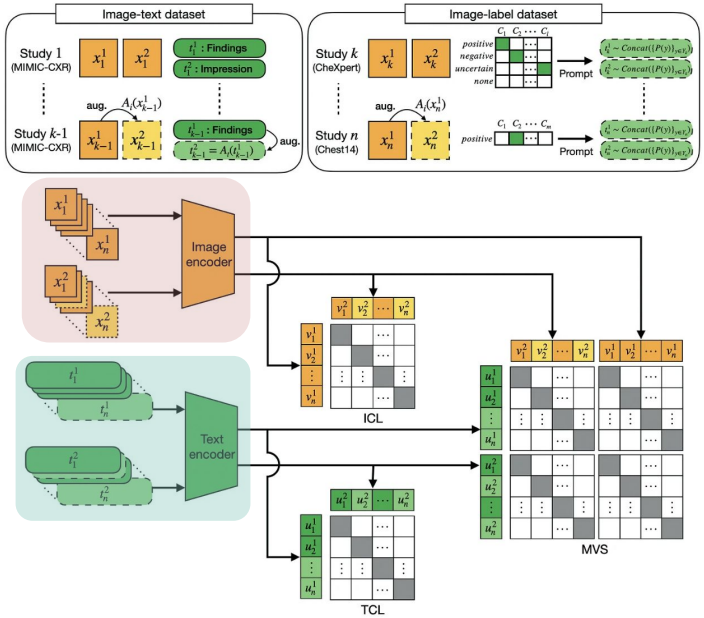
CXR-CLIP

Expanding image-label pairs

Following prior methods, the embeddings can be obtained as:

$$\mathcal{V}^1 = \{\mathbf{v}^1\}_{b=1}^{B}$$
$$\mathcal{V}^2 = \{\mathbf{v}^2\}_{b=1}^{B}$$

$$\mathcal{Z}^1 = \{\mathbf{z}^1\}_{b=1}^{B}$$
$$\mathcal{Z}^2 = \{\mathbf{z}^2\}_{b=1}^{B}$$



You et *al*. CXR-CLIP: Toward large scale chest X-ray language-image pre-training, MICCAI'23

Improving the pre-training

CXR-CLIP

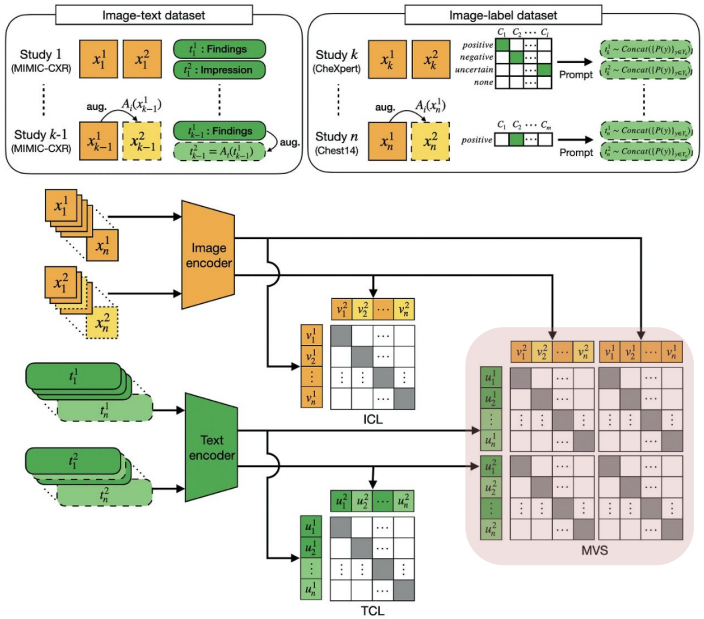
Expanding image-label pairs

Following prior methods, the embeddings can be obtained as:

$$\mathcal{V}^1 = \{\mathbf{v}^1\}_{b=1}^{B}$$
$$\mathcal{V}^2 = \{\mathbf{v}^2\}_{b=1}^{B}$$

$$\mathcal{Z}^1 = \{\mathbf{z}^1\}_{b=1}^{B}$$
$$\mathcal{Z}^2 = \{\mathbf{z}^2\}_{b=1}^{B}$$

$$\mathcal{L}_{MVS} = \frac{1}{4}(\mathcal{L}_{CLIP}(\mathcal{V}^1, \mathcal{Z}^1) + (\mathcal{L}_{CLIP}(\mathcal{V}^1, \mathcal{Z}^2) + (\mathcal{L}_{CLIP}(\mathcal{V}^2, \mathcal{Z}^1) + (\mathcal{L}_{CLIP}(\mathcal{V}^2, \mathcal{Z}^2))$$



You et *al*. CXR-CLIP: Toward large scale chest X-ray language-image pre-training, MICCAI'23

Improving the pre-training

CXR-CLIP

Expanding image-label pairs

Following prior methods, the embeddings can be obtained as:

$$\mathcal{V}^1 = \{\mathbf{v}^1\}_{b=1}^{B} \qquad \mathcal{Z}^1 = \{\mathbf{z}^1\}_{b=1}^{B}$$

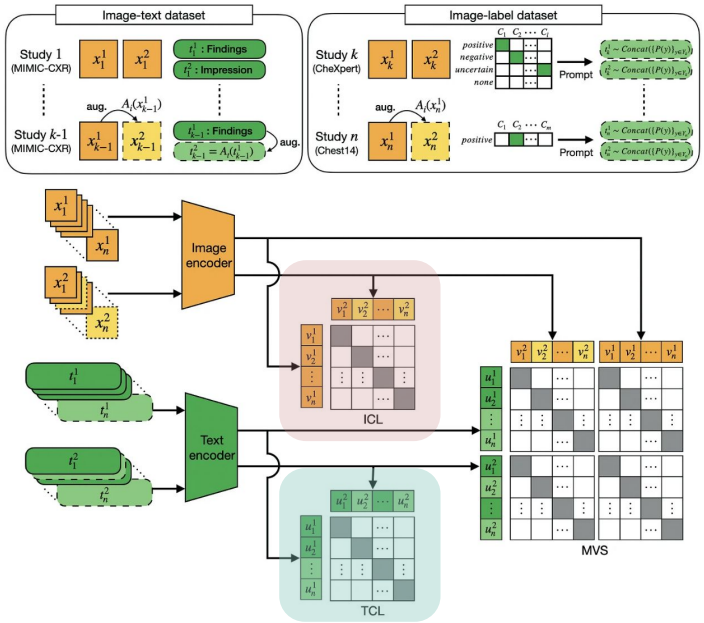
$$\mathcal{V}^2 = \{\mathbf{v}^2\}_{b=1}^{B} \qquad \mathcal{Z}^2 = \{\mathbf{z}^2\}_{b=1}^{B}$$

$$\mathcal{L}_{MVS} = \frac{1}{4}(\mathcal{L}_{CLIP}(\mathcal{V}^1, \mathcal{Z}^1) + (\mathcal{L}_{CLIP}(\mathcal{V}^1, \mathcal{Z}^2) + (\mathcal{L}_{CLIP}(\mathcal{V}^2, \mathcal{Z}^1) + (\mathcal{L}_{CLIP}(\mathcal{V}^2, \mathcal{Z}^2))$$

To learn modality specific characteristics:

Image: $\mathcal{L}_{ICL} = \mathcal{L}_{CLIP}(\mathcal{V}^1, \mathcal{V}^2)$

Text: $\mathcal{L}_{TCL} = \mathcal{L}_{CLIP}(\mathcal{Z}^1, \mathcal{Z}^2)$



You et *al*. CXR-CLIP: Toward large scale chest X-ray language-image pre-training, MICCAI'23

Improving the pre-training

CXR-CLIP

Expanding image-label pairs

Following prior methods, the embeddings can be obtained as:

$$\mathcal{V}^1 = \{\mathbf{v}^1\}_{b=1}^B \qquad \mathcal{Z}^1 = \{\mathbf{z}^1\}_{b=1}^B$$

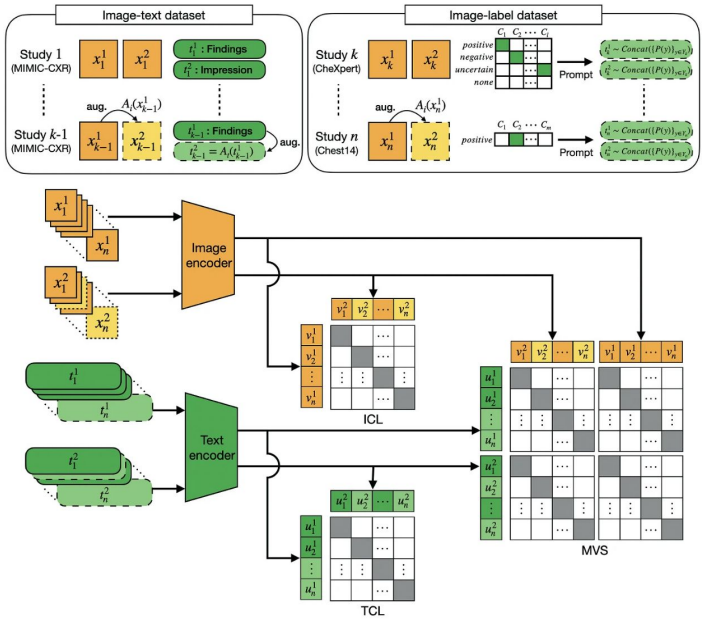
$$\mathcal{V}^2 = \{\mathbf{v}^2\}_{b=1}^B \qquad \mathcal{Z}^2 = \{\mathbf{z}^2\}_{b=1}^B$$

$$\mathcal{L}_{MVS} = \frac{1}{4}(\mathcal{L}_{CLIP}(\mathcal{V}^1, \mathcal{Z}^1) + (\mathcal{L}_{CLIP}(\mathcal{V}^1, \mathcal{Z}^2) + (\mathcal{L}_{CLIP}(\mathcal{V}^2, \mathcal{Z}^1) + (\mathcal{L}_{CLIP}(\mathcal{V}^2, \mathcal{Z}^2))$$

To learn modality specific characteristics:

Image: $\mathcal{L}_{ICL} = \mathcal{L}_{CLIP}(\mathcal{V}^1, \mathcal{V}^2)$

Text: $\mathcal{L}_{TCL} = \mathcal{L}_{CLIP}(\mathcal{Z}^1, \mathcal{Z}^2)$



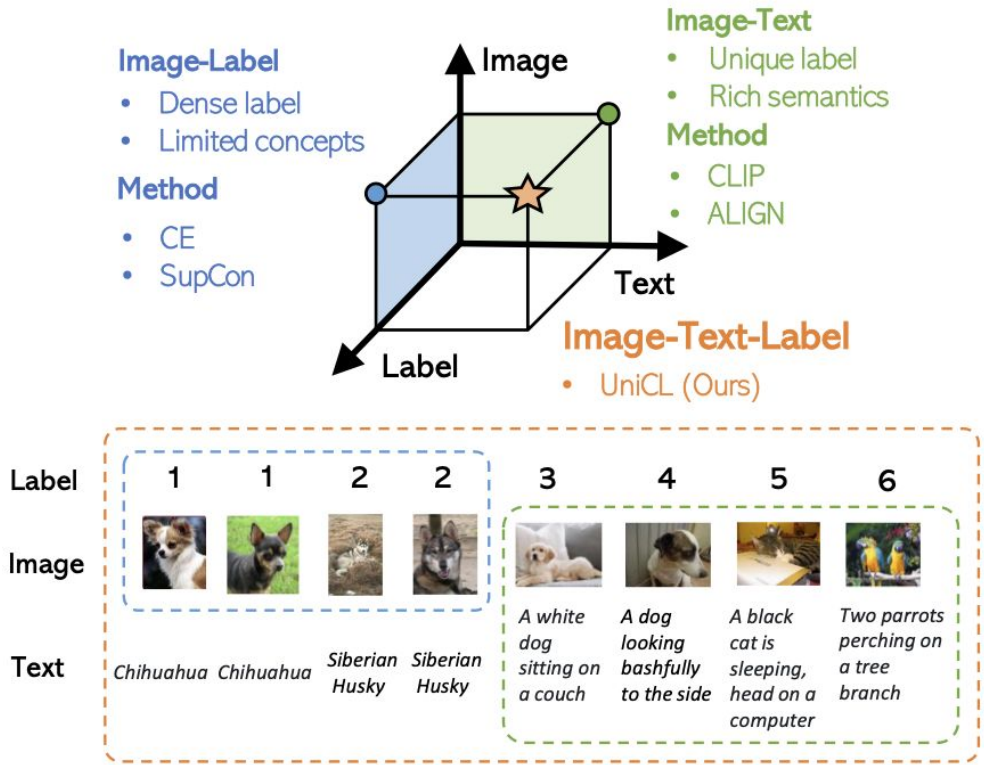
$$\boxed{\mathcal{L} = \mathcal{L}_{MVS} + \lambda_1 \mathcal{L}_{ICL} + \lambda_2 \mathcal{L}_{TCL}}$$

You et *al*. CXR-CLIP: Toward large scale chest X-ray language-image pre-training, MICCAI'23

Improving the pre-training

UniCL

A common image-text-label space

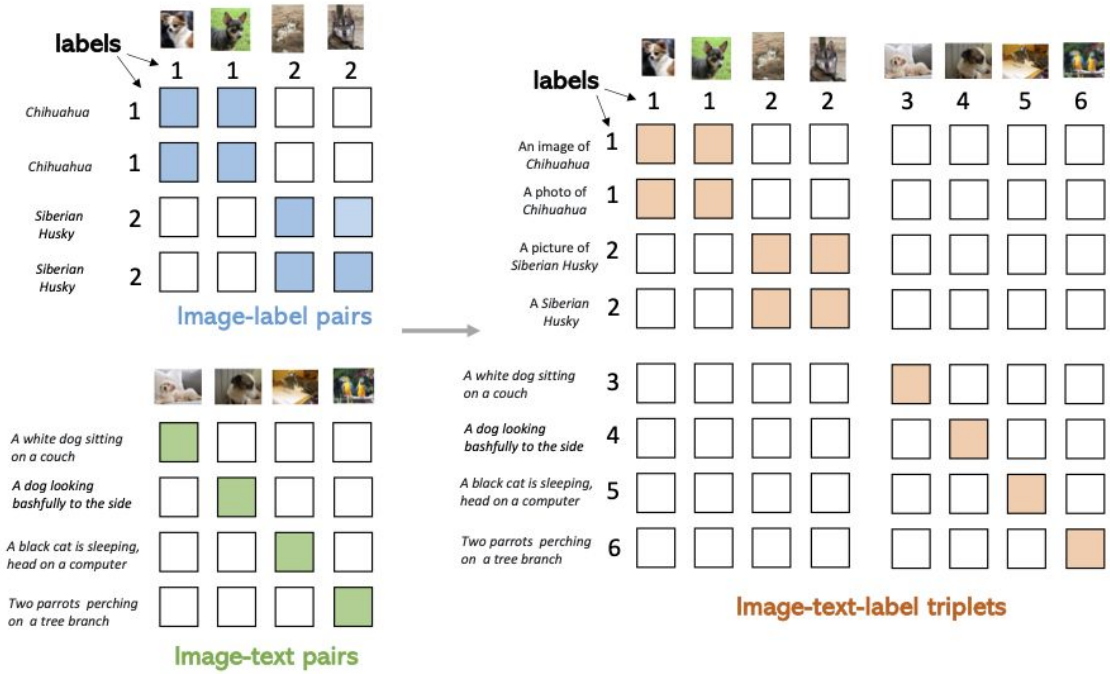


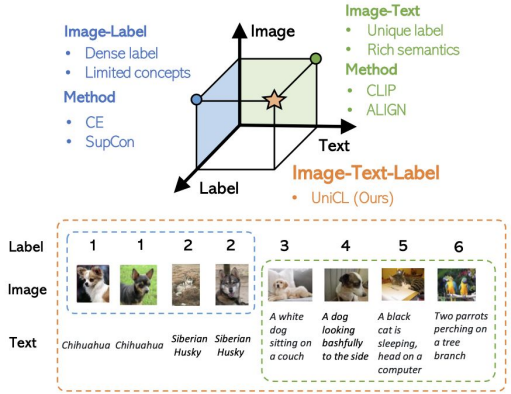
Yang et *al*. Unified Contrastive Learning in Image-Text-Label Space, CVPR'22

Improving the pre-training

UniCL

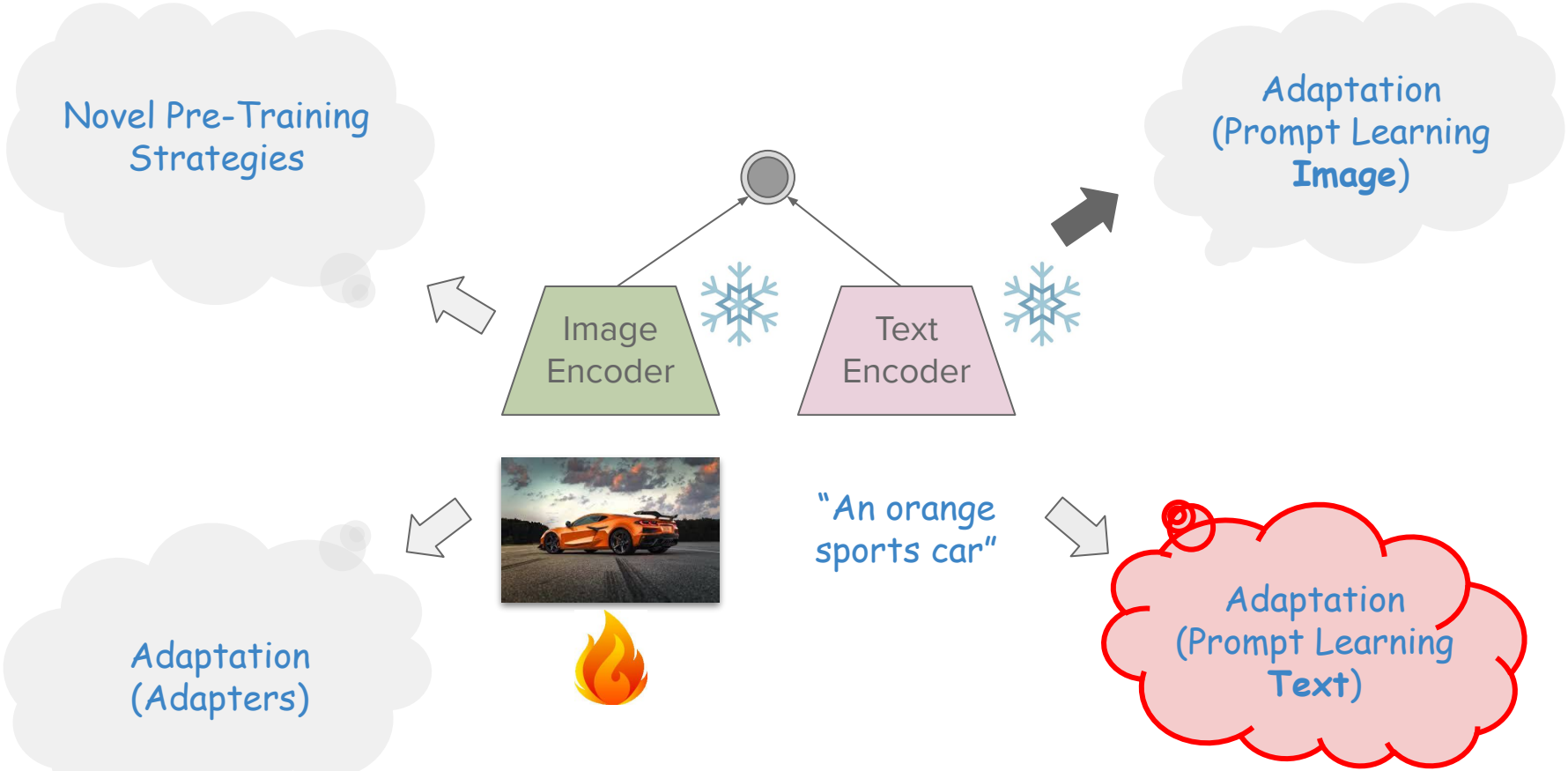
A common image-text-label space





Yang et *al*. Unified Contrastive Learning in Image-Text-Label Space, CVPR'22

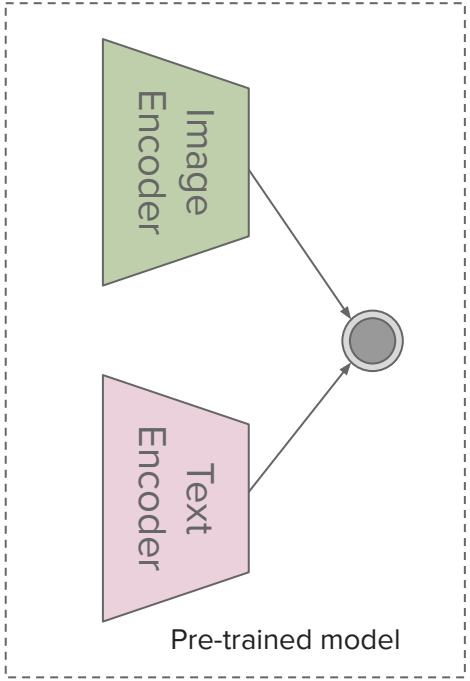
Where we go from CLIP?



