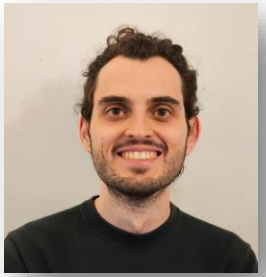
ETH zürich MICCAI2026

Semi-Supervised Few-Shot Adaptation of Vision-Language Models

Julio Silva-Rodríguez



Ender Konukoglu

Outline

A. Vision-Language Models (VLMs)

- Contrastive Language-Image Pre-training (CLIP).
- Zero-shot and few-shot inference.
- Medical vision-language models.

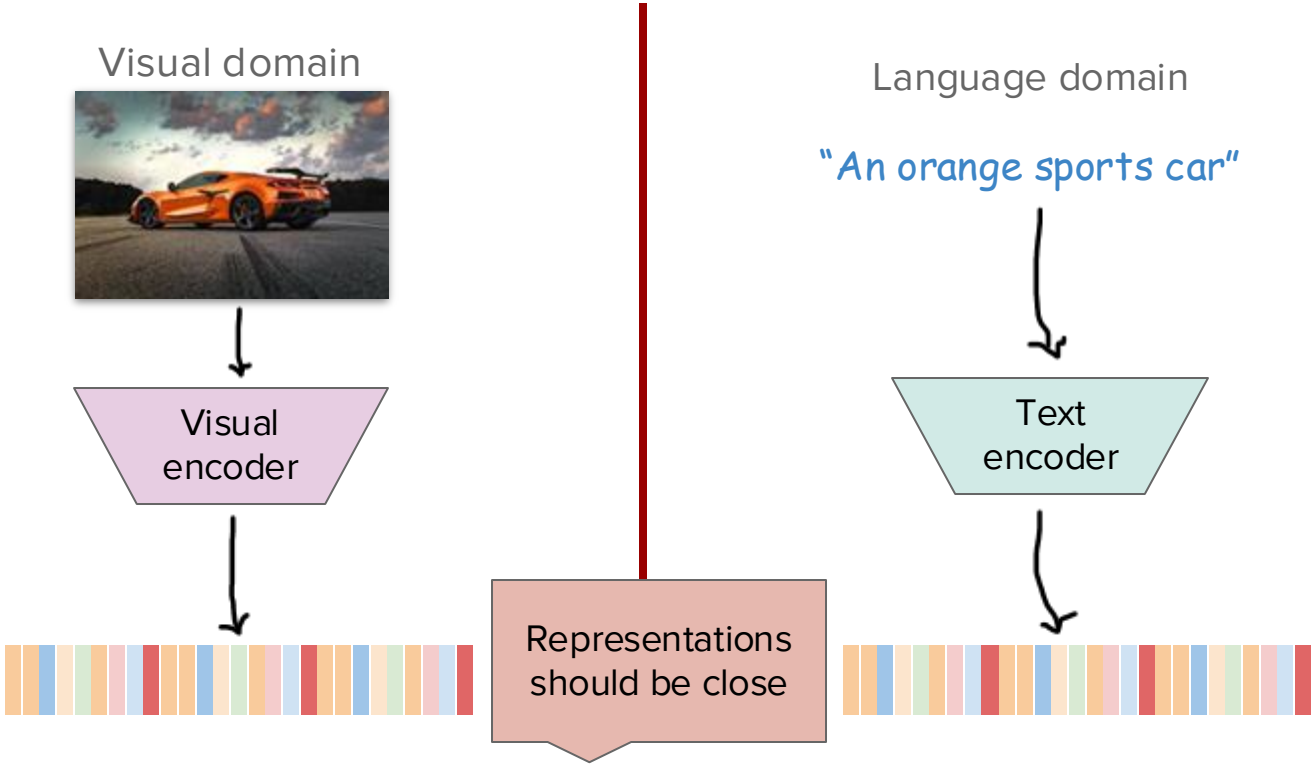
B. Few-shot adaptation

- Text-informed linear probes.