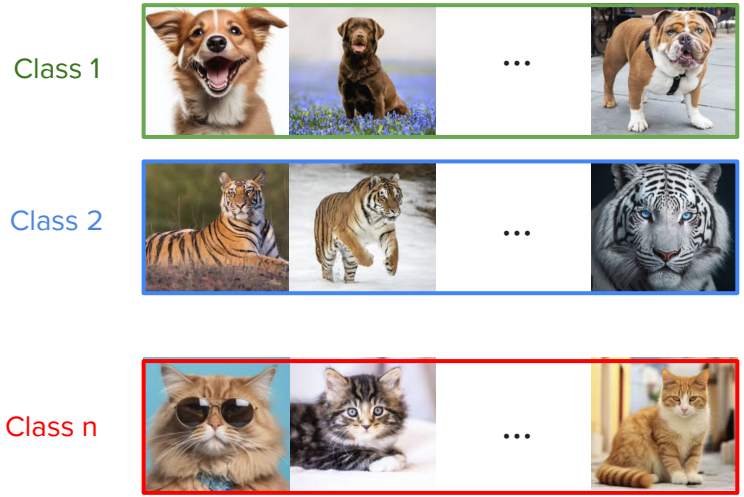
Preliminaries

Setting

Few-shot adaptation



Adapt it to novel classes
(only a few labeled samples)



Only *K* labeled samples per class

$$\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{S}$$

Prompt learning (text)

CoOp

Learning the text prompts

Motivation

Caltech101	Prompt	Accuracy
	a [CLASS].	82.68
	a photo of [CLASS].	80.81
	a photo of a [CLASS].	86.29

Flowers102	Prompt	Accuracy
	a photo of a [CLASS].	60.86
	a flower photo of a [CLASS].	65.81
	a photo of a [CLASS], a type of flower.	66.14

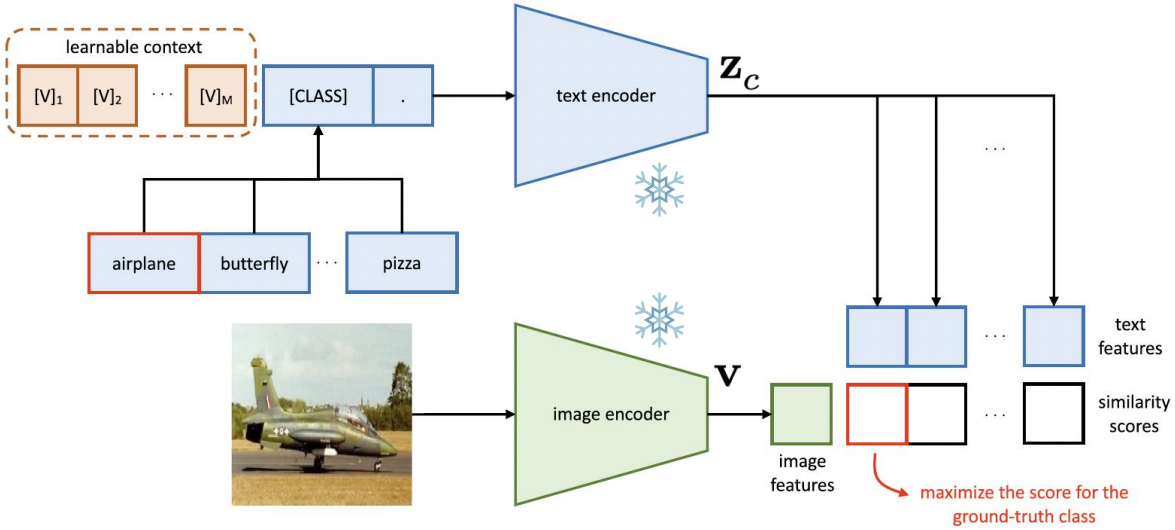
Zhou et *al*. Learning to prompt for vision-language models, IJCV'22

Prompt learning (text)

CoOp

Learning the text prompts

$$\hat{y}_c = \frac{\exp(\mathbf{v} \cdot \mathbf{z}_c^\top / \tau)}{\sum_{k=1}^{K} \exp(\mathbf{v} \cdot \mathbf{z}_k^\top / \tau)}$$



Main idea

$$\boldsymbol{t} = [\mathrm{V}]_1[\mathrm{V}]_2 \ldots [\mathrm{V}]_M[\mathrm{CLASS}],$$

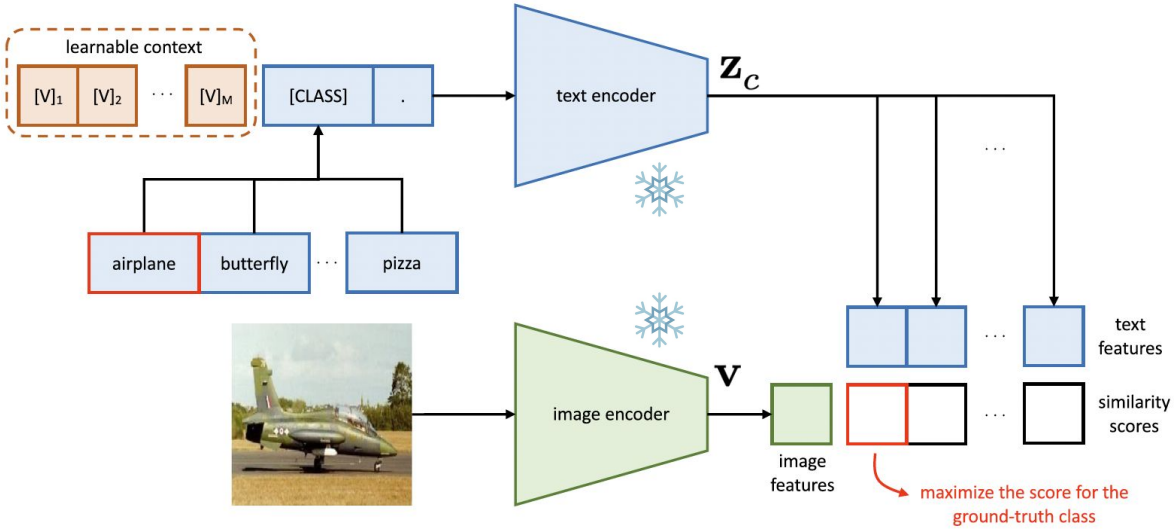
Zhou et *al*. Learning to prompt for vision-language models, IJCV'22

Prompt learning (text)

CoOp

Learning the text prompts

$$\hat{y}_c = \frac{\exp(\mathbf{v} \cdot \mathbf{z}_c^\top / \tau)}{\sum_{k=1}^{K} \exp(\mathbf{v} \cdot \mathbf{z}_k^\top / \tau)}$$



Main idea

$$\boldsymbol{t} = [\mathrm{V}]_1[\mathrm{V}]_2 \ldots [\mathrm{V}]_M[\mathrm{CLASS}],$$

Unified context

Same context across classes

$$\boldsymbol{t} = [\mathrm{V}]_1[\mathrm{V}]_2 \ldots [\mathrm{V}]_M[\mathrm{CLASS}],$$

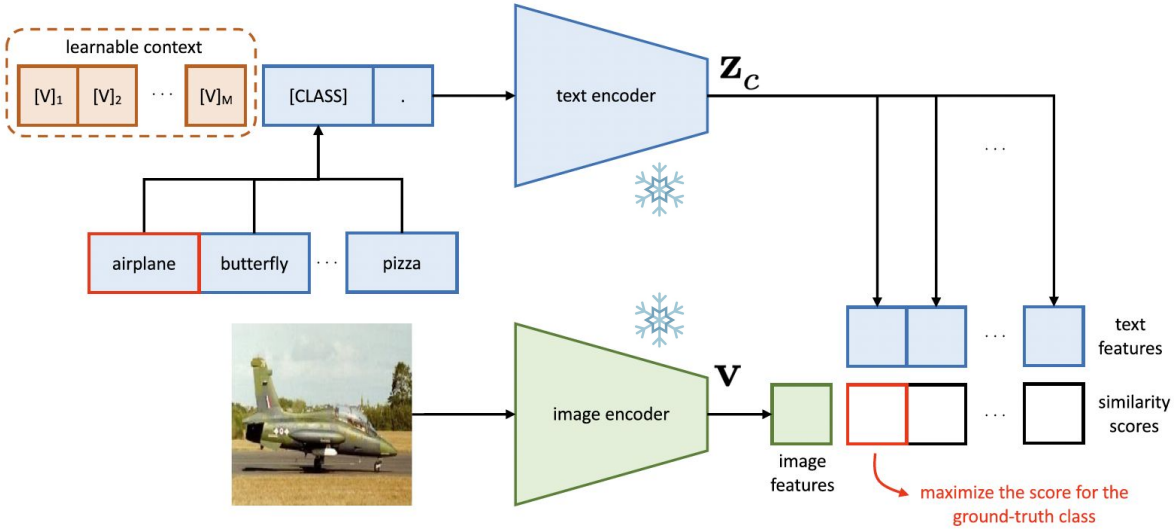
Zhou et *al*. Learning to prompt for vision-language models, IJCV'22

Prompt learning (text)

CoOp

Learning the text prompts

$$\hat{y}_c = \frac{\exp(\mathbf{v} \cdot \mathbf{z}_c^\top / \tau)}{\sum_{k=1}^{K} \exp(\mathbf{v} \cdot \mathbf{z}_k^\top / \tau)}$$



Main idea

$$\boldsymbol{t} = [\mathrm{V}]_1[\mathrm{V}]_2 \ldots [\mathrm{V}]_M[\mathrm{CLASS}],$$

Unified context

Same context across classes

$$\boldsymbol{t} = [\mathrm{V}]_1[\mathrm{V}]_2 \ldots [\mathrm{V}]_M[\mathrm{CLASS}],$$

Class-specific context

Context changes per class

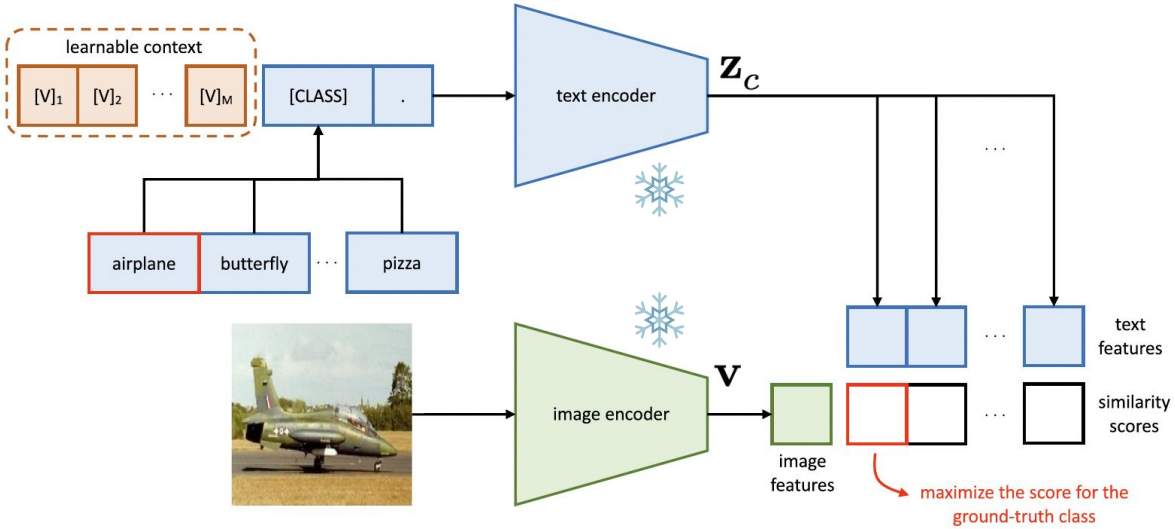
$$[\mathrm{V}]_1^i[\mathrm{V}]_2^i \ldots [\mathrm{V}]_M^i \neq [\mathrm{V}]_1^j[\mathrm{V}]_2^j \ldots [\mathrm{V}]_M^j \quad \text{for } i \neq j \text{ and } i, j \in \{1, \ldots, K\}.$$

Zhou et *al*. Learning to prompt for vision-language models, IJCV'22

Prompt learning (text)

CoOp

Learning the text prompts



$$\hat{y}_c = \frac{\exp(\mathbf{v} \cdot \mathbf{z}_c^\top / \tau)}{\sum_{k=1}^{K} \exp(\mathbf{v} \cdot \mathbf{z}_k^\top / \tau)}$$

Training

$$\mathcal{L} = \sum_{i \in \mathcal{S}} \mathcal{H}(\mathbf{y}_i, \hat{\mathbf{y}}_i)$$

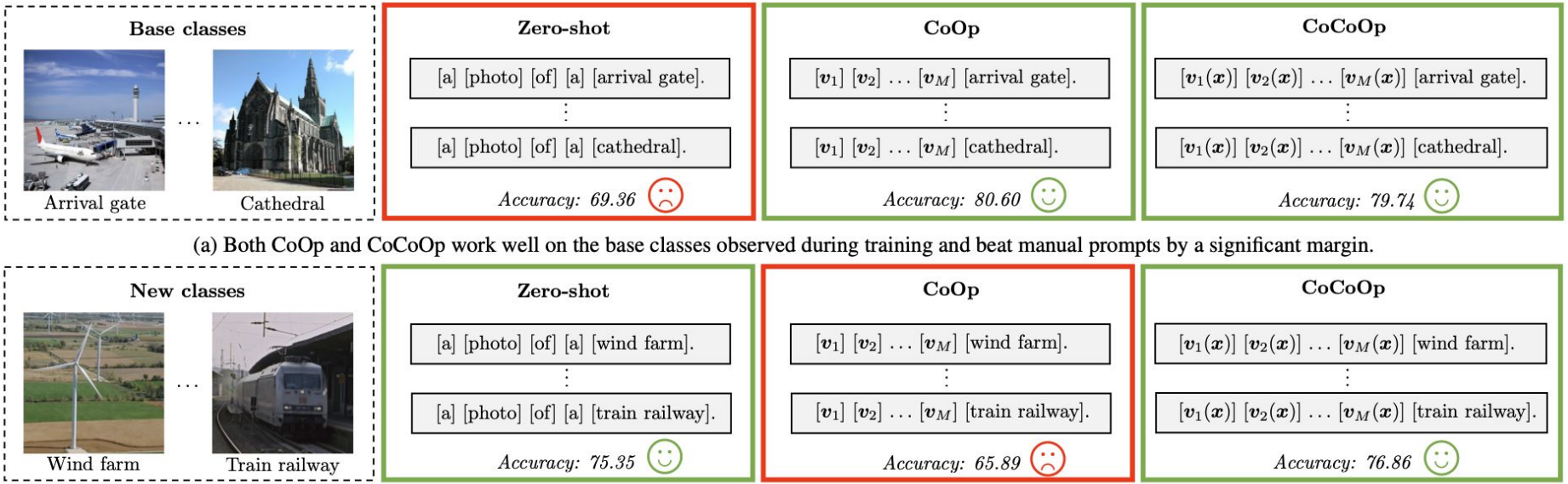
Ground truth for *i*-th element $\mathbf{y}_i = [0, 0, 1, ..., 0]$

Zhou et *al*. Learning to prompt for vision-language models, IJCV'22

Prompt learning (text)

CoCoOp

Learning the text prompts



(a) Both CoOp and CoCoOp work well on the base classes observed during training and beat manual prompts by a significant margin.

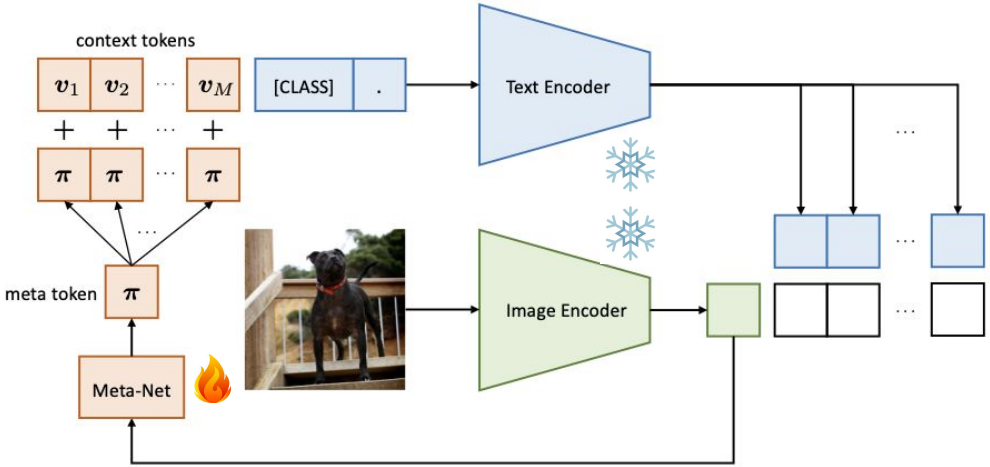
(b) The instance-conditional prompts learned by CoCoOp are much more generalizable than CoOp to the unseen classes.

Zhou et *al*. Conditional prompt learning for vision-language models, CVPR'22

Prompt learning (text)

CoCoOp

Learning the text prompts

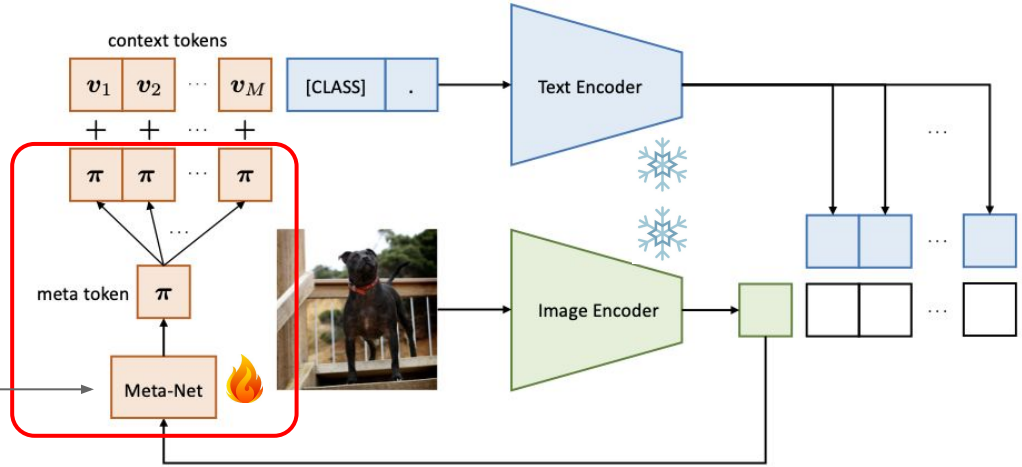


Zhou et *al*. Conditional prompt learning for vision-language models, CVPR'22

Prompt learning (text)

CoCoOp

Learning the text prompts



Addition compared to CoOp

Neural network that generates (for each image) an input-conditional token

$$\boldsymbol{\pi} = h_{\boldsymbol{\theta}}(\boldsymbol{x})$$

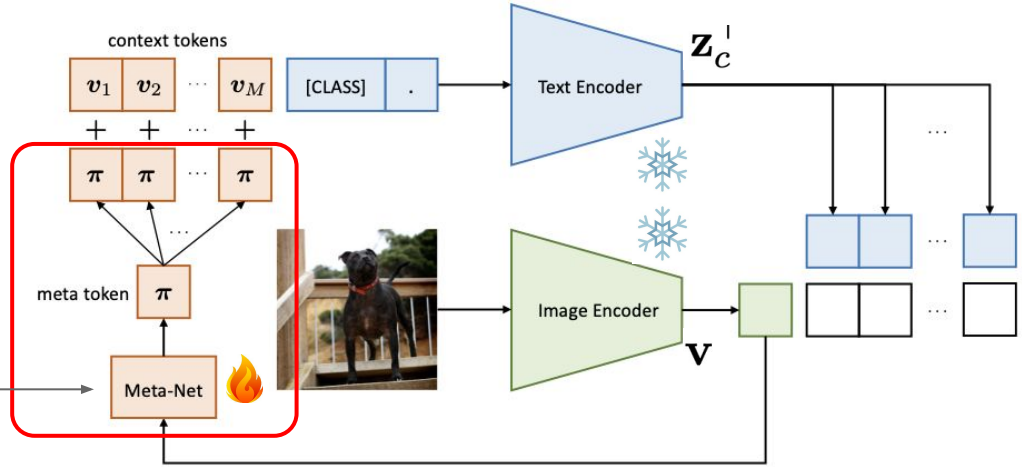
$$\boldsymbol{v}_m(\boldsymbol{x}) = \boldsymbol{v}_m + \boldsymbol{\pi}$$

Zhou et *al*. Conditional prompt learning for vision-language models, CVPR'22

Prompt learning (text)

CoCoOp

Learning the text prompts



Addition compared to CoOp

Neural network that generates (for each image) an input-conditional token

$$\boldsymbol{\pi} = h_{\boldsymbol{\theta}}(\boldsymbol{x})$$

$$\boldsymbol{v}_m(\boldsymbol{x}) = \boldsymbol{v}_m + \boldsymbol{\pi}$$

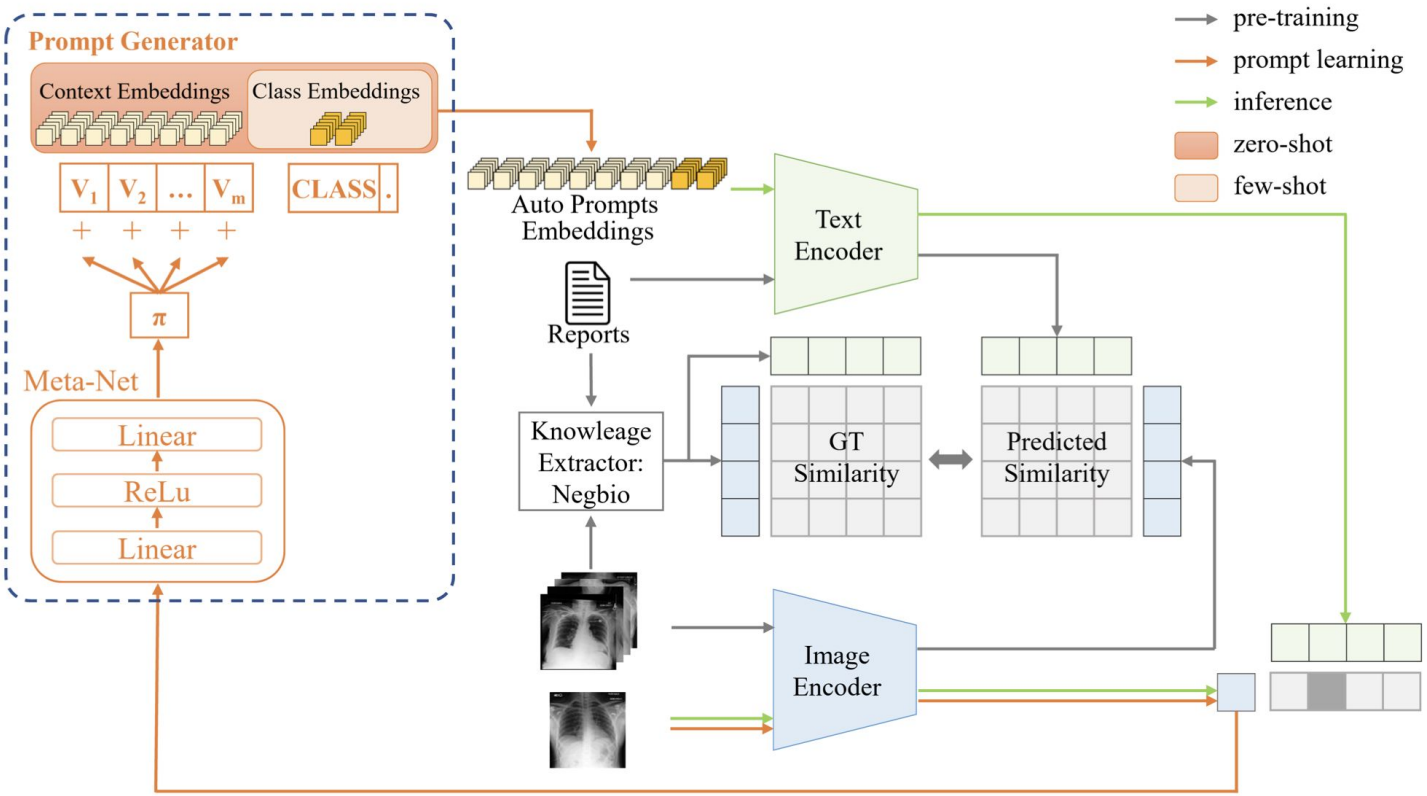
$$\hat{y}_c = \frac{\exp(\mathbf{v} \cdot \mathbf{z}_c^\top / \tau)}{\sum_{k=1}^{K} \exp(\mathbf{v} \cdot \mathbf{z}_k^\top / \tau)}$$

Zhou et *al*. Conditional prompt learning for vision-language models, CVPR'22

Prompt learning (text)

MedPrompt

MedCLIP + CoCoOp

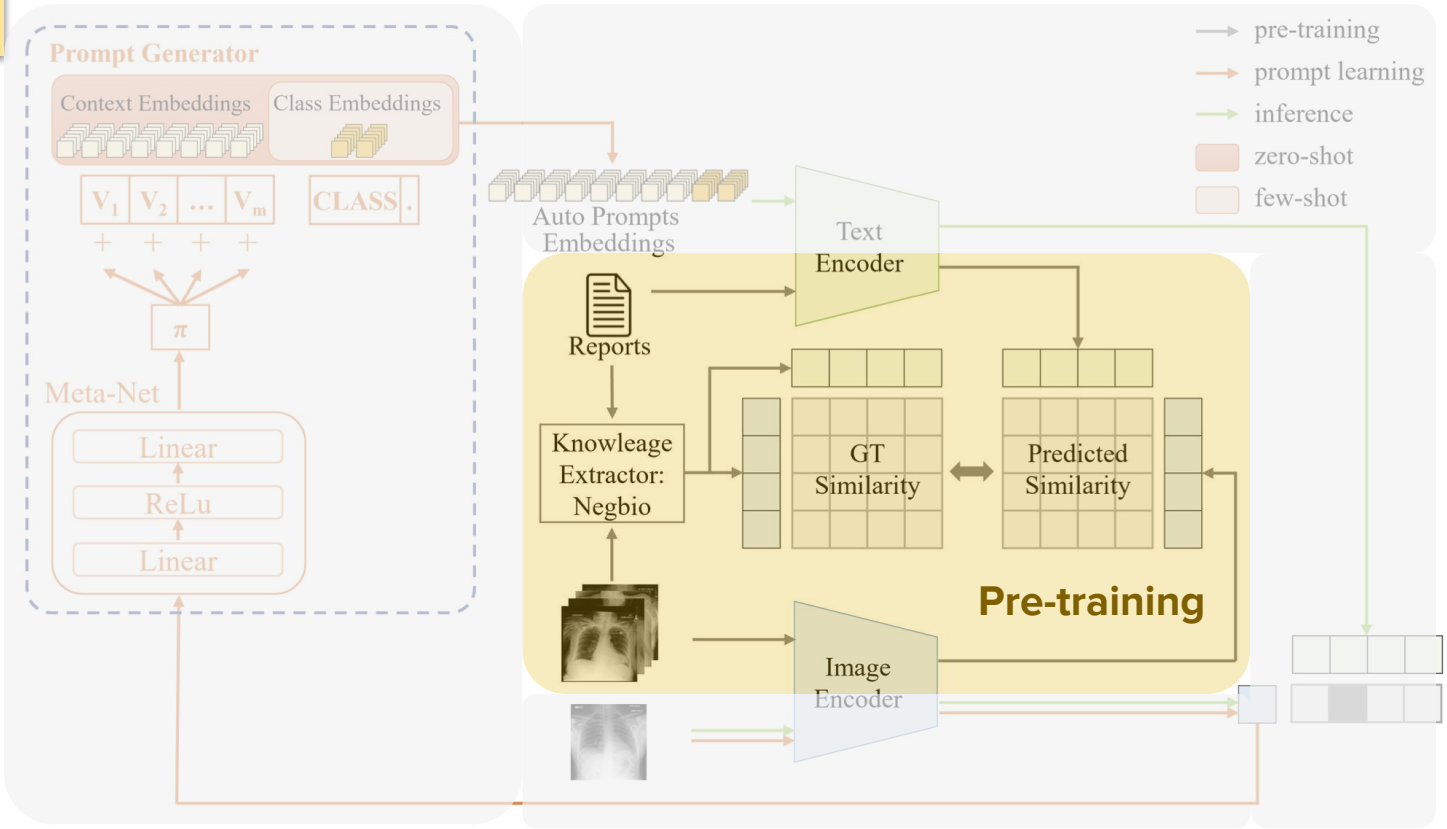


Zheng et al, Exploring low-resource Medical image classification with weakly supervised prompt learning, Pattern Recognition, 2024

Prompt learning (text)

MedPrompt

MedCLIP + CoCoOp

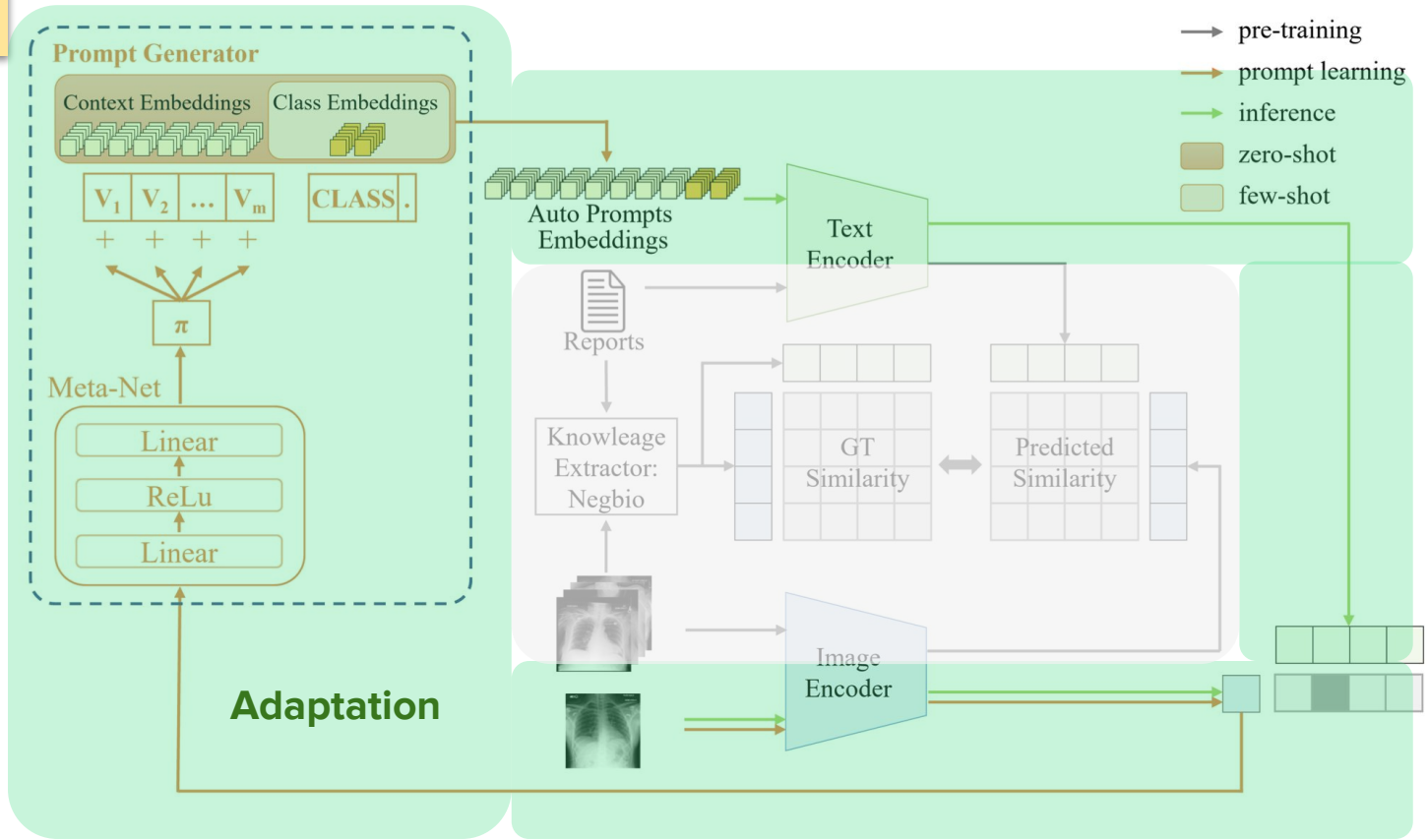


Zheng et al, Exploring low-resource Medical image classification with weakly supervised prompt learning, Pattern Recognition, 2024

Prompt learning (text)

MedPrompt

MedCLIP + CoCoOp



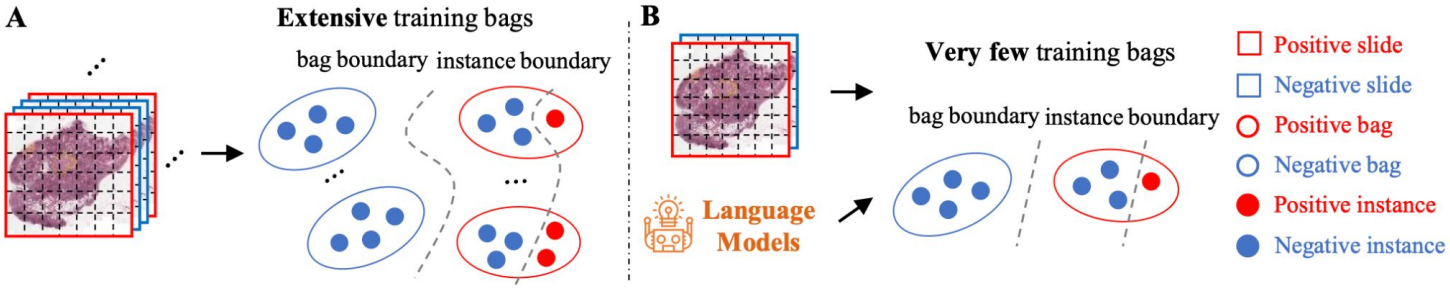
Zheng et al, Exploring low-resource Medical image classification with weakly supervised prompt learning, Pattern Recognition, 2024

Prompt learning (text)

FSWC

MIL based approach

Motivation



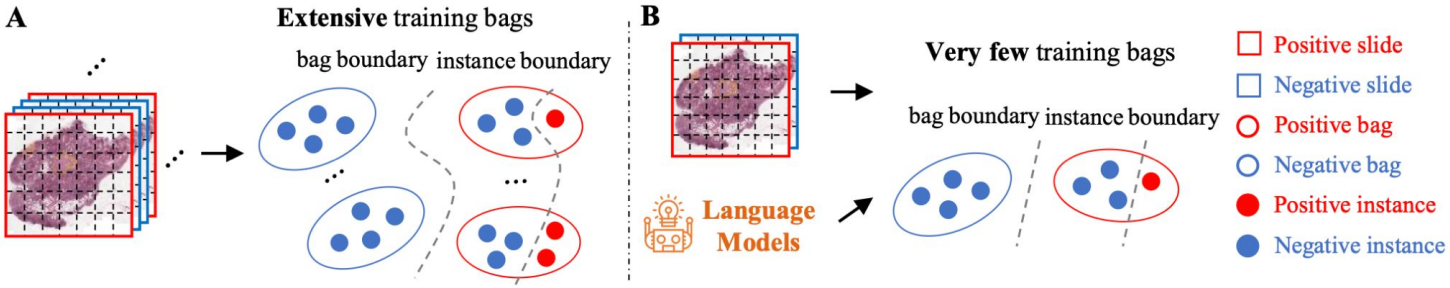
Qu et al, The rise of AI Language pathologists: Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. NeurIPS'23.

Prompt learning (text)

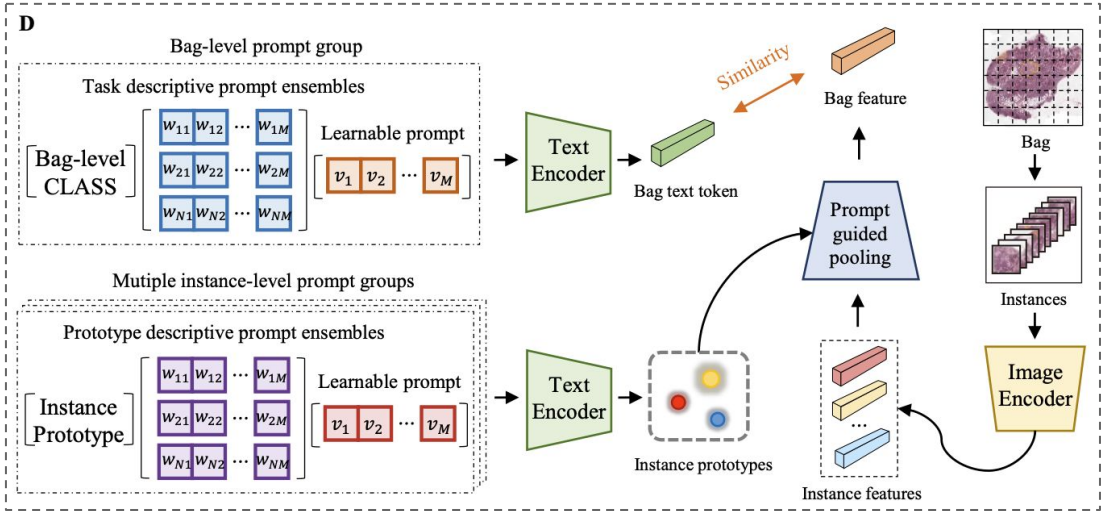
FSWC

MIL based approach

Motivation



Solution



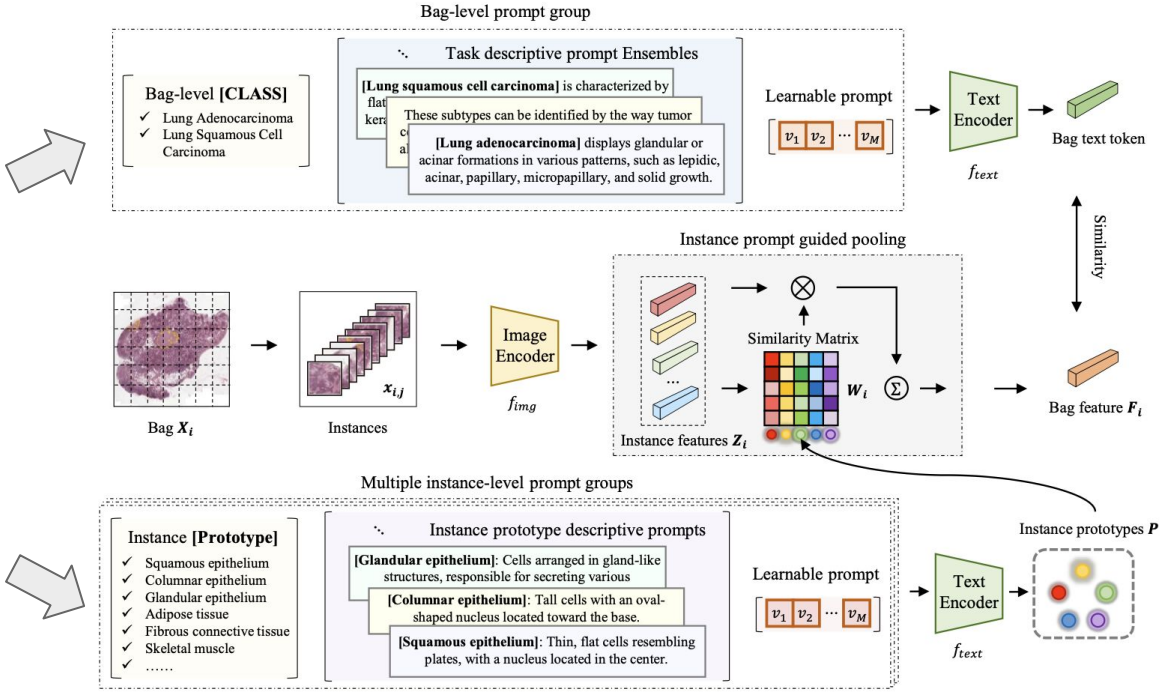
Qu et al, The rise of AI Language pathologists: Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. NeurIPS'23.

Prompt learning (text)

FSWC

MIL based approach

Construction of instance and bag-level prompts (with Chat GPT-4)



Qu et al, The rise of AI Language pathologists: Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. NeurIPS'23.

Prompt learning (text)

FSWC

MIL based approach

Construction of instance and bag-level prompts (with Chat GPT-4)

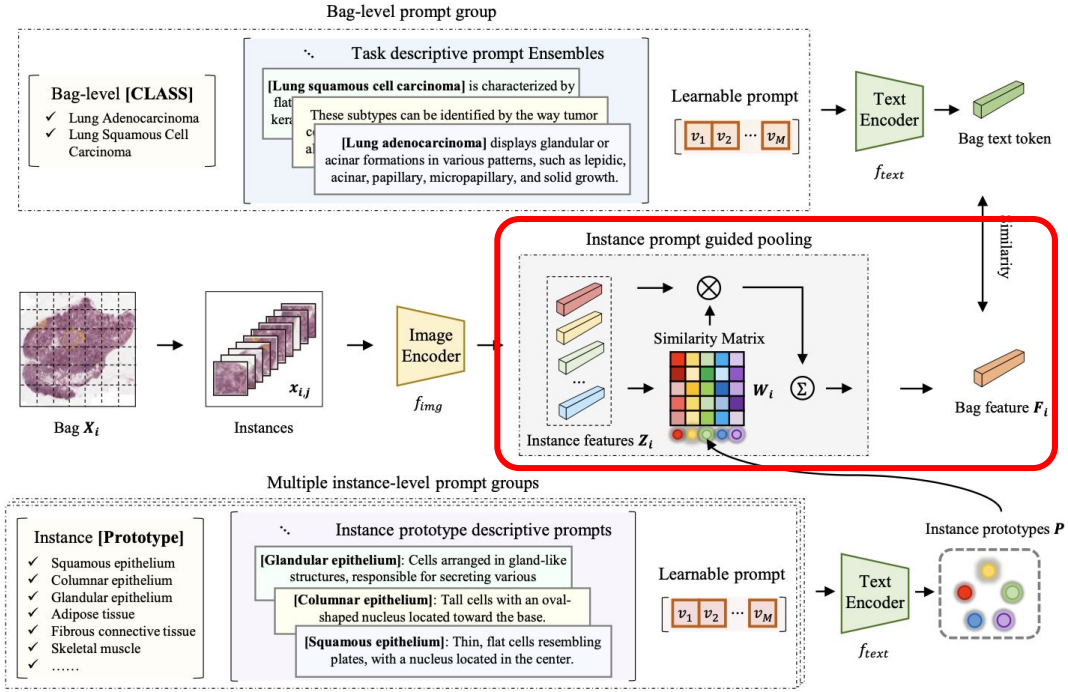
Instance prompt guided pooling

$$\mathbf{V}_i = f_\theta^v(\mathbf{X}_i)$$

$$\mathbf{Z}_i = f_\theta^t(\mathbf{T}_i)$$

$$\mathbf{W}_i = \sigma(\mathbf{V}_i \mathbf{Z}^\top)$$

$$\mathbf{F}_i = \mathrm{MEAN}(\mathbf{W}_i^\top \mathbf{Z}_i)$$



Qu et al, The rise of AI Language pathologists: Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. NeurIPS'23.

Prompt learning (text)

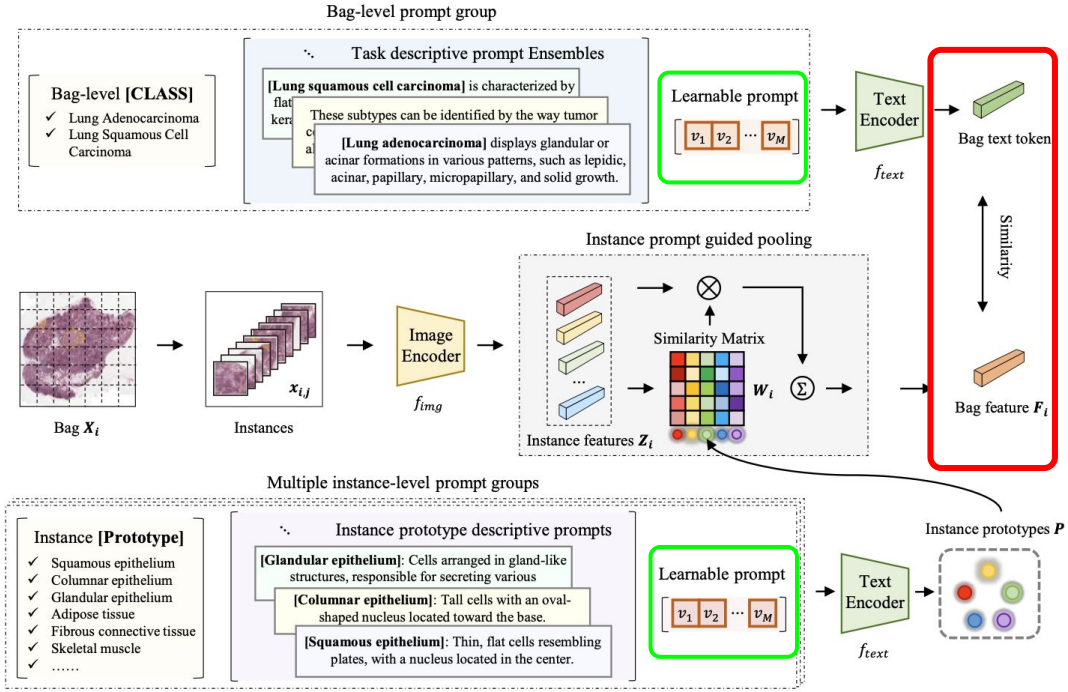
FSWC

MIL based approach

Construction of instance and bag-level prompts (with Chat GPT-4)

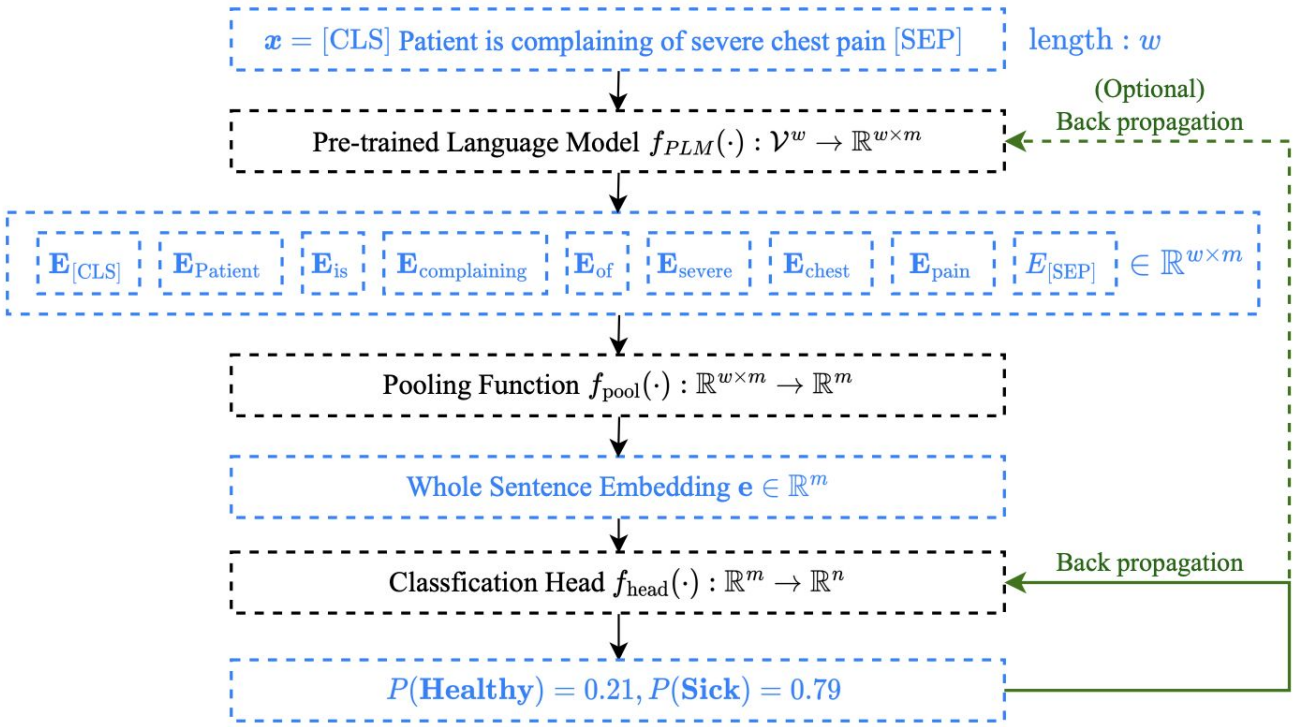
Instance prompt guided pooling

Bag-level text and image similarity
(Only the prompts are updated)



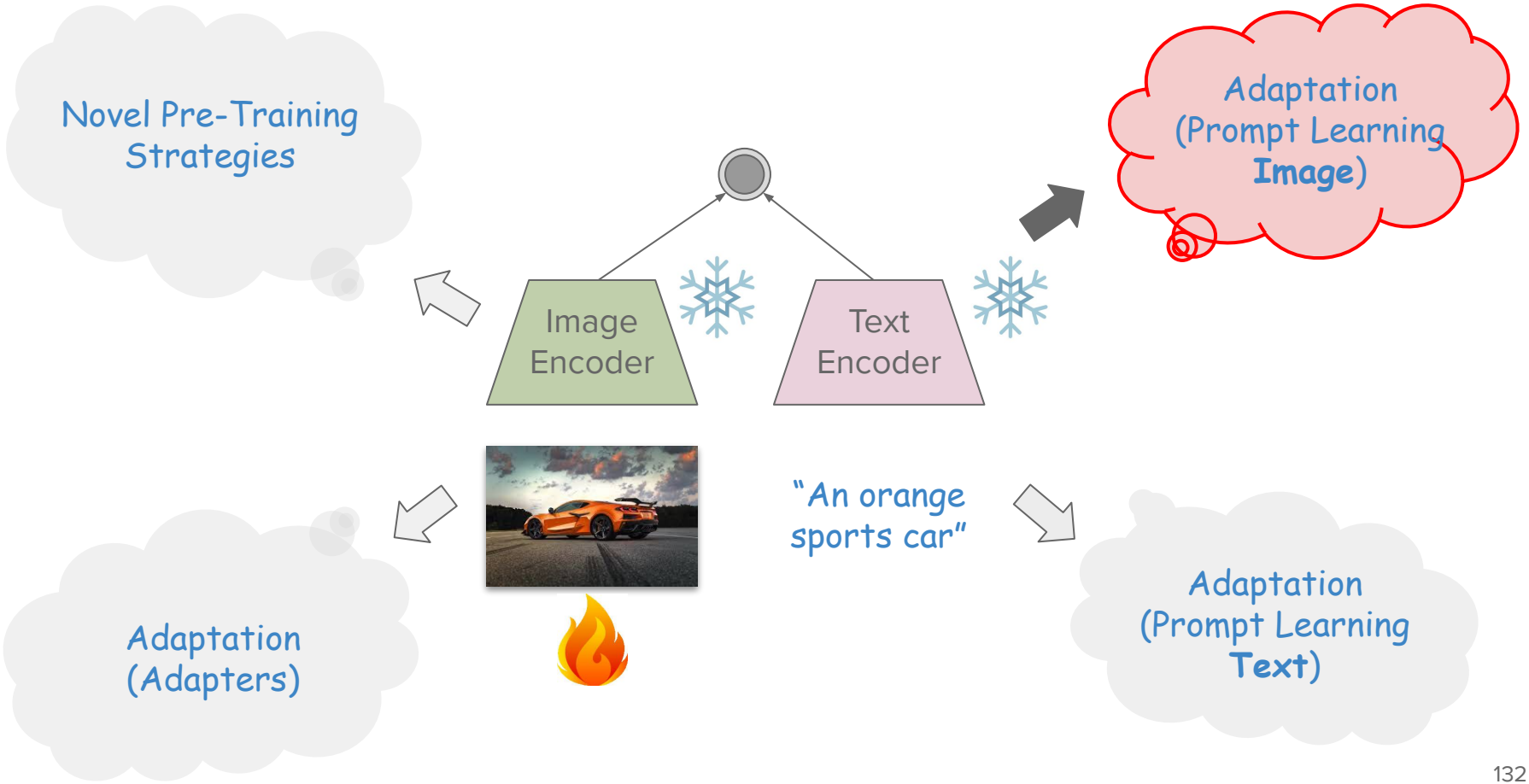
Qu et al, The rise of AI Language pathologists: Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. NeurIPS'23.