C. Semi-supervised few-shot adaptation

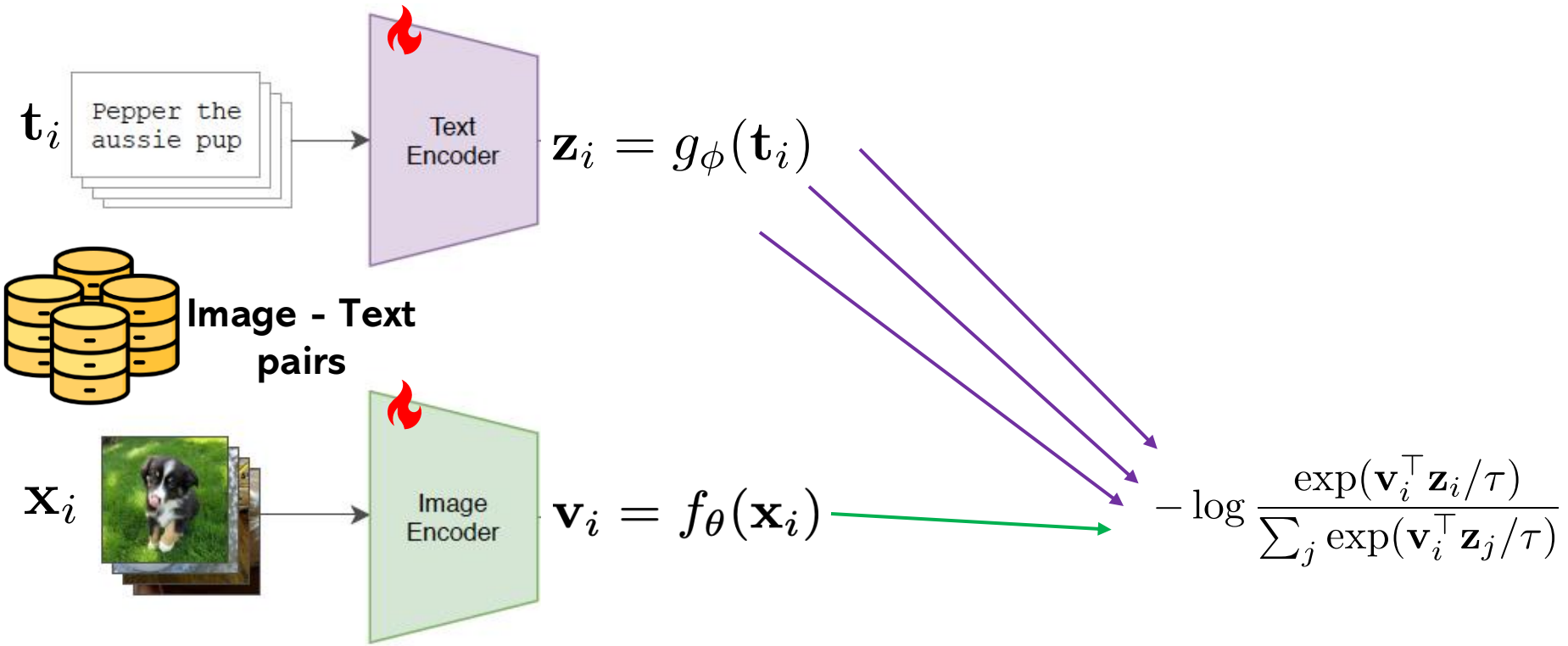
- Motivation.
- Learning from unlabeled data.
- SS-Text-U: objective and block-wise optimization.
- Results & exploratory analysis.