Prompt learning (text)



Zheng et al, Exploring low-resource Medical image classification with weakly supervised prompt learning, Pattern Recognition, 2024

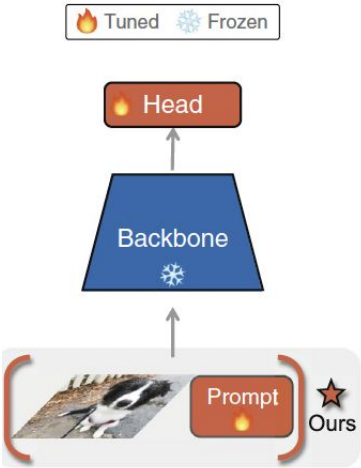
Where we go from CLIP?



Prompt learning (image)

VPT

Learning the text **(visual)** prompts

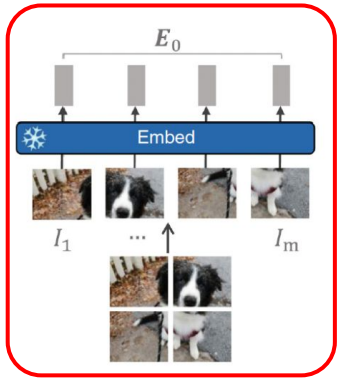


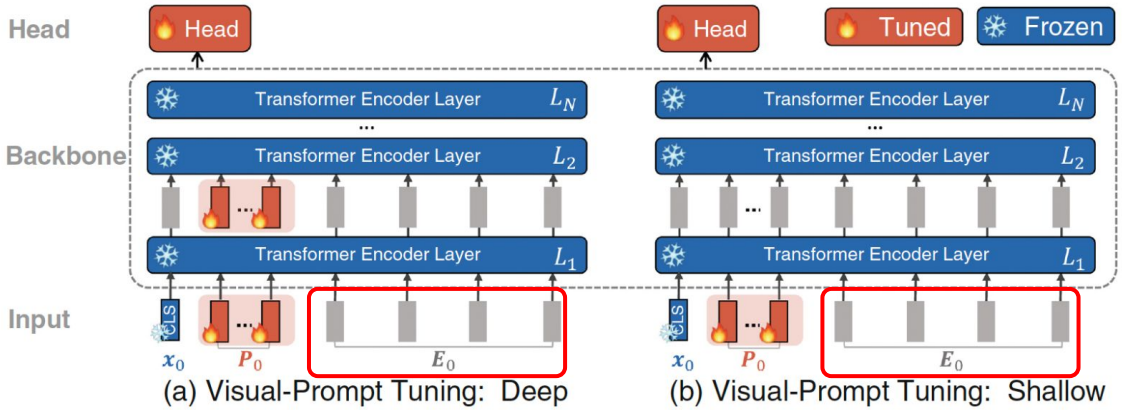
Jia et *al*. Visual Prompt Tuning, ECCV'22

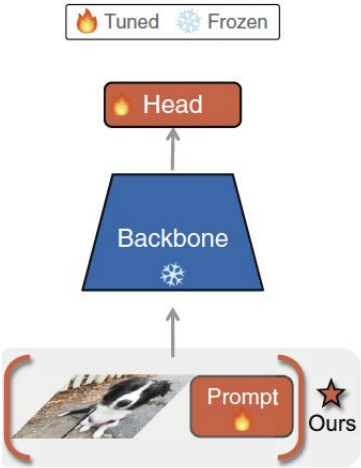
Prompt learning (image)

VPT

Learning the text **(visual)** prompts







- Divide an image into *m* patches (and fed them to the ViT)

$$\mathbf{e}_0^j = \mathtt{Embed}(I_j) \qquad \mathbf{e}_0^j \in \mathbb{R}^d, j = 1, 2, \ldots m$$

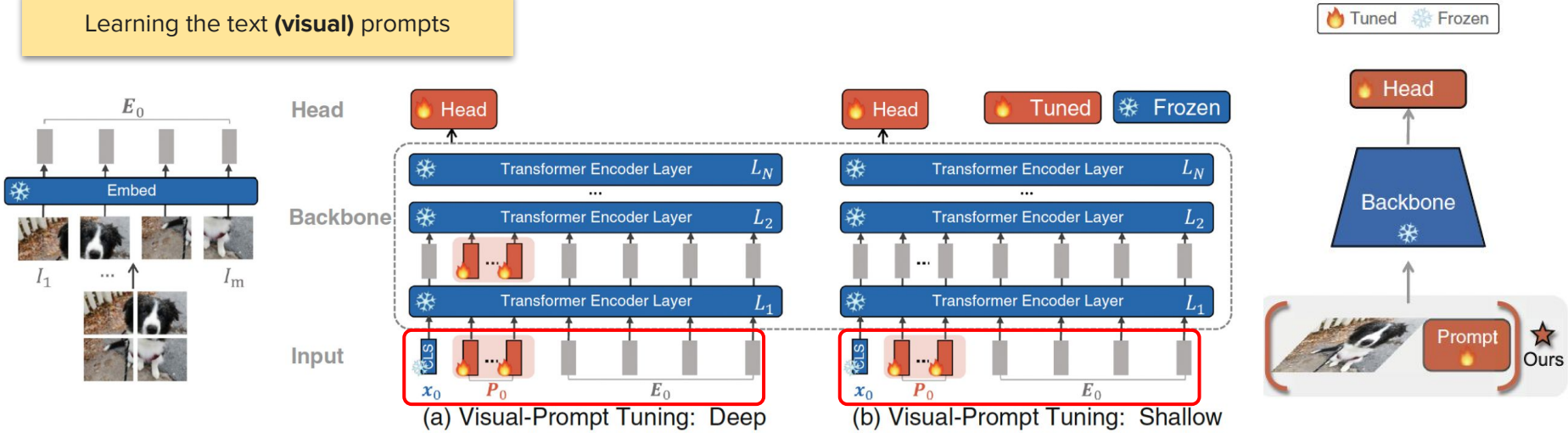
$$\mathbf{E}_i = \{\mathbf{e}_i^j \in \mathbb{R}^d | j \in \mathbb{N}, 1 \leq j \leq m\}$$ In the i-th layer

Jia et *al*. Visual Prompt Tuning, ECCV'22

Prompt learning (image)

VPT

Learning the text **(visual)** prompts



- Divide an image into *m* patches (and fed them to the ViT)

$$\mathbf{e}_0^j = \mathtt{Embed}(I_j) \qquad \mathbf{e}_0^j \in \mathbb{R}^d, j = 1, 2, \ldots m$$

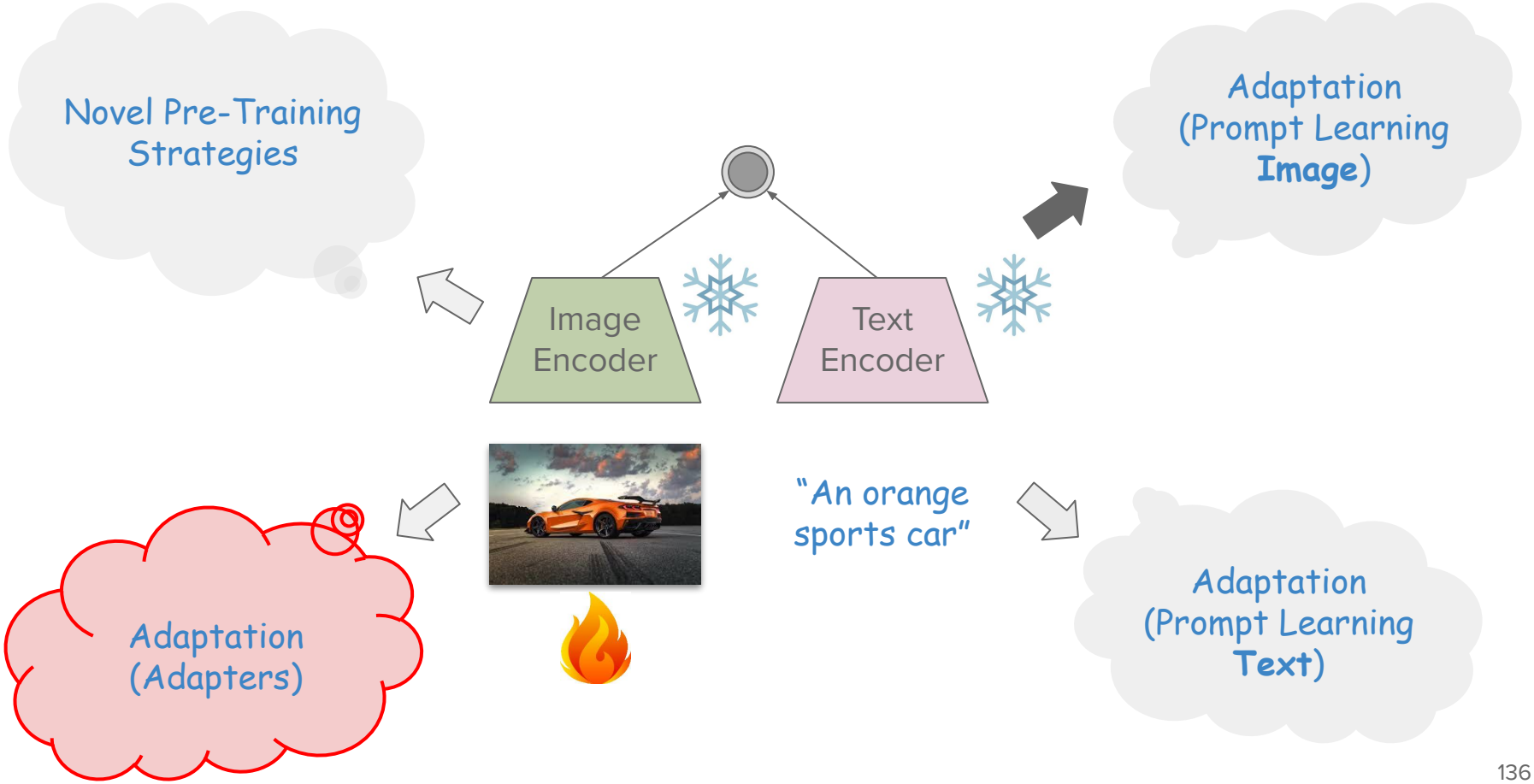
$$\mathbf{E}_i = \{\mathbf{e}_i^j \in \mathbb{R}^d | j \in \mathbb{N}, 1 \leq j \leq m\}$$

In the i-th layer

$$[\mathbf{x}_1, \mathbf{Z}_1, \mathbf{E}_1] = L_1([\mathbf{x}_0, \mathbf{P}, \mathbf{E}_0])$$
$$[\mathbf{x}_i, \mathbf{Z}_i, \mathbf{E}_i] = L_i([\mathbf{x}_{i-1}, \mathbf{Z}_{i-1}, \mathbf{E}_{i-1}])$$
$$\mathbf{y} = \mathtt{Head}(\mathbf{x}_N) \ ,$$

Jia et *al*. Visual Prompt Tuning, ECCV'22

Where we go from CLIP?



Adapters (parameter efficient)

Revisiting the predictions

Softmax probability of class *k* for a given image *i*

$$\hat{y}_{i,k} = \frac{\exp\left(\mathbf{v}_i \cdot \mathbf{z}_k^\top / \tau\right)}{\sum_{j=1}^{K} \exp\left(\mathbf{v}_i \cdot \mathbf{z}_j^\top / \tau\right)}$$

$$\mathbf{l}_i = \{l_{i1}, l_{i2}, ..., l_{ik}\}$$

Logits

Learning objective (for all methods)

$$\mathcal{L} = \sum_{i \in \mathcal{S}} \mathcal{H}(\mathbf{y}_i, \hat{\mathbf{y}}_i)$$

Adapters (parameter efficient)

Revisiting the predictions

Softmax probability of class *k* for a given image *i*

$$\hat{y}_{i,k} = \frac{\exp\left(\mathbf{v}_i \cdot \mathbf{z}_k^\top / \tau\right)}{\sum_{j=1}^{K} \exp\left(\mathbf{v}_i \cdot \mathbf{z}_j^\top / \tau\right)}$$

$$\mathbf{l}_i = \{l_{i1}, l_{i2}, \ldots, l_{ik}\}$$

Logits

Learning objective (for all methods)

$$\mathcal{L} = \sum_{i \in \mathcal{S}} \mathcal{H}(\mathbf{y}_i, \hat{\mathbf{y}}_i)$$

Objective of adapters

$$l_{ik} = \boldsymbol{\theta}_a(\mathbf{v}_i, \mathbf{z}_k, \tau)$$

Set of learnable parameters

Adapters (parameter efficient)

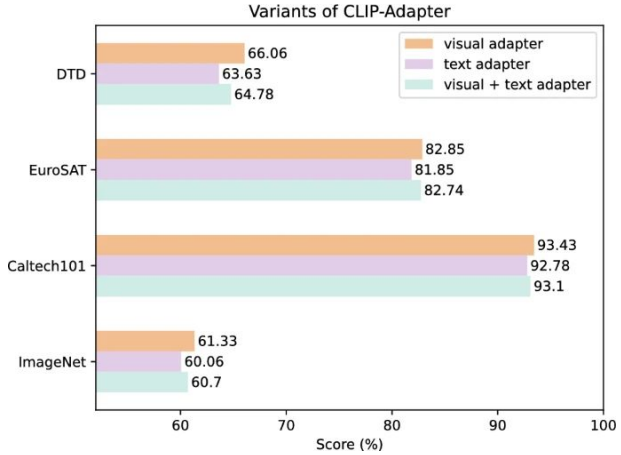
CLIP-Adapter

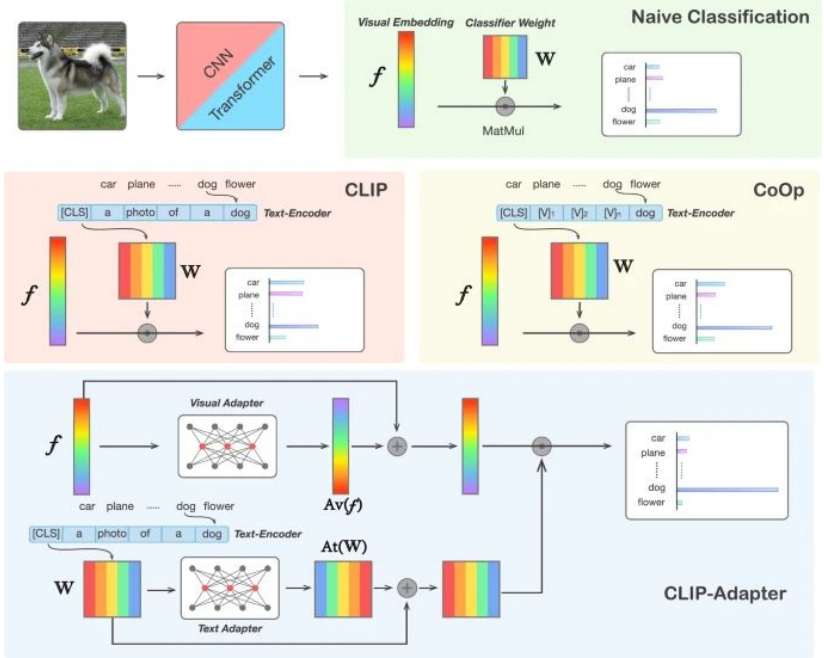
$$l_{ik} = \boldsymbol{\theta}_a(\mathbf{v}_i, \mathbf{z}_k, \tau)$$

How visual features are modified:

$$\mathbf{v}' = \mathbf{v} + \alpha_{\mathrm{r}} f_{\psi}(\mathbf{v})$$

* Even though this residual is presented for the text features also, best results are obtained with only visual adaptation





Gao et *al*. Clip-adapter: Better vision-language models with feature adapters, IJCV'24

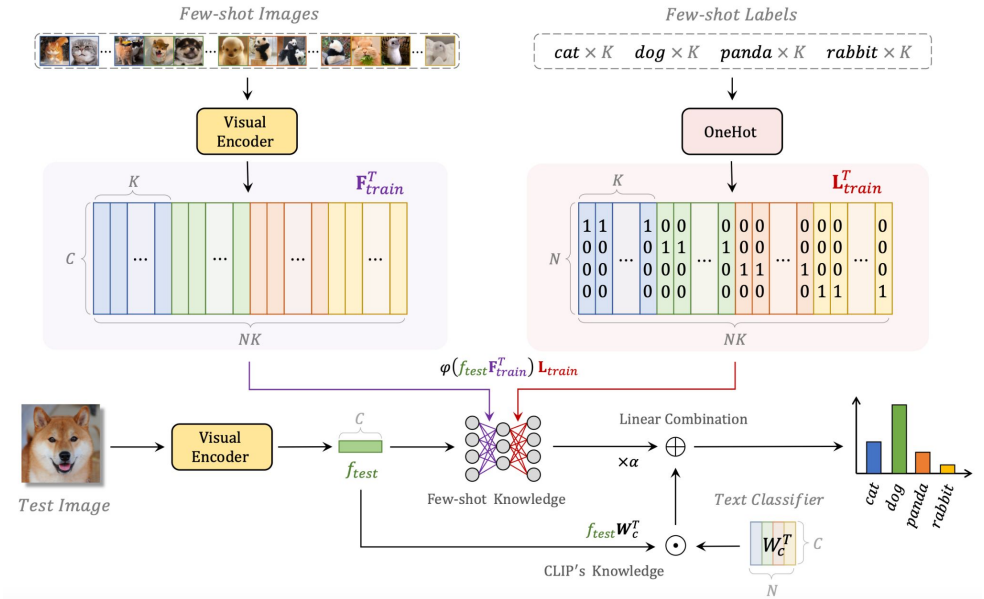
Adapters (parameter efficient)

TIP-Adapter

$$l_{ik} = \boldsymbol{\theta}_a(\mathbf{v}_i, \mathbf{z}_k, \tau)$$

How visual features are modified:

$$\mathbf{l}_c = \alpha_{\text{tipA}} f_\psi(\mathbf{v}, \beta) + \mathbf{v} \cdot \mathbf{t}_c^\top / \tau$$



Zhang et *al*. Tip-Adapter: Training-free clip-adapter for better vision-language models, ECCV'22

Adapters (parameter efficient)

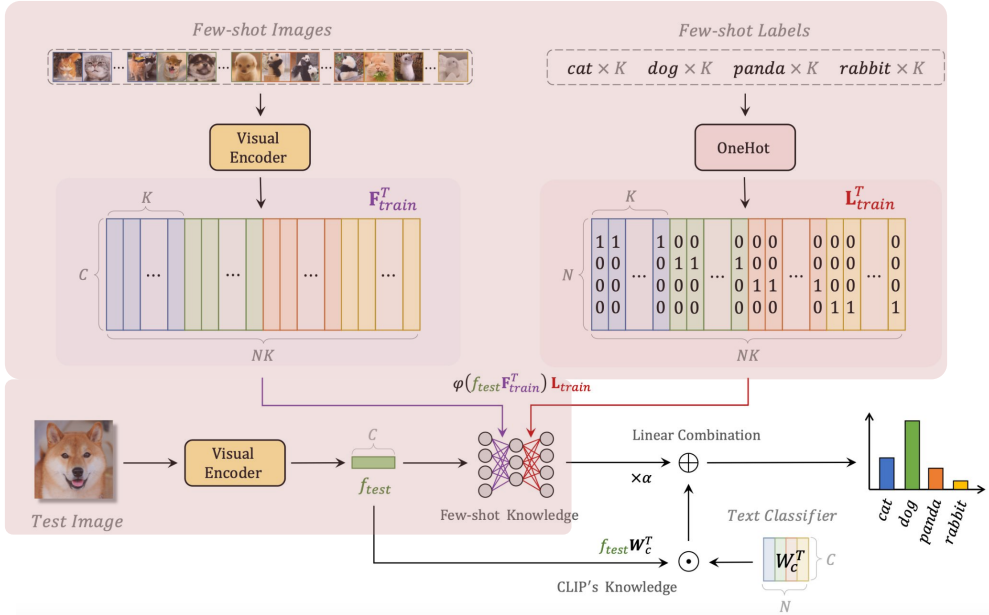
TIP-Adapter

$$l_{ik} = \boldsymbol{\theta}_a(\mathbf{v}_i, \mathbf{z}_k, \tau)$$

How visual features are modified:

$$\mathbf{l}_c = \alpha_{\text{tipA}} \boxed{f_\psi(\mathbf{v}, \beta)} + \mathbf{v} \cdot \mathbf{t}_c^\top / \tau$$

Similarity to support samples



Zhang et *al*. Tip-Adapter: Training-free clip-adapter for better vision-language models, ECCV'22

Adapters (parameter efficient)

TIP-Adapter

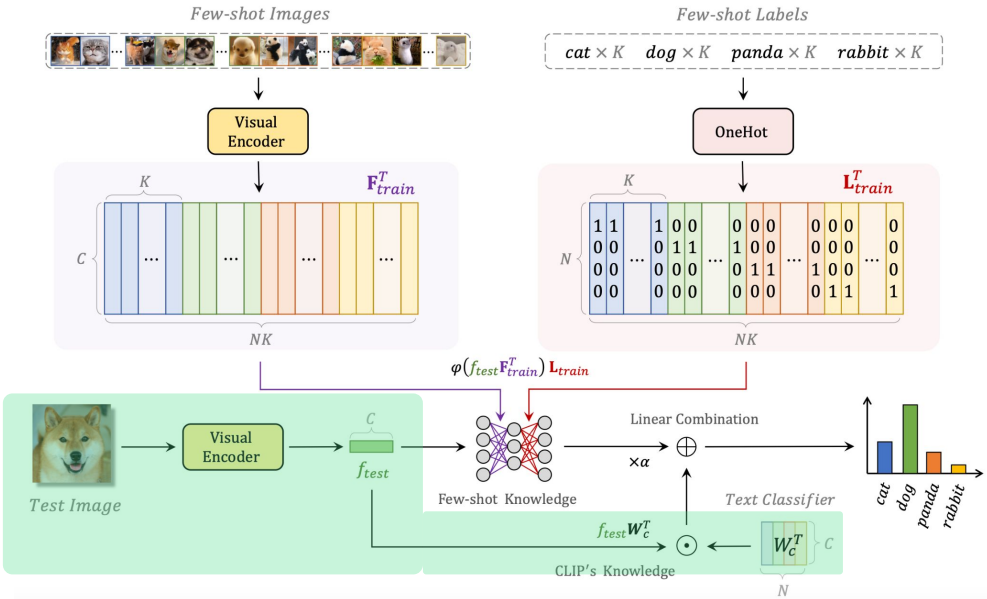
$$l_{ik} = \boldsymbol{\theta}_a(\mathbf{v}_i, \mathbf{z}_k, \tau)$$

How visual features are modified:

$$\mathbf{l}_c = \alpha_{\text{tipA}} f_\psi(\mathbf{v}, \beta) + \boxed{\mathbf{v} \cdot \mathbf{t}_c^\top / \tau}$$

Similarity to support samples

Similarity to the class text prototypes (i.e., zero-shot prototypes)



Zhang et *al*. Tip-Adapter: Training-free clip-adapter for better vision-language models, ECCV'22

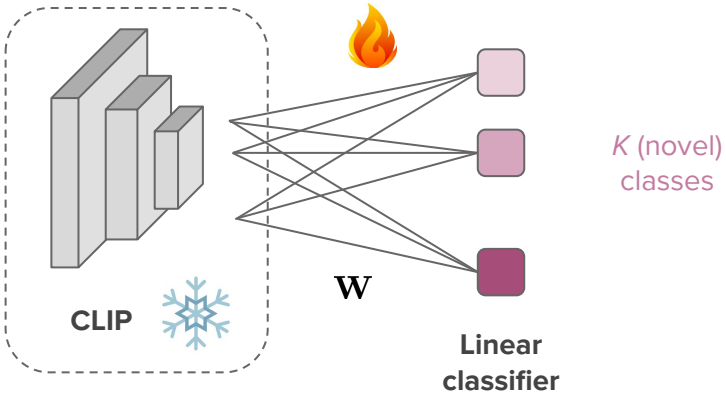
Adapters (parameter efficient)

Limitations of existing methods

$$l_{ik} = \boldsymbol{\theta}_a(\mathbf{v}_i, \mathbf{z}_k, \tau)$$

1- They do not show strong generability to other tasks

Preliminaries: What is Linear Probing (LP)?



Optimize parameters only with the support set

$$\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{S}$$

Silva-Rodriguez et *al*. A closer look at the few shot adaptation of large vision-language models. CVPR'24

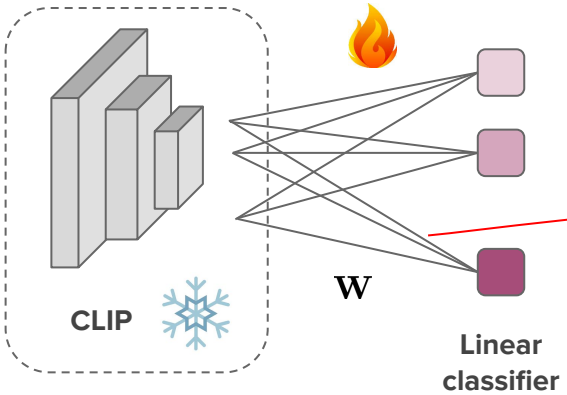
Adapters (parameter efficient)

Limitations of existing methods

$$l_{ik} = \boldsymbol{\theta}_a(\mathbf{v}_i, \mathbf{z}_k, \tau)$$

1- They do not show strong generability to other tasks

Preliminaries: What is Linear Probing (LP)?



K (novel) classes

These weights are equivalent to class prototypes

Optimize parameters only with the support set

$$\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{S}$$

$$\hat{y}_c = \frac{\exp(\mathbf{v} \cdot \mathbf{w}_c^\top / \tau)}{\sum_{i=1}^{C} \exp(\mathbf{v} \cdot \mathbf{w}_i^\top / \tau)}$$

Silva-Rodriguez et *al*. A closer look at the few shot adaptation of large vision-language models. CVPR'24

Adapters (parameter efficient)

Limitations of existing methods

$$l_{ik} = \boldsymbol{\theta}_a(\mathbf{v}_i, \mathbf{z}_k, \tau)$$

1- They do not show strong generability to other tasks

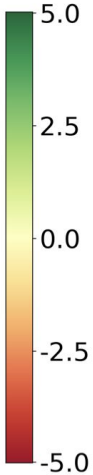
Relative improvement over Linear Probing when the hyperparameters are tuned on OPe dataset

CLIP-Adapter

	ImN	Cal	OPe	SCa	Flw	Foo	FGV	SUN	DTD	EuS	UCF
ImN	4.1	-2.1	3.1	-3.6	-21	1.7	-6.6	-0.8	-9.9	-3	-6.3
Cal	1.9	0.4	3.6	-1.4	-11	1.4	-4.1	0.5	-1.4	-0.9	-2.4
OPe	1.9	0.4	3.6	-1.4	-11	1.4	-4.1	0.5	-1.4	-0.9	-2.4
SCa	1.9	0.4	3.6	-1.4	-11	1.4	-4.1	0.5	-1.4	-0.9	-2.4
Flw	-4.7	-1.6	-2.2	-6	-1.3	-5.3	-1.6	-3.8	-1.7	0.9	-2
Foo	4.1	-2.1	3.1	-3.6	-21	1.7	-6.6	-0.8	-9.9	-3	-6.3
FGV	-4.7	-1.6	-2.2	-6	-1.3	-5.3	-1.6	-3.8	-1.7	0.9	-2
SUN	1.9	0.4	3.6	-1.4	-11	1.4	-4.1	0.5	-1.4	-0.9	-2.4
DTD	-3.2	-1.1	-1.4	-4.6	-2.2	-3.5	-2	-2.8	0.1	0.7	-1
EuS	-4.7	-1.6	-2.2	-6	-1.3	-5.3	-1.6	-3.8	-1.7	0.9	-1
UCF	-0.4	-0.1	1.9	-1.8	-20	-1.1	-2.3	-0.4	-0	0.4	0.4

TIP-Adapter

	ImN	Cal	OPe	SCa	Flw	Foo	FGV	SUN	DTD	EuS	UCF
ImN	3.7	-2.4	3.1	-1.3	-17	0.8	-5.1	-1.5	-6.9	-6	-6.3
Cal	-3.8	-2.3	-3.9	-7.2	-2.6	-9.1	-2.6	-3.1	-3.2	-10	-7
OPe	3.7	-2.4	3.1	-1.3	-17	0.8	-5.1	-1.5	-6.9	-6	-6.3
SCa	3.7	-2.4	3.1	-1.3	-17	0.8	-5.1	-1.5	-6.9	-6	-6.3
Flw	-3.8	-2.3	-3.9	-7.2	-2.6	-9.1	-2.6	-3.1	-3.2	-10	-7
Foo	3.7	-2.4	3.1	-1.3	-17	0.8	-5.1	-1.5	-6.9	-6	-6.3
FGV	-3.8	-2.3	-3.9	-7.2	-2.6	-9.1	-2.6	-3.1	-3.2	-10	-7
SUN	3.7	-2.4	3.1	-1.3	-17	0.8	-5.1	-1.5	-6.9	-6	-6.3
DTD	-3.8	-2.3	-3.9	-7.2	-2.6	-9.1	-2.6	-3.1	-3.2	-10	-7
EuS	3.7	-2.4	3.1	-1.3	-17	0.8	-5.1	-1.5	-6.9	-6	-6.3
UCF	3.7	-2.4	3.1	-1.3	-17	0.8	-5.1	-1.5	-6.9	-6	-6.3



Silva-Rodriguez et *al*. A closer look at the few shot adaptation of large vision-language models. CVPR'24

Adapters (parameter efficient)

Limitations of existing methods

$$l_{ik} = \boldsymbol{\theta}_a(\mathbf{v}_i, \mathbf{z}_k, \tau)$$

1- They do not show strong generability to other tasks

Relative improvement over Linear Probing when the hyperparameters are tuned on OPe dataset

CLIP-Adapter

	ImN	Cal	OPe	SCa	Flw	Foo	FGV	SUN	DTD	EuS	UCF
ImN	4.1	-2.1	3.1	-3.6	-21	1.7	-6.6	-0.8	-9.9	-3	-6.3
Cal	1.9	0.4	3.6	-1.4	-11	1.4	-4.1	0.5	-1.4	-0.9	-2.4
OPe	1.9	0.4	3.6	-1.4	-11	1.4	-4.1	0.5	-1.4	0.9	2.4
SCa	1.9	0.4	3.6	-1.4	-11	1.4	-4.1	0.5	-1.4		
Flw	-4.7	-1.6	-2.2	-6	-1.3	-5.3	-1.6	-3.8	-1.7		
Foo	4.1	-2.1	3.1	-3.6	-21	1.7	-6.6	-0.8	-9.9		
FGV	-4.7	-1.6	-2.2	-6	-1.3	-5.3	-1.6	-3.8	-1.7		
SUN	1.9	0.4	3.6	-1.4	-11	1.4	-4.1	0.5	-1.4		
DTD	-3.2	-1.1	-1.4	-4.6	-2.2	-3.5	-2	-2.8	0.1	0.7	-1
EuS	-4.7	-1.6	-2.2	-6	-1.3	-5.3	-1.6	-3.8	-1.7	0.9	-1
UCF	-0.4	-0.1	1.9	-1.8	-20	-1.1	-2.3	-0.4	-0	0.4	0.4

TIP-Adapter

	ImN	Cal	OPe	SCa	Flw	Foo	FGV	SUN	DTD	EuS	UCF
ImN	3.7	-2.4	3.1	-1.3	-17	0.8	-5.1	-1.5	-6.9	-6	-6.3
Cal	-3.8	-2.3	-3.9	-7.2	-2.6	-9.1	-2.6	-3.1	-3.2	-10	-7
OPe	3.7	-2.4	3.1	-1.3	-17	0.8	-5.1	-1.5	-6.9	-6	-6.3
SCa											-6.3
Flw											-7
Foo											-6.3
FGV											-7
SUN											-6.3
DTD	-3.8	-2.3	-3.9	-7.2	-2.6	-9.1	-2.6	-3.1	-3.2	-10	-7
EuS	3.7	-2.4	3.1	-1.3	-17	0.8	-5.1	-1.5	-6.9	-6	-6.3
UCF	3.7	-2.4	3.1	-1.3	-17	0.8	-5.1	-1.5	-6.9	-6	-6.3

5.0 … -2.5 … -5.0

They require an additional validation set to tune their hyperparameters

Silva-Rodriguez et *al*. A closer look at the few shot adaptation of large vision-language models. CVPR'24

Adapters (parameter efficient)

CLAP

$$l_{ik} = \boldsymbol{\theta}_a(\mathbf{v}_i, \mathbf{z}_k, \tau)$$

$$\min_{\mathcal{W}} \quad \frac{1}{\|\mathcal{S}\|} \sum_{i \in \mathcal{S}} \mathcal{H}(\mathbf{y}_i, \hat{\mathbf{y}}_i)$$

Cross-entropy on the support samples

$$\text{s.t.} \quad \mathbf{w}_k = \mathbf{t}_k \quad \forall k \in \{1, \ldots, K\},$$

Constraint (i.e., force similarity to text prototypes)

Silva-Rodriguez et *al*. A closer look at the few shot adaptation of large vision-language models. CVPR'24

Adapters (parameter efficient)

CLAP

$$l_{ik} = \boldsymbol{\theta}_a(\mathbf{v}_i, \mathbf{z}_k, \tau)$$

$$\min_{\mathcal{W}} \quad \frac{1}{\|\mathcal{S}\|} \sum_{i \in \mathcal{S}} \mathcal{H}(\mathbf{y}_i, \hat{\mathbf{y}}_i)$$

Cross-entropy on the support samples

$$\text{s.t.} \quad \mathbf{w}_k = \mathbf{t}_k \quad \forall k \in \{1, \ldots, K\},$$

Constraint (i.e., force similarity to text prototypes)

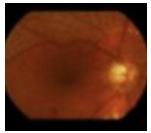
$$\min_{\mathcal{W}} \quad \sum_{i \in \mathcal{S}} \mathcal{H}(\mathbf{y}_i, \hat{\mathbf{y}}_i) + \lambda_k \sum_{k=1}^{K} ||\mathbf{t}_k - \mathbf{w}_k^{(m)}||_2^2,$$

Initialization: Average of the zero-shot softmax for all support samples for class k

Silva-Rodriguez et *al*. A closer look at the few shot adaptation of large vision-language models. CVPR'24

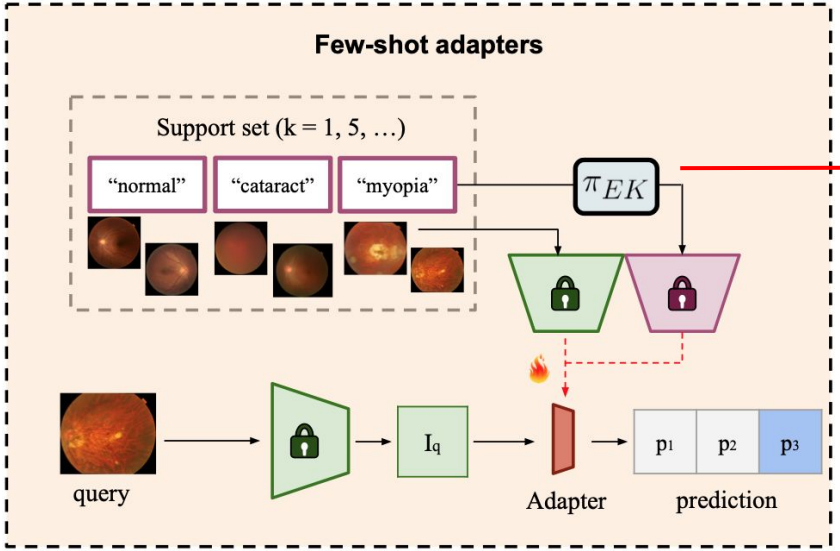
Adapters (parameter efficient)

FLAIR



Incorporating expert-domain knowledge

$$l_{ik} = \boldsymbol{\theta}_a(\mathbf{v}_i, \mathbf{z}_k, \tau)$$



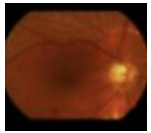
"A fundus image of possible signs of hemorrhage with blot, dot, or flame-shaped"

Expert Knowledge

Silva-Rodriguez et *al*. A Foundation LAnguage-Image model of the Retina (FLAIR): Encoding expert knowledge in text supervision. MedIA'24 (Under review)

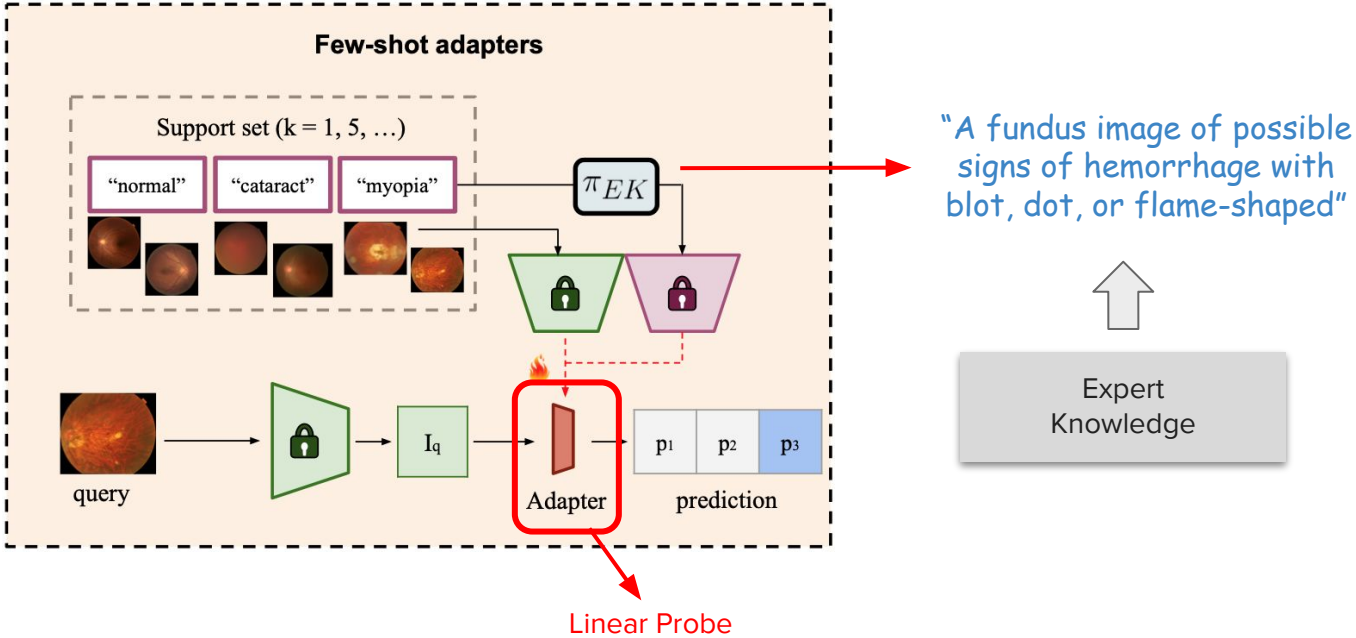
Adapters (parameter efficient)

FLAIR



Incorporating expert-domain knowledge

$$l_{ik} = \boldsymbol{\theta}_a(\mathbf{v}_i, \mathbf{z}_k, \tau)$$

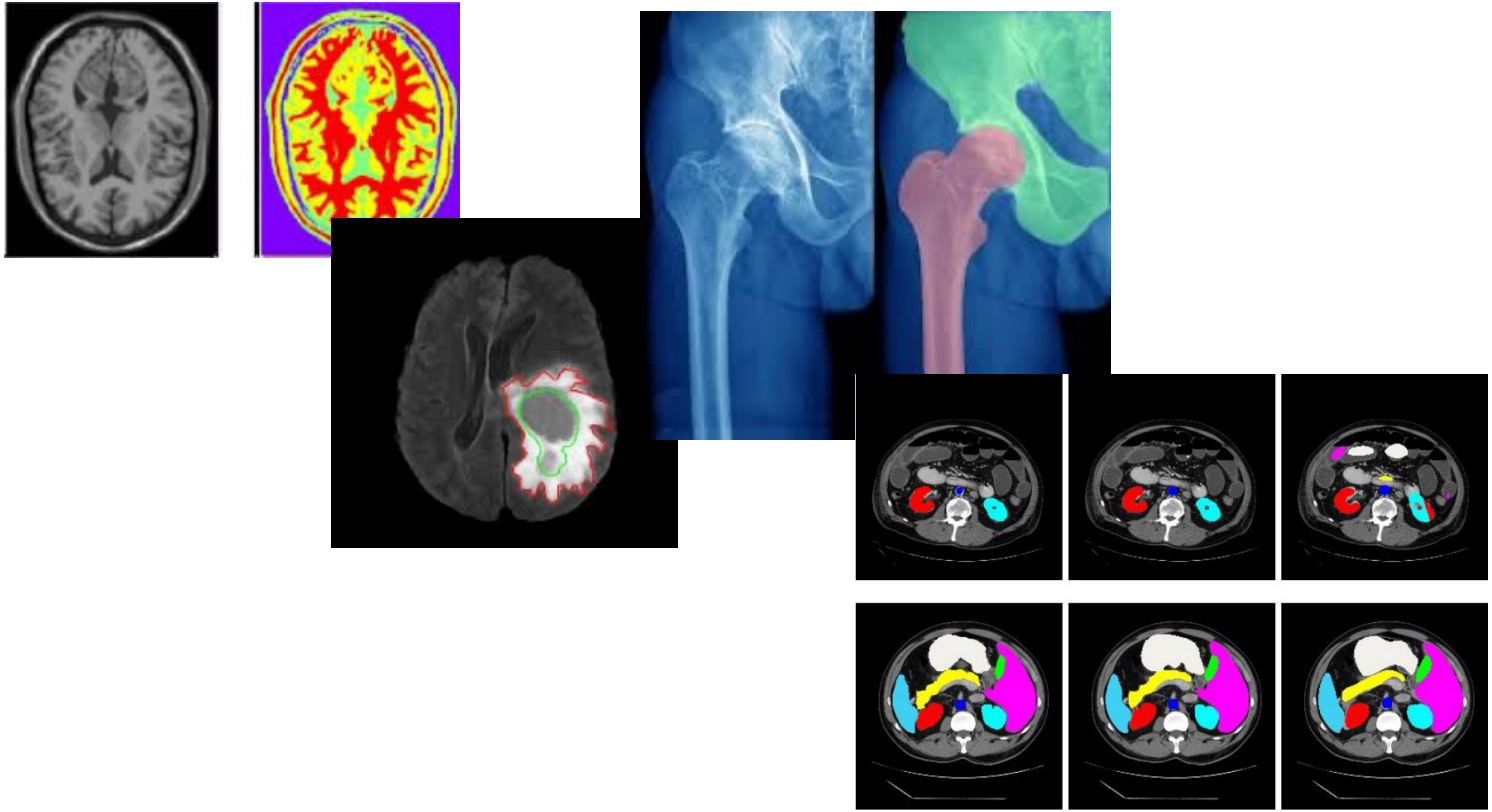


"A fundus image of possible signs of hemorrhage with blot, dot, or flame-shaped"

Expert Knowledge

Silva-Rodriguez et *al*. A Foundation LAnguage-Image model of the Retina (FLAIR): Encoding expert knowledge in text supervision. MedIA'24 (Under review)

What can we do in image segmentation?

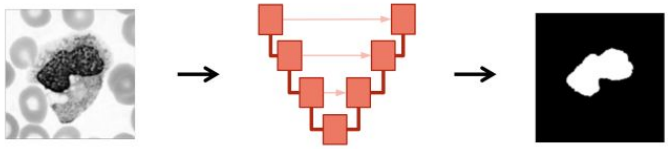


Foundational models for segmentation

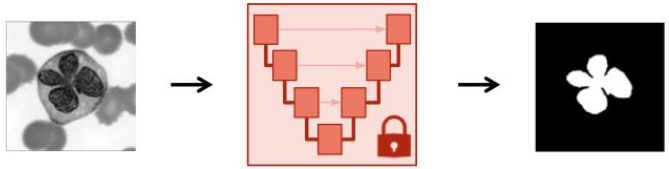
UniSeg

Traditional Approach

1. Design and train a task-specific model.

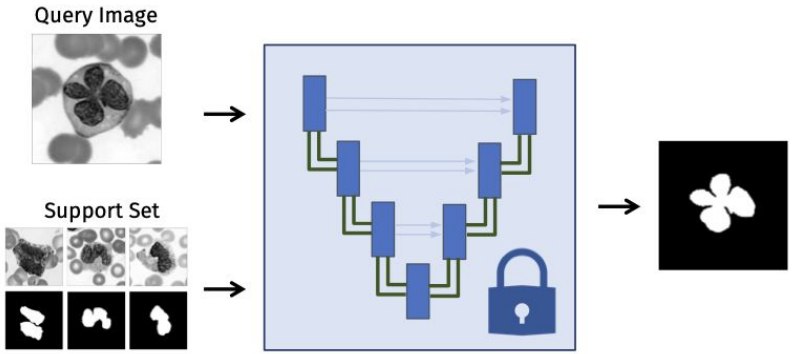


2. Predict new images with the trained model.

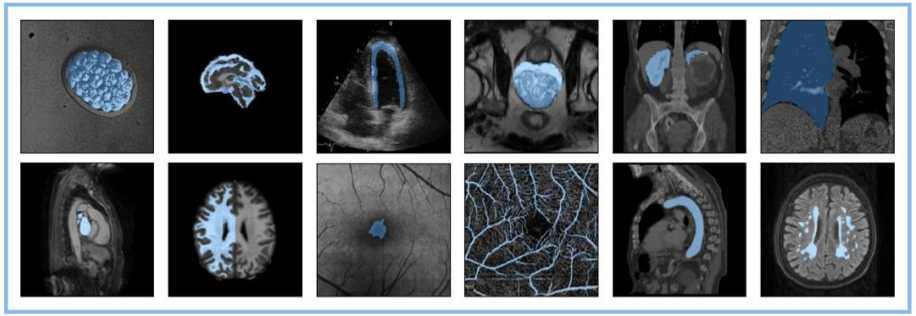


UniverSeg Approach

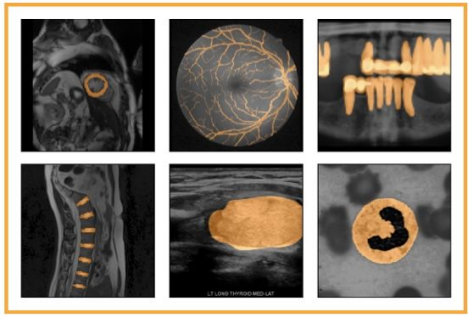
With a trained UniverSeg model, predict new images for the new task from a few labeled pairs without retraining.