Embedding semantic knowledge at larger scale trough text supervision



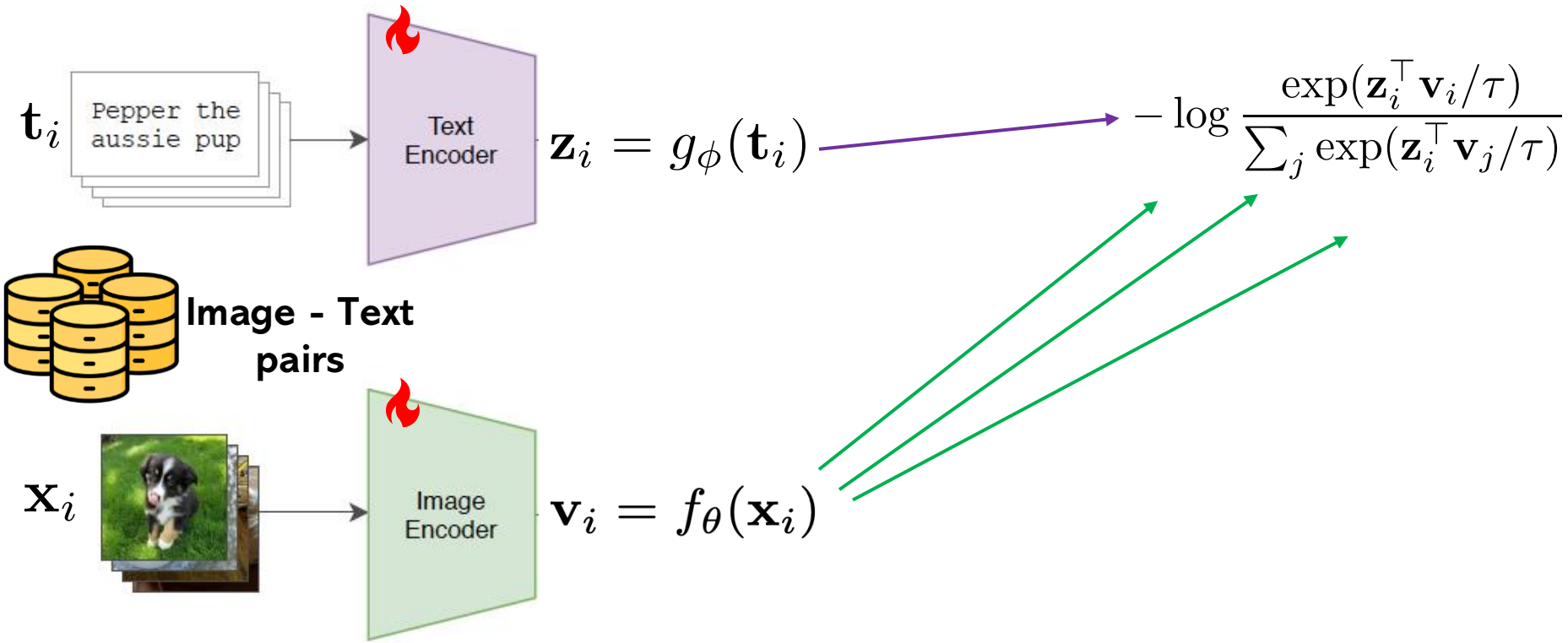
Radford et *al*. Learning transferable visual models from natural language supervision, **ICML'21**

Contrastive Language-Image Pre-training (CLIP)



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

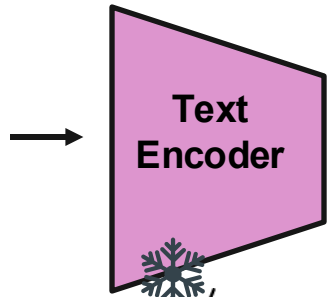
Contrastive Language-Image Pre-training (CLIP)



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

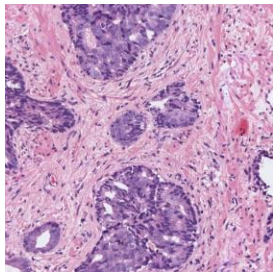
“Zero-shot” text-driven prediction for new tasks

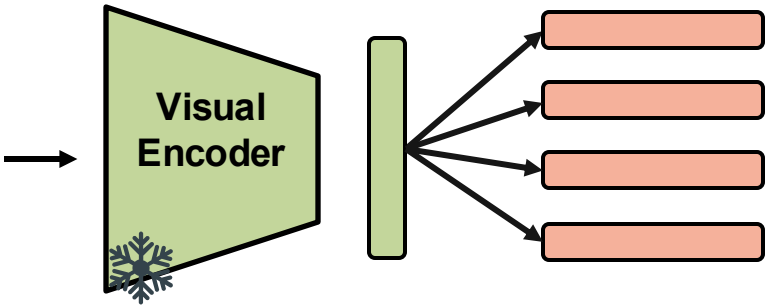
“A prostate histology image of [CLS] tissue.”



$$f_\phi([\text{Description}]_\mathbf{c}) = \mathbf{t_c}$$

$$\langle \mathbf{v}, \mathbf{w}_c \rangle$$





Non-cancer	0.05
Grade 3	0.16
Grade 4	**0.55**
Grade 5	0.24

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

$$p_c(\mathbf{W}) = \frac{\exp(\mathbf{v}^\top \mathbf{w}_c/\tau)}{\sum_{j=1}^{C} \exp(\mathbf{v}^\top \mathbf{w}_j/\tau)}$$