Train Segmentation Tasks



Test Segmentation Tasks

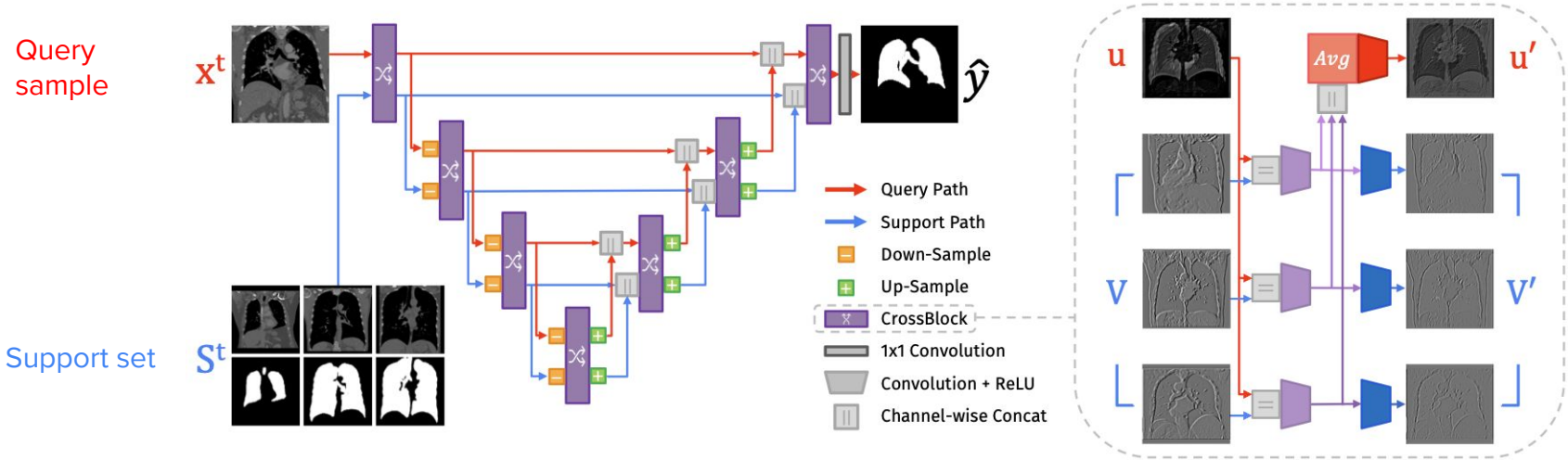


Butoi et *al*. Universeg: Universal medical image segmentation. ICCV'23

Foundational models for segmentation

UniSeg

Main Idea

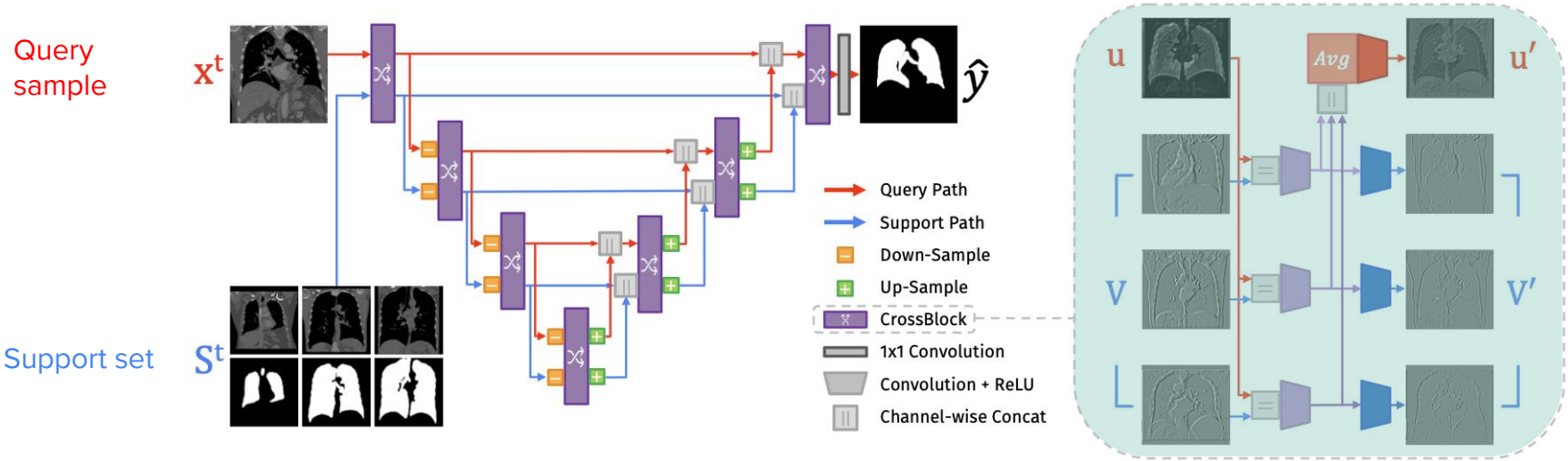


Butoi et *al*. Universeg: Universal medical image segmentation. ICCV'23

Foundational models for segmentation

UniSeg

Main Idea



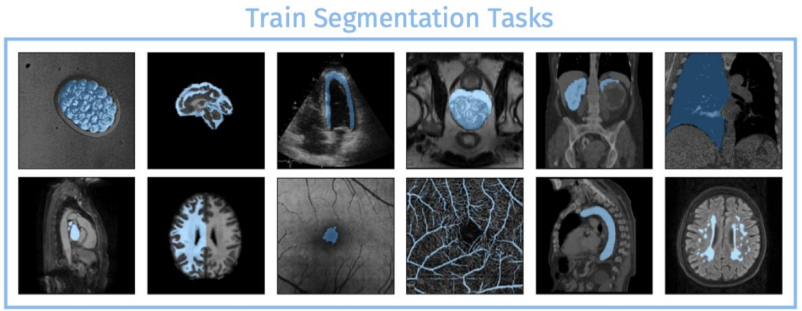
The representations from the query and support samples can interact at multiple scales

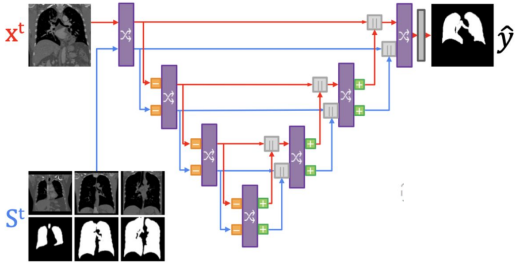
Butoi et *al*. Universeg: Universal medical image segmentation. ICCV'23

Foundational models for segmentation

UniSeg

How this is trained? (Hint: based on meta-learning or *learning-to-learn*)





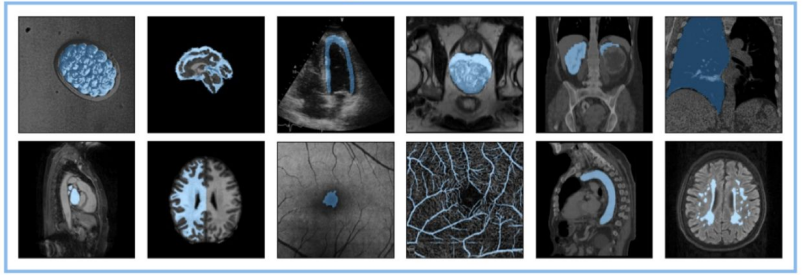
Butoi et *al*. Universeg: Universal medical image segmentation. ICCV'23

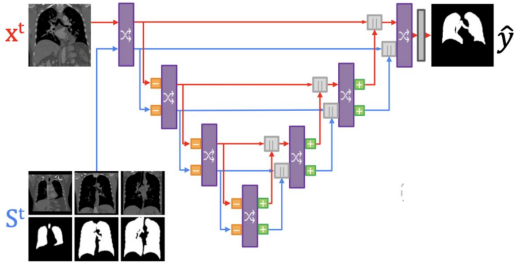
Foundational models for segmentation

UniSeg

How this is trained? (Hint: based on meta-learning or *learning-to-learn*)

Train Segmentation Tasks





for $k = 1, \ldots,$ NumTrainSteps **do**

$t \sim \mathcal{T}$ ▷ Sample Task

$(x_i^t, y_i^t) \sim t$ ▷ Sample Query

$S^t \leftarrow \{(x_j^t, y_j^t)\}_{j \neq i}^n$ ▷ Sample Support

$x_i^t, y_i^t \leftarrow \text{Aug}_t(x_i^t, y_i^t)$ ▷ Augment Query

$S^t \leftarrow \{\text{Aug}_t(x_j^t, y_j^t)\}_j^n$ ▷ Augment Support

$x_i^t, y_i^t, S^t \leftarrow \text{Aug}_T(x_i^t, y_i^t, S^t)$ ▷ Task Aug

$\hat{y}_i \leftarrow f_\theta(x_i^t, S^t)$ ▷ Predict label map

$\ell \leftarrow \mathcal{L}_{\text{seg}}(\hat{y}_i, y_i^t)$ ▷ Compute loss

$\theta \leftarrow \theta - \eta \nabla_\theta \ell$ ▷ Gradient step

end for

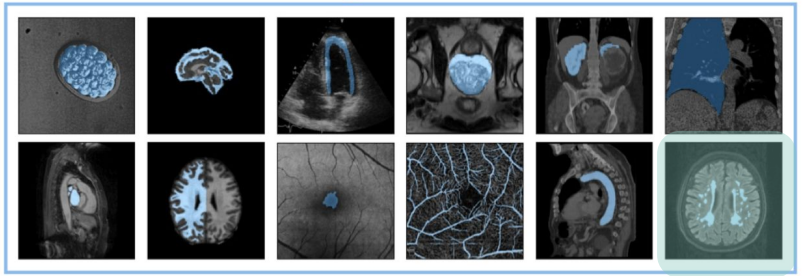
Butoi et *al*. Universeg: Universal medical image segmentation. ICCV'23

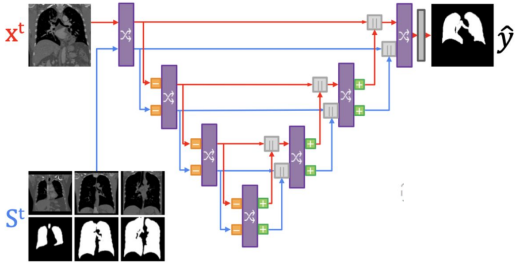
Foundational models for segmentation

UniSeg

How this is trained? (Hint: based on meta-learning or *learning-to-learn*)

Train Segmentation Tasks





for $k = 1, \ldots,$ NumTrainSteps **do**

$t \sim \mathcal{T}$ ▷ Sample Task

$(x_i^t, y_i^t) \sim t$ ▷ Sample Query

$S^t \leftarrow \{(x_j^t, y_j^t)\}_{j \neq i}^n$ ▷ Sample Support

$x_i^t, y_i^t \leftarrow \text{Aug}_t(x_i^t, y_i^t)$ ▷ Augment Query

$S^t \leftarrow \{\text{Aug}_t(x_j^t, y_j^t)\}_j^n$ ▷ Augment Support

$x_i^t, y_i^t, S^t \leftarrow \text{Aug}_T(x_i^t, y_i^t, S^t)$ ▷ Task Aug

$\hat{y}_i \leftarrow f_\theta(x_i^t, S^t)$ ▷ Predict label map

$\ell \leftarrow \mathcal{L}_{\text{seg}}(\hat{y}_i, y_i^t)$ ▷ Compute loss

$\theta \leftarrow \theta - \eta \nabla_\theta \ell$ ▷ Gradient step

end for

Among all training tasks

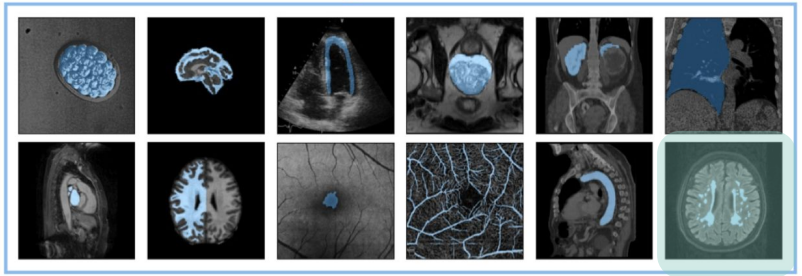
Butoi et *al*. Universeg: Universal medical image segmentation. ICCV'23

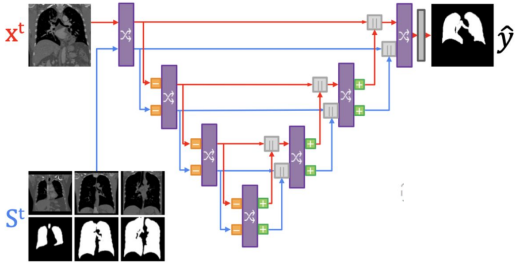
Foundational models for segmentation

UniSeg

How this is trained? (Hint: based on meta-learning or *learning-to-learn*)

Train Segmentation Tasks





for $k = 1, \ldots,$ NumTrainSteps **do**

$t \sim \mathcal{T}$ ▷ Sample Task

$(x_i^t, y_i^t) \sim t$ ▷ Sample Query

$S^t \leftarrow \{(x_j^t, y_j^t)\}_{j \neq i}^n$ ▷ Sample Support

$x_i^t, y_i^t \leftarrow \text{Aug}_t(x_i^t, y_i^t)$ ▷ Augment Query

$S^t \leftarrow \{\text{Aug}_t(x_j^t, y_j^t)\}_j^n$ ▷ Augment Support

$x_i^t, y_i^t, S^t \leftarrow \text{Aug}_T(x_i^t, y_i^t, S^t)$ ▷ Task Aug

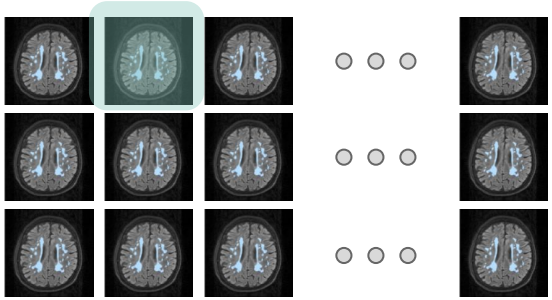
$\hat{y}_i \leftarrow f_\theta(x_i^t, S^t)$ ▷ Predict label map

$\ell \leftarrow \mathcal{L}_{\text{seg}}(\hat{y}_i, y_i^t)$ ▷ Compute loss

$\theta \leftarrow \theta - \eta \nabla_\theta \ell$ ▷ Gradient step

end for

Among all training samples from that task



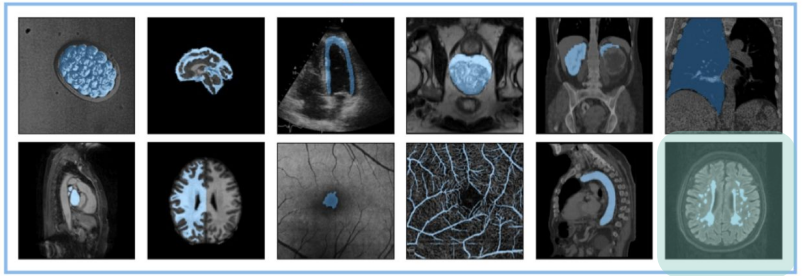
Butoi et *al*. Universeg: Universal medical image segmentation. ICCV'23

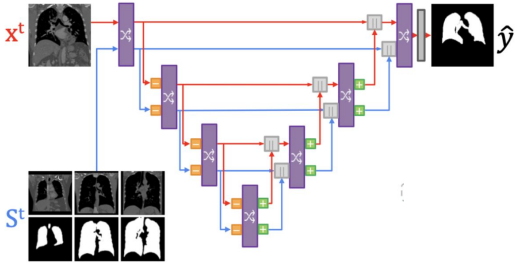
Foundational models for segmentation

UniSeg

How this is trained? (Hint: based on meta-learning or *learning-to-learn*)

Train Segmentation Tasks





for $k = 1, \ldots,$ NumTrainSteps **do**

$t \sim \mathcal{T}$ ▷ Sample Task

$(x_i^t, y_i^t) \sim t$ ▷ Sample Query

$S^t \leftarrow \{(x_j^t, y_j^t)\}_{j \neq i}^n$ ▷ Sample Support

$x_i^t, y_i^t \leftarrow \text{Aug}_t(x_i^t, y_i^t)$ ▷ Augment Query

$S^t \leftarrow \{\text{Aug}_t(x_j^t, y_j^t)\}_j^n$ ▷ Augment Support

$x_i^t, y_i^t, S^t \leftarrow \text{Aug}_T(x_i^t, y_i^t, S^t)$ ▷ Task Aug

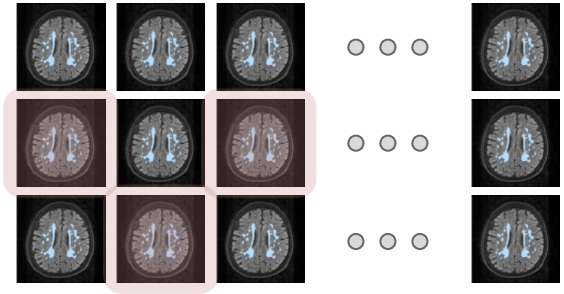
$\hat{y}_i \leftarrow f_\theta(x_i^t, S^t)$ ▷ Predict label map

$\ell \leftarrow \mathcal{L}_{\text{seg}}(\hat{y}_i, y_i^t)$ ▷ Compute loss

$\theta \leftarrow \theta - \eta \nabla_\theta \ell$ ▷ Gradient step

end for

Among all training samples from that task



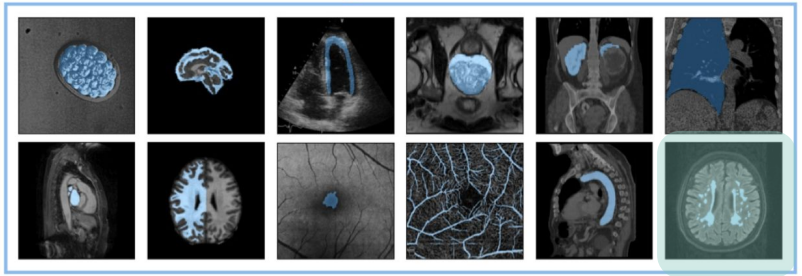
Butoi et *al*. Universeg: Universal medical image segmentation. ICCV'23

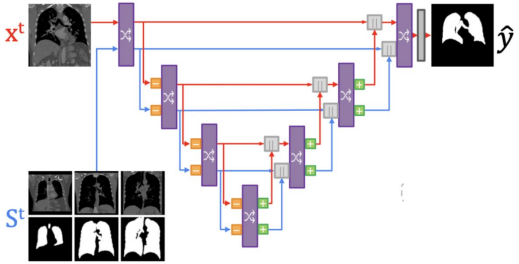
Foundational models for segmentation

UniSeg

How this is trained? (Hint: based on meta-learning or *learning-to-learn*)

Train Segmentation Tasks





for $k = 1, \ldots,$ NumTrainSteps **do**

$t \sim \mathcal{T}$ ▷ Sample Task

$(x_i^t, y_i^t) \sim t$ ▷ Sample Query

$S^t \leftarrow \{(x_j^t, y_j^t)\}_{j \neq i}^n$ ▷ Sample Support

$x_i^t, y_i^t \leftarrow \mathrm{Aug}_t(x_i^t, y_i^t)$ ▷ Augment Query

$S^t \leftarrow \{\mathrm{Aug}_t(x_j^t, y_j^t)\}_j^n$ ▷ Augment Support

$x_i^t, y_i^t, S^t \leftarrow \mathrm{Aug}_T(x_i^t, y_i^t, S^t)$ ▷ Task Aug

$\hat{y}_i \leftarrow f_\theta(x_i^t, S^t)$ ▷ Predict label map

$\ell \leftarrow \mathcal{L}_{\text{seg}}(\hat{y}_i, y_i^t)$ ▷ Compute loss

$\theta \leftarrow \theta - \eta \nabla_\theta \ell$ ▷ Gradient step

end for

Perform augmentations

In-task augmentations $\mathrm{Aug}_t(x, y)$

(All done independently)

Standard augmentation: affine transformations, elastic deformations, adding noise, etc

Task augmentations $\mathrm{Aug}_T(x, y, S)$

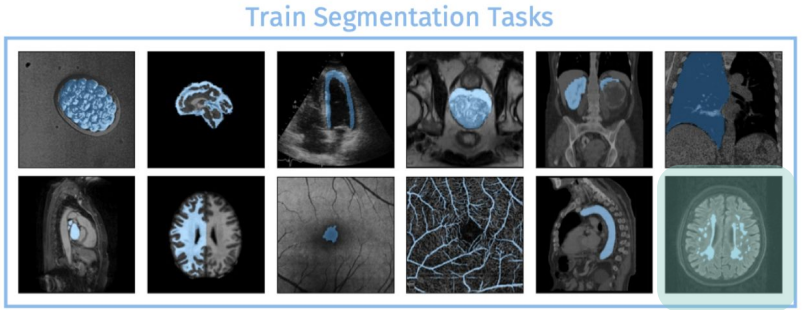
(Same type of augmentations for all images in a task)

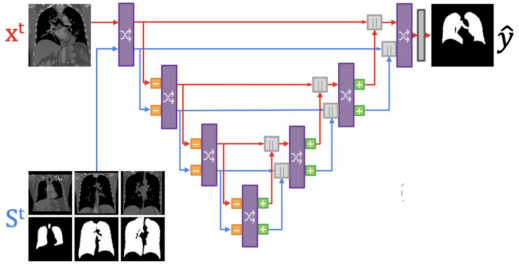
Butoi et *al*. Universeg: Universal medical image segmentation. ICCV'23

Foundational models for segmentation

UniSeg

How this is trained? (Hint: based on meta-learning or *learning-to-learn*)





for $k = 1, \ldots,$ NumTrainSteps **do**

$t \sim \mathcal{T}$ ▷ Sample Task

$(x_i^t, y_i^t) \sim t$ ▷ Sample Query

$S^t \leftarrow \{(x_j^t, y_j^t)\}_{j \neq i}^n$ ▷ Sample Support

$x_i^t, y_i^t \leftarrow \text{Aug}_t(x_i^t, y_i^t)$ ▷ Augment Query

$S^t \leftarrow \{\text{Aug}_t(x_j^t, y_j^t)\}_j^n$ ▷ Augment Support

$x_i^t, y_i^t, S^t \leftarrow \text{Aug}_T(x_i^t, y_i^t, S^t)$ ▷ Task Aug

$\hat{y}_i \leftarrow f_\theta(x_i^t, S^t)$ ▷ Predict label map

$\ell \leftarrow \mathcal{L}_{\text{seg}}(\hat{y}_i, y_i^t)$ ▷ Compute loss

$\theta \leftarrow \theta - \eta \nabla_\theta \ell$ ▷ Gradient step

end for

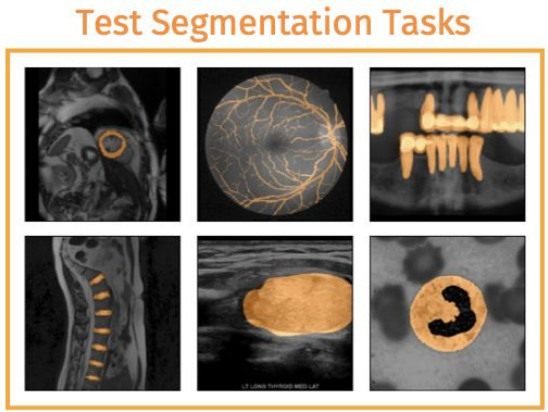
Standard training steps

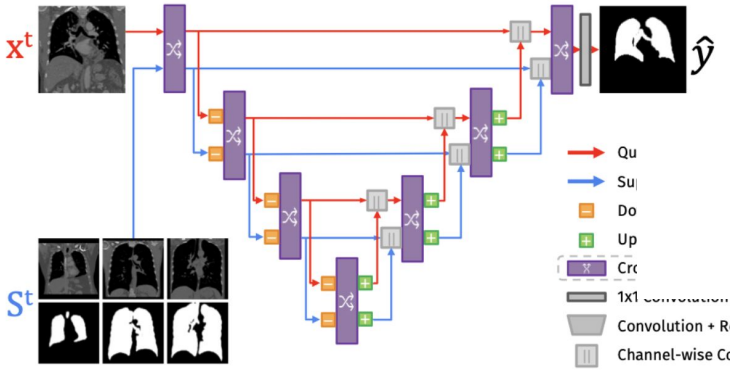
Butoi et *al*. Universeg: Universal medical image segmentation. ICCV'23

Foundational models for segmentation

UniSeg

And what about inference?





For a given image x^t $\hat{y} = f_\theta(x^t, S^t)$

To make it more robust, multiple support sets are employed

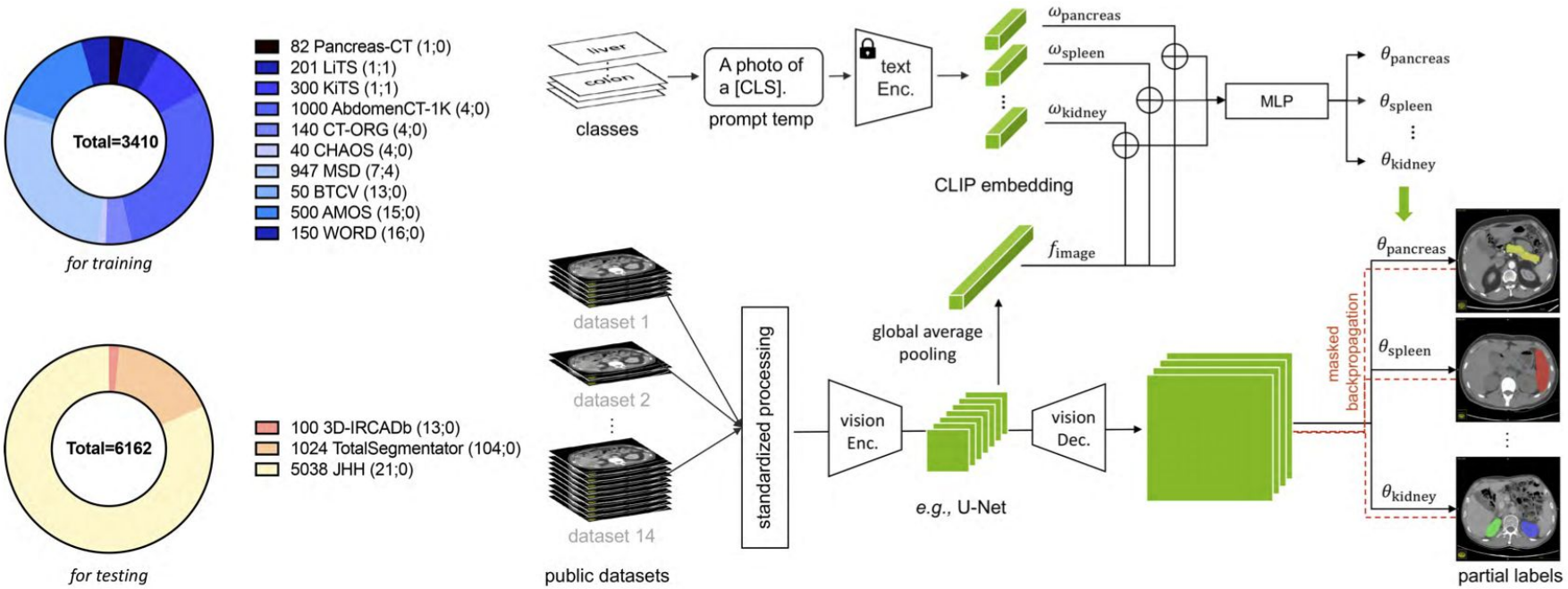
$$\hat{y} = \frac{1}{M}\sum_{m=1}^{M} f_\theta(x^t, S_m^t)$$

Butoi et *al*. Universeg: Universal medical image segmentation. ICCV'23

Foundational models for segmentation

CLIP-Driven

Main idea



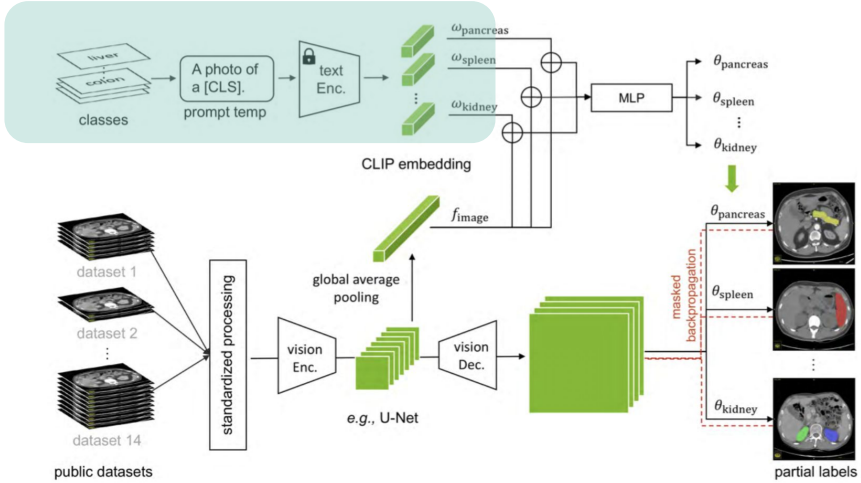
Liu et *al*. CLIP-Driven universal model for organ segmentation and tumor detection. ICCV'23

Foundational models for segmentation

CLIP-Driven

Main idea

Text branch
(generates text embedding for class k) $\mathbf{w}_k$



Liu et *al*. CLIP-Driven universal model for organ segmentation and tumor detection. ICCV'23

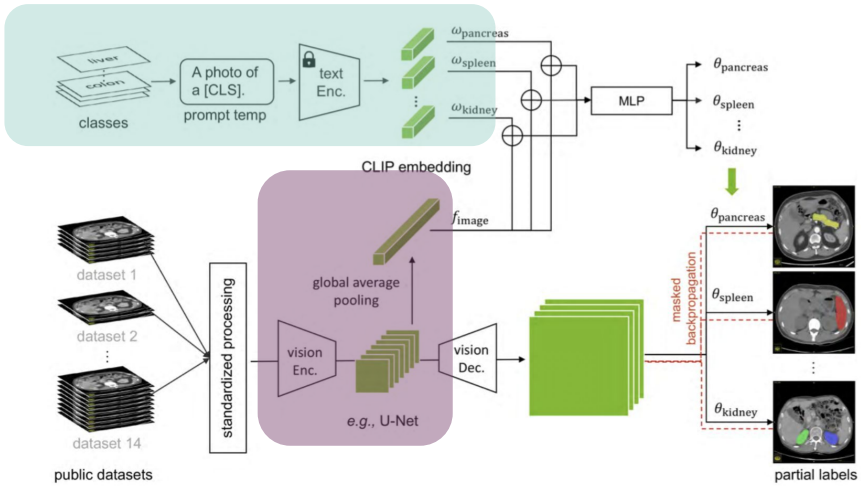
Foundational models for segmentation

CLIP-Driven

Main idea

Text branch
(generates text embedding for class k) $\mathbf{w}_k$

Visual branch-encoder
(generates visual embedding for image x) $\mathbf{f}$



Liu et *al*. CLIP-Driven universal model for organ segmentation and tumor detection. ICCV'23

Foundational models for segmentation

CLIP-Driven

Main idea

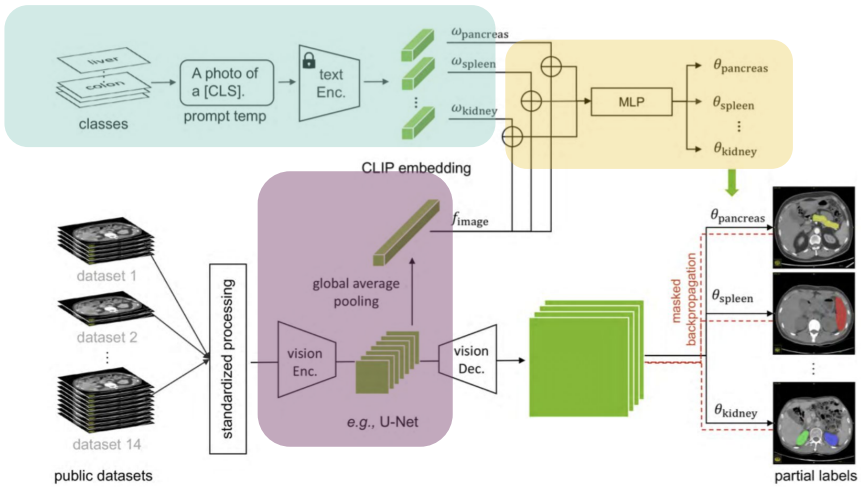
Text branch
(generates text embedding for class k) $\mathbf{w}_k$

Visual branch-encoder
(generates visual embedding for image x) $\mathbf{f}$

Text-based controller MLP
(generates class parameters)

$$\boldsymbol{\theta}_k = MLP(\mathbf{w}_k \oplus \mathbf{f})$$

$$\boldsymbol{\theta}_k = \{\boldsymbol{\theta}_{k_1}, \boldsymbol{\theta}_{k_2}, \boldsymbol{\theta}_{k_3}\}$$



Liu et *al*. CLIP-Driven universal model for organ segmentation and tumor detection. ICCV'23

Foundational models for segmentation

CLIP-Driven

Main idea

Text branch
(generates text embedding for class k) $\mathbf{w}_k$

Visual branch-encoder
(generates visual embedding for image x) $\mathbf{f}$

Text-based controller MLP
(generates class parameters)

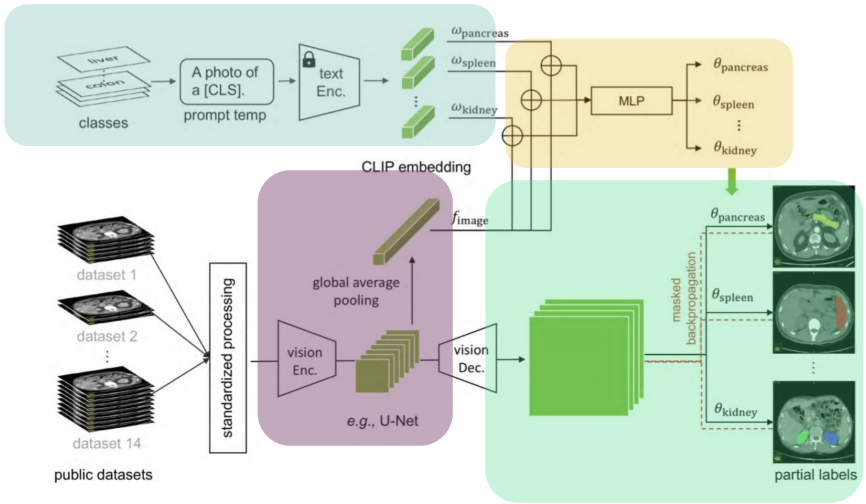
$$\boldsymbol{\theta_k} = MLP(\mathbf{w}_k \oplus \mathbf{f})$$

$$\boldsymbol{\theta}_k = \{\boldsymbol{\theta}_{k_1}, \boldsymbol{\theta}_{k_2}, \boldsymbol{\theta}_{k_3}\}$$

Visual branch-decoder
(generates visual embedding for image x)

$$\mathbf{P}_k = \text{sigmoid}(((\mathbf{F} * \boldsymbol{\theta}_{k_1}) * \boldsymbol{\theta}_{k_2}) * \boldsymbol{\theta}_{k_3})$$

It represents foreground class k *vs* background



Liu et *al*. CLIP-Driven universal model for organ segmentation and tumor detection. ICCV'23

Foundational models for segmentation

CLIP-Driven

Main idea

Text branch
(generates text embedding for class k) $\mathbf{w}_k$

Visual branch-encoder
(generates visual embedding for image x) $\mathbf{f}$

Text-based controller MLP
(generates class parameters)

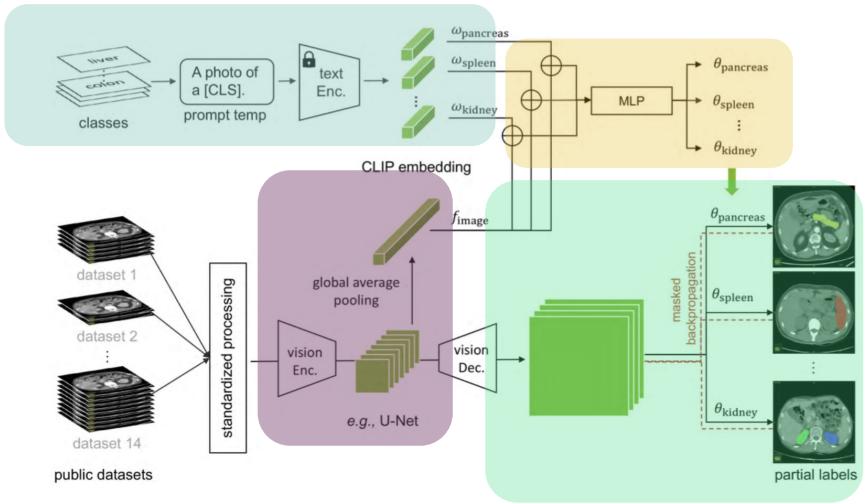
$$\boldsymbol{\theta}_k = MLP(\mathbf{w}_k \oplus \mathbf{f})$$

$$\boldsymbol{\theta}_k = \{\boldsymbol{\theta}_{k_1}, \boldsymbol{\theta}_{k_2}, \boldsymbol{\theta}_{k_3}\}$$

Visual branch-decoder
(generates visual embedding for image x)

$$\mathbf{P}_k = \text{sigmoid}(((\mathbf{F} * \boldsymbol{\theta}_{k_1}) * \boldsymbol{\theta}_{k_2}) * \boldsymbol{\theta}_{k_3})$$

It represents foreground class k *vs* background



Training loss

Binary cross-entropy per class (and terms masked for those classes not present)

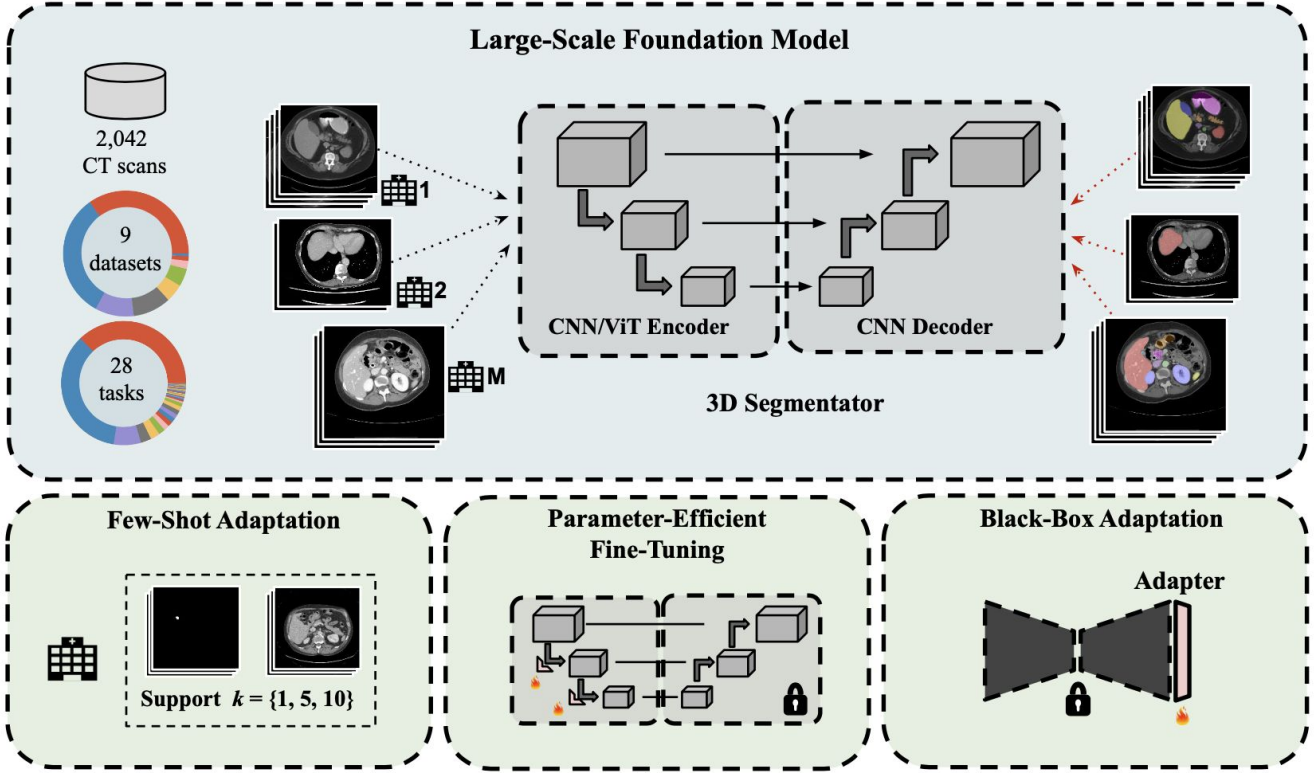
$$\mathcal{L} = \sum_{k=1}^{K} \mathbf{1}_{\{k \in y\}} \cdot \text{BCE}_k$$

Liu et *al*. CLIP-Driven universal model for organ segmentation and tumor detection. ICCV'23

Foundational models for segmentation

FSEFT-Adapters

Main idea (how to adapt a pre-trained large-scale model)



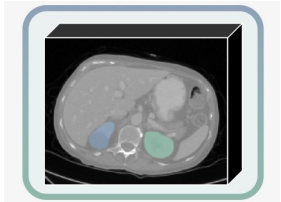
Silva-Rodriguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MedIA'24 (Under Review)

Foundational models for segmentation

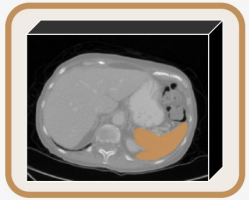
FSEFT-Adapters

Pre-training Large-scale foundation model

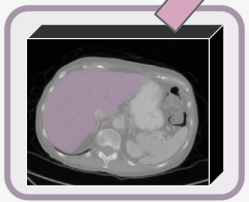
Assembling the dataset



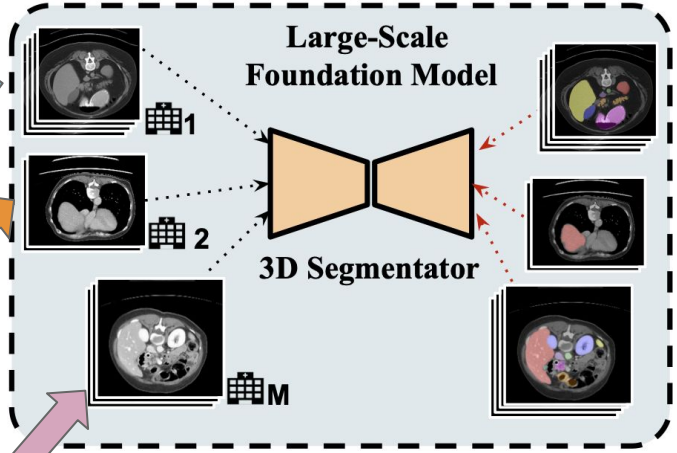
Dataset A: kidney



Dataset B: spleen



Dataset D: liver



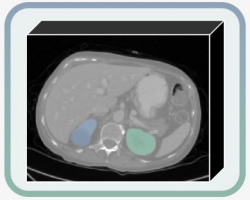
Silva-Rodriguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MedIA'24 (Under Review)

Foundational models for segmentation

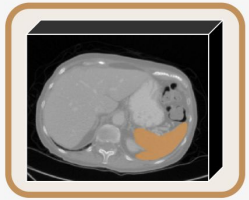
FSEFT-Adapters

Pre-training Large-scale foundation model

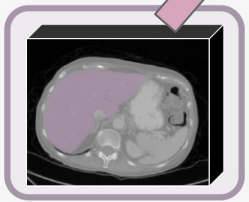
Assembling the dataset



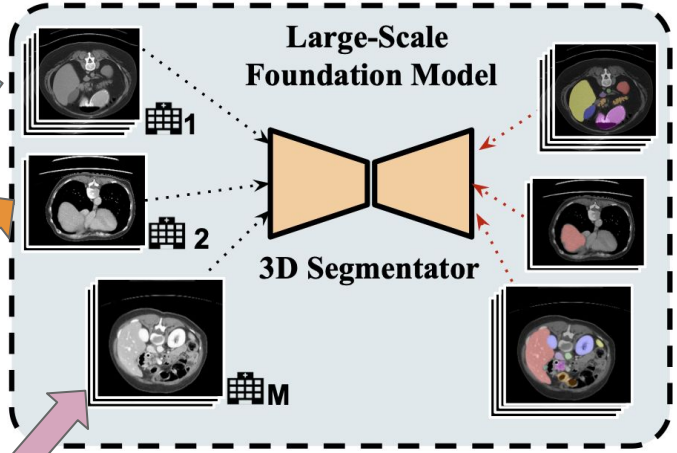
Dataset A: kidney



Dataset B: spleen



Dataset D: liver



$$\mathcal{D}_T = \{(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{w}_n)\}_{n=1}^{N}$$

Input image

Partial Mask

Multi-label one-hot encoding

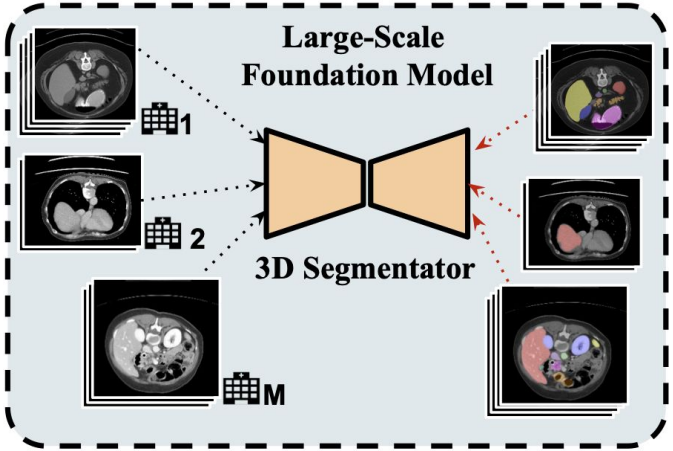
Silva-Rodriguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MedIA'24 (Under Review)

Foundational models for segmentation

FSEFT-Adapters

Pre-training Large-scale foundation model

Training



$$\mathcal{D}_T = \{(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{w}_n)\}_{n=1}^{N}$$

Loss function

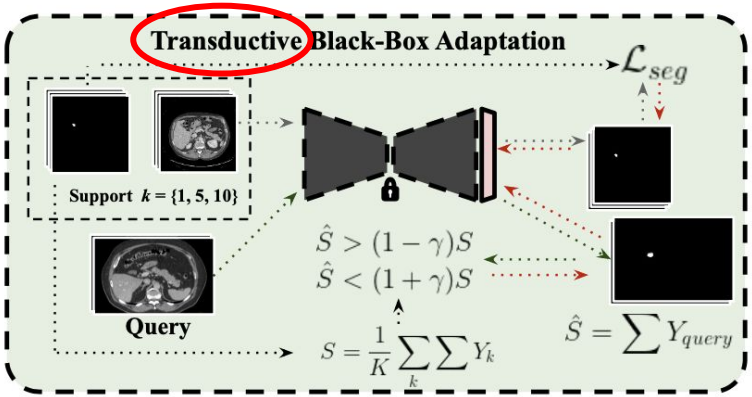
$$\min_{\theta_f, \theta_c} \quad \frac{1}{\sum_k \mathbf{w}_{n,k}} \sum_k \mathbf{w}_{n,k} \mathcal{L}_{SEG}(\mathbf{Y}_{n,k}, \hat{\mathbf{Y}}_{n,k}), \quad n = 1, \ldots, N$$

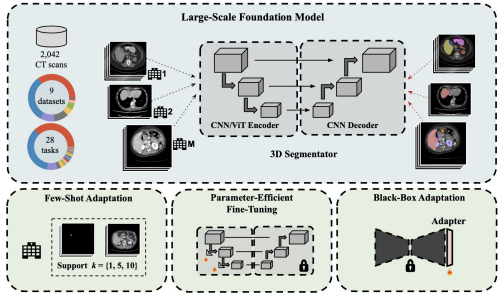
Silva-Rodriguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MedIA'24 (Under Review)

Foundational models for segmentation

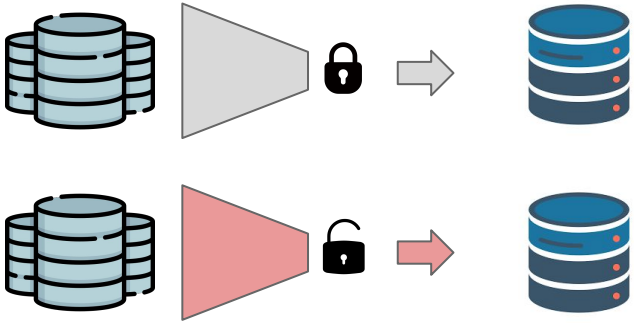
FSEFT-Adapters

Adaptation (a few labeled samples)





Inductive *vs.* Transductive



Train set

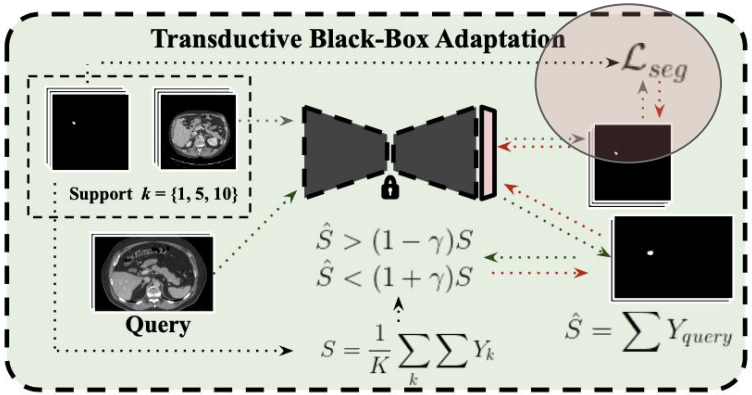
Test set

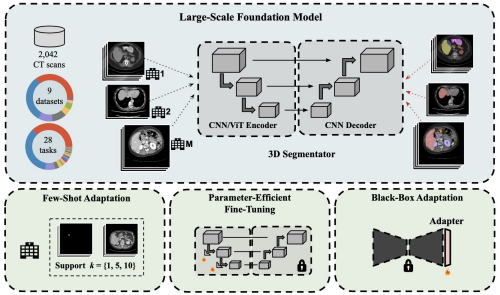
Silva-Rodriguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MedIA'24 (Under Review)

Foundational models for segmentation

FSEFT-Adapters

Adaptation (a few labeled samples)





$$\min_{\phi_f, \phi_c} \mathcal{L}_{SEG}(Y_k, \hat{Y}_k) \qquad k = 1, \ldots, K$$

Adapters parameters
(we come to this soon....)

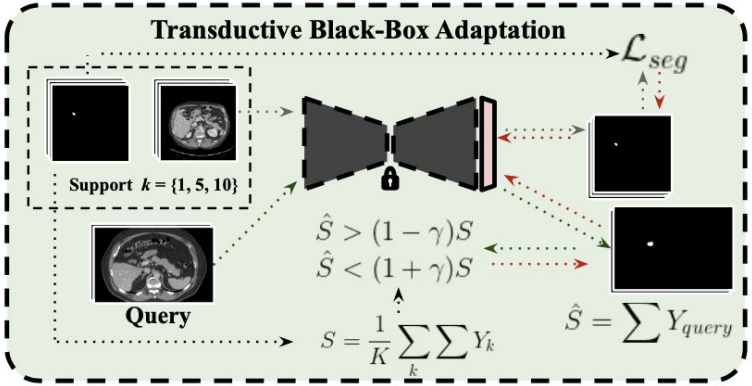
Standard segmentation
loss on the few shots

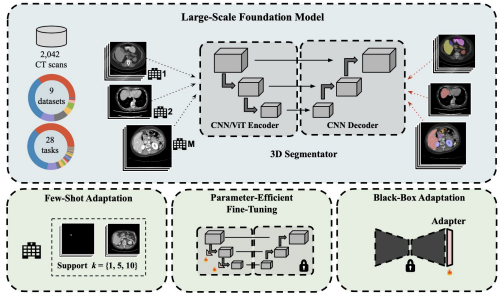
Silva-Rodriguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MedIA'24 (Under Review)

Foundational models for segmentation

FSEFT-Adapters

Adaptation (a few labeled samples)





Can we incorporate additional knowledge during optimization?

$$\min_{\phi_f, \phi_c} \mathcal{L}_{SEG}(Y_k, \hat{Y}_k) \qquad k = 1, \ldots, K$$

Adapters parameters (we come to this soon....)

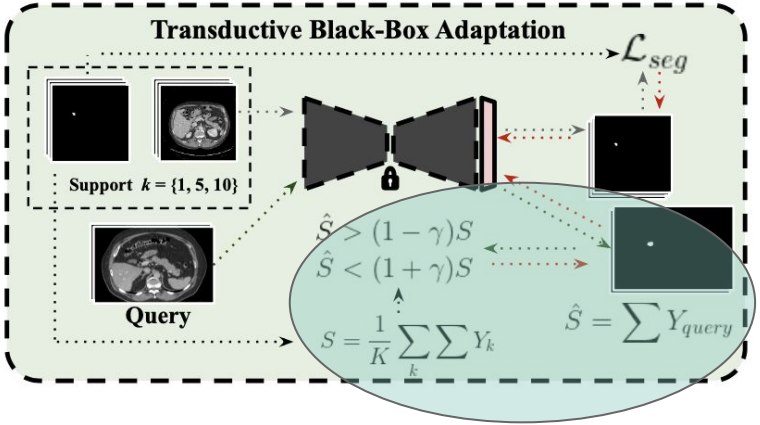
Standard segmentation loss on the few shots

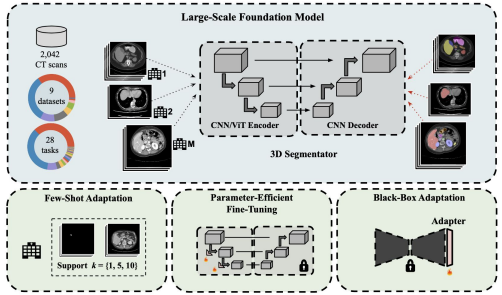
Silva-Rodriguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MedIA'24 (Under Review)

Foundational models for segmentation

FSEFT-Adapters

Adaptation (a few labeled samples)





Can we incorporate additional knowledge during optimization?

$$\min_{\phi_f, \phi_c} \quad \mathcal{L}_{SEG}(Y_k, \hat{Y}_k) + \lambda \mathcal{L}_{TI}(S, \hat{S}_{query}), \quad k = 1, \ldots, K$$

$$\mathcal{L}_{TI} = \begin{cases} |\hat{S} - (1-m)S|, & \text{if } \hat{S} < (1-m)S \\ |\hat{S} - (1+m)S|, & \text{if } \hat{S} > (1+m)S \\ 0, & \text{otherwise} \end{cases}$$

Adapters parameters (we come to this soon....)

Standard segmentation loss on the few shots

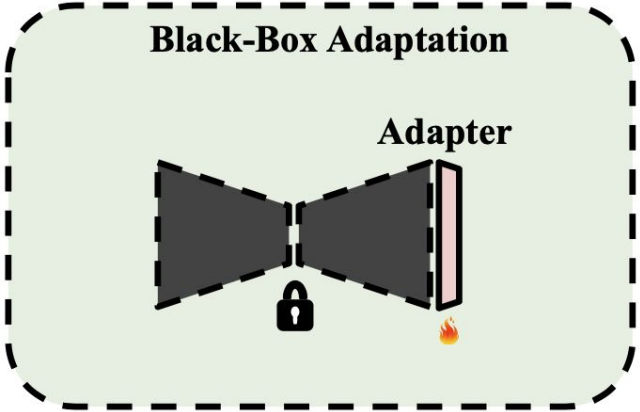
Anatomical priors (e.g., size)

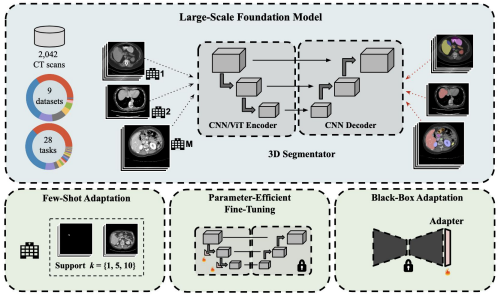
Silva-Rodriguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MedIA'24 (Under Review)

Foundational models for segmentation

FSEFT-Adapters

Adaptation (a few labeled samples)





What to modify?

$$\mathbf{Z}' = \mathbf{Z} + \phi_f(\mathbf{Z})$$

Modifies original projections

$$\phi_c = \{\mathcal{W}^{D\times 1}, \mathcal{B}^1\}$$

Classifier head

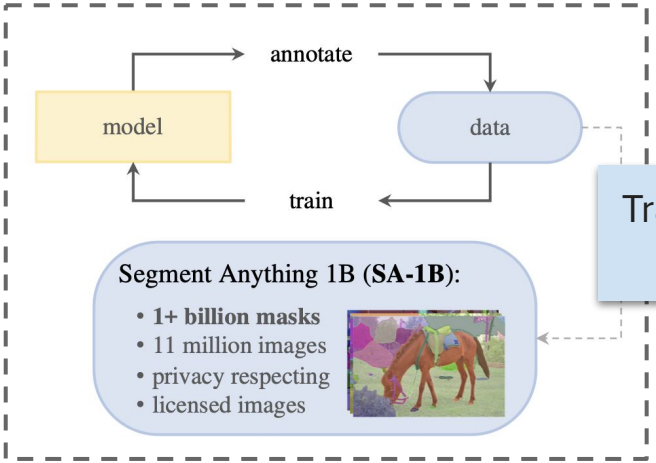
Silva-Rodriguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MedIA'24 (Under Review)

Foundational models for segmentation

What about SAM (Segment Anything Model)?

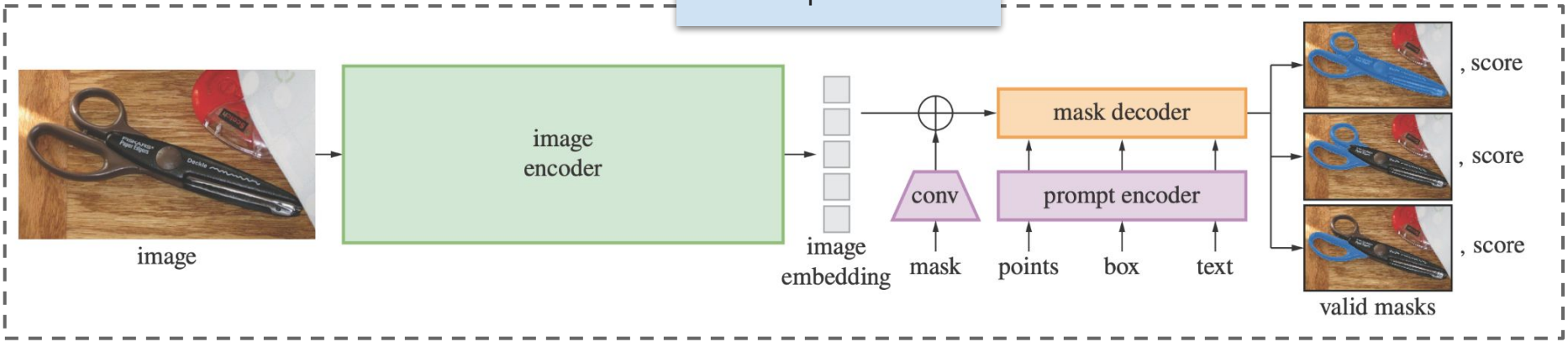
Foundational models for segmentation

SAM



Trained on an huge amount of data

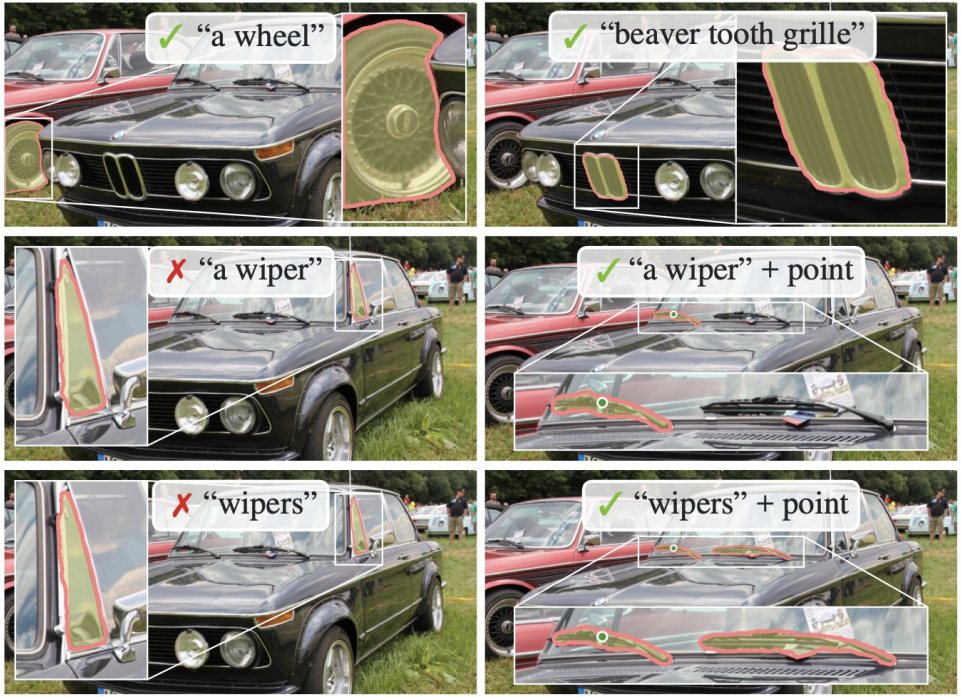
Pipeline

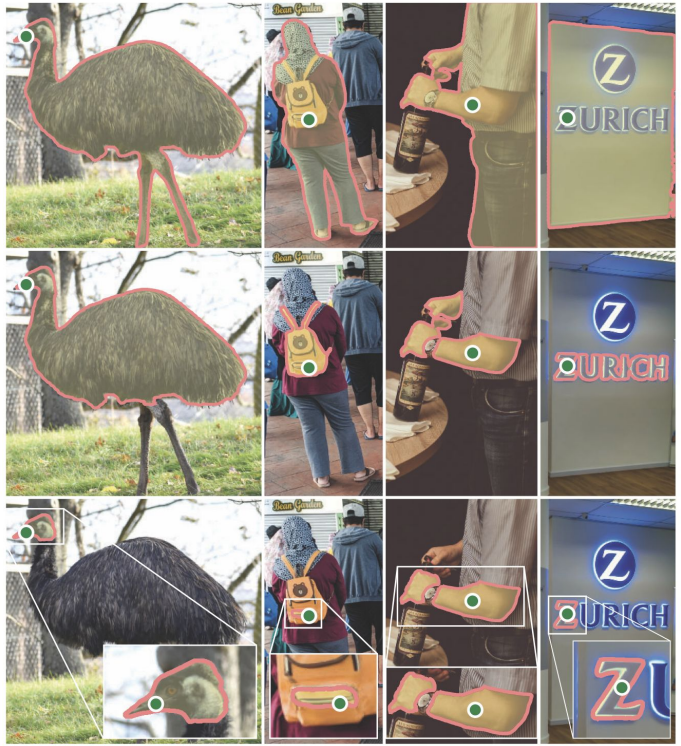


Kirillow et *al*. Segment Anything, ICCV'23

Foundational models for segmentation

SAM

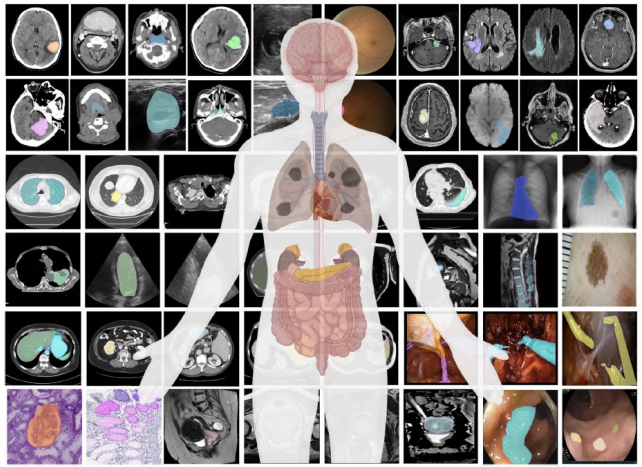




Kirillow et *al*. Segment Anything, ICCV'23

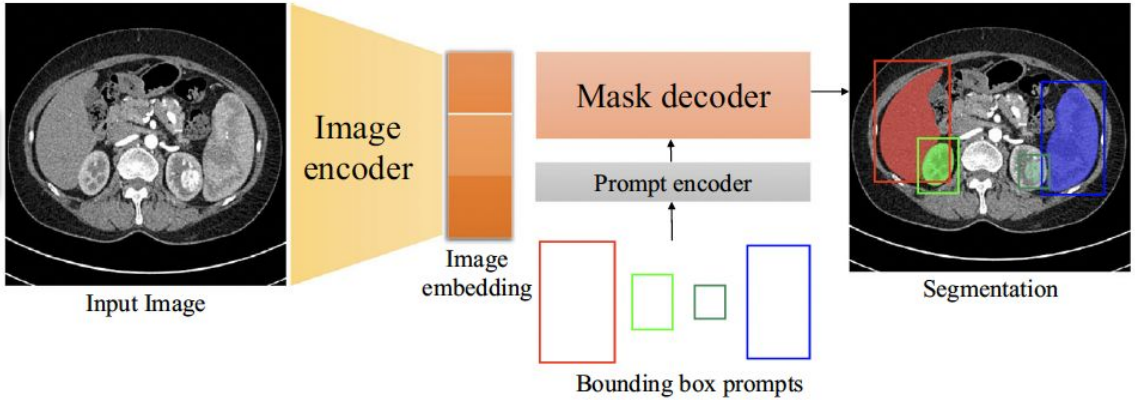
Foundational models for segmentation

MedSAM



Trained on an huge amount of data

Pipeline



Ma et *al*. Segment Anything in medical images, Nature Communications 2024

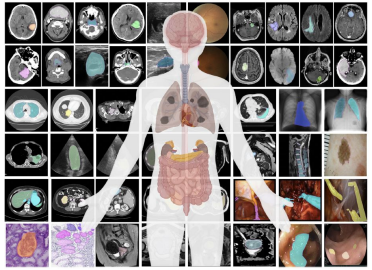
Foundational models for segmentation

SAM-related

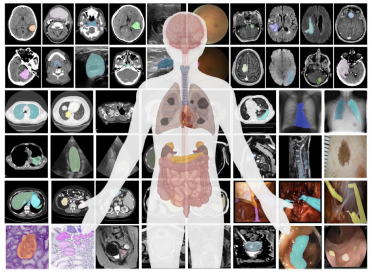
(Some) current questions

Better start with SAM or MedSAM (or similar)?

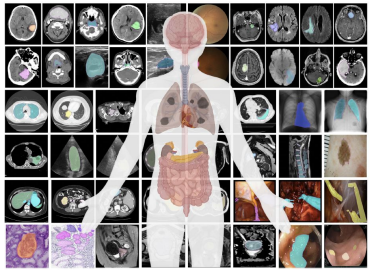
SAM



Medical domain



MedSAM



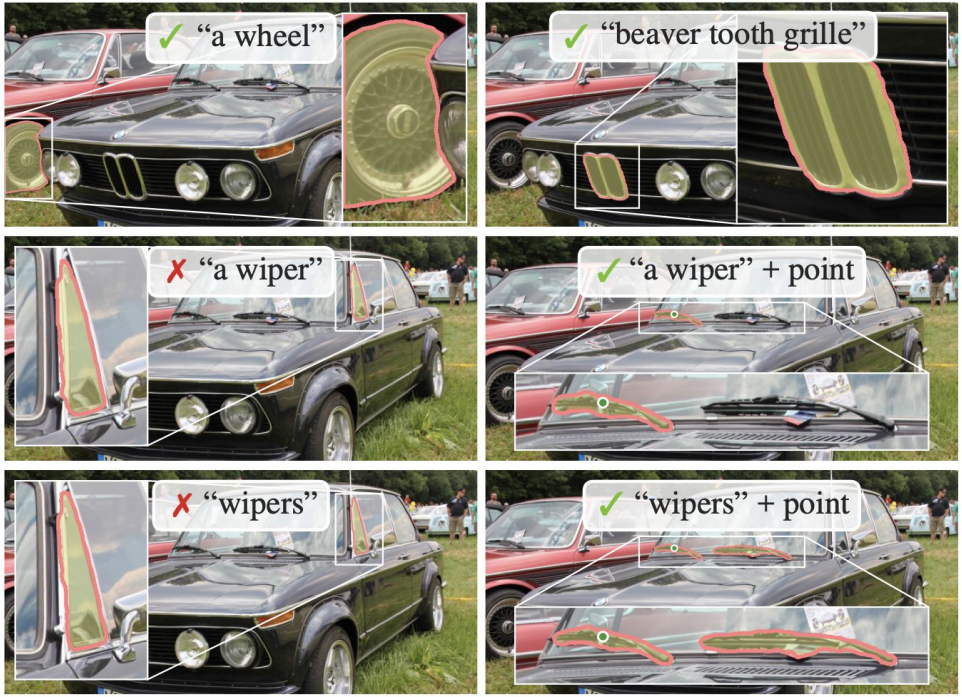
Medical domain

Foundational models for segmentation

SAM-related

(Some) current questions

How to optimize the prompts?

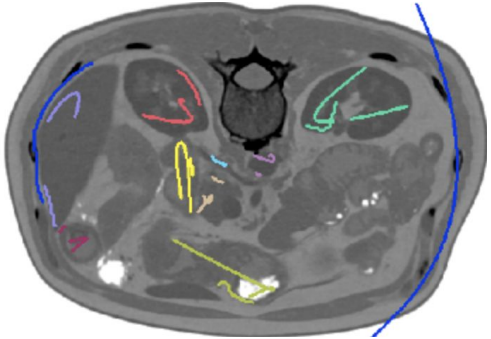


Foundational models for segmentation

SAM-related

(Some) current questions

SAM-based or task-specific?



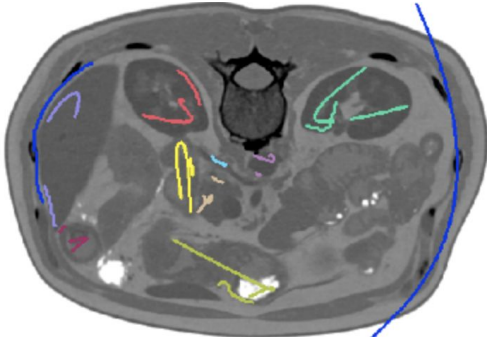
SAM is promptable
(i.e., requires user interaction
per EACH test image)

Foundational models for segmentation

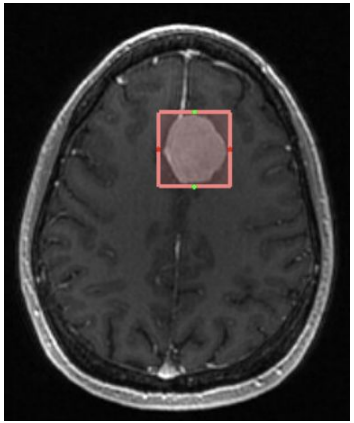
SAM-related

(Some) current questions

SAM-based or task-specific?



SAM is promptable
(i.e., requires user interaction
per EACH test image)



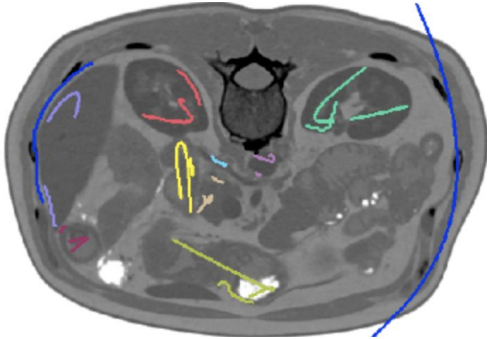
SAM only handles
binary segmentation
(one class at a time)

Foundational models for segmentation

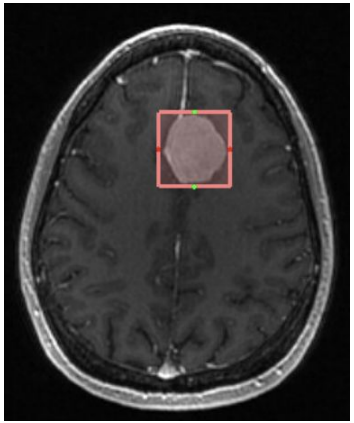
SAM-related

(Some) current questions

SAM-based or task-specific?



SAM is promptable
(i.e., requires user interaction
per EACH test image)



SAM only handles
binary segmentation
(one class at a time)

SAM yields sometimes
similar results to
task-specific models