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

Few-shot adaptation via linear probing

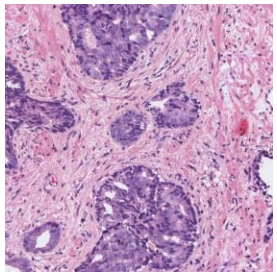
Support set

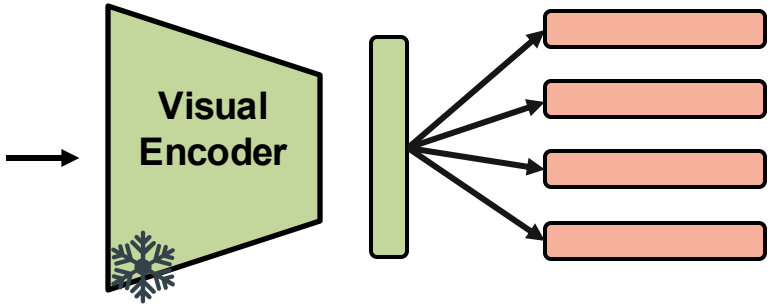
$\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic} \ln(p_{ic}(\mathbf{W}))$$

$N = C \times K; K \in \{1, 2, 4, 8, 16\}$

$\langle \mathbf{v}, \mathbf{w}_c \rangle$





Non-cancer
Grade 3
Grade 4
Grade 5

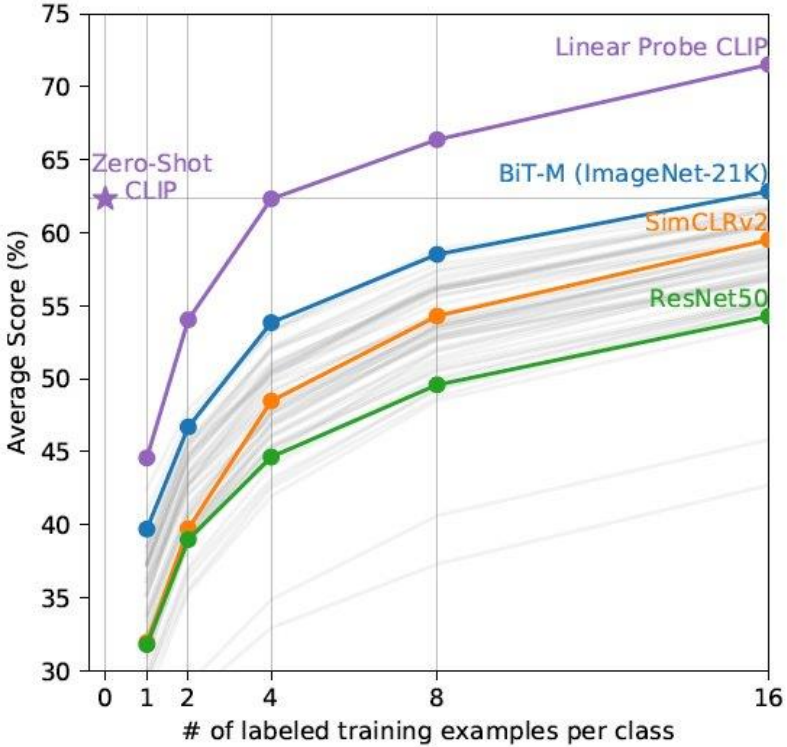
0.05
0.16
0.55
0.24

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

$$p_c(\mathbf{W}) = \frac{\exp(\mathbf{v}^\top \mathbf{w}_c/\tau)}{\sum_{j=1}^{C}\exp(\mathbf{v}^\top \mathbf{w}_j/\tau)}$$

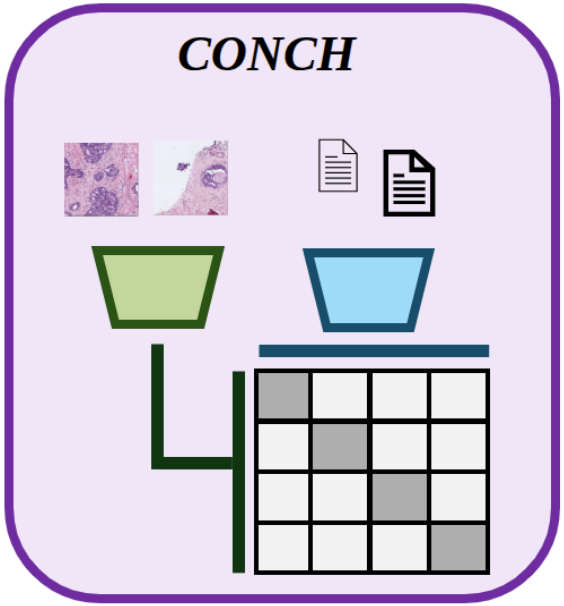
Radford et *al*. Learning transferable visual models from natural language supervision, **ICML'21**

Few-shot adaptation via linear probing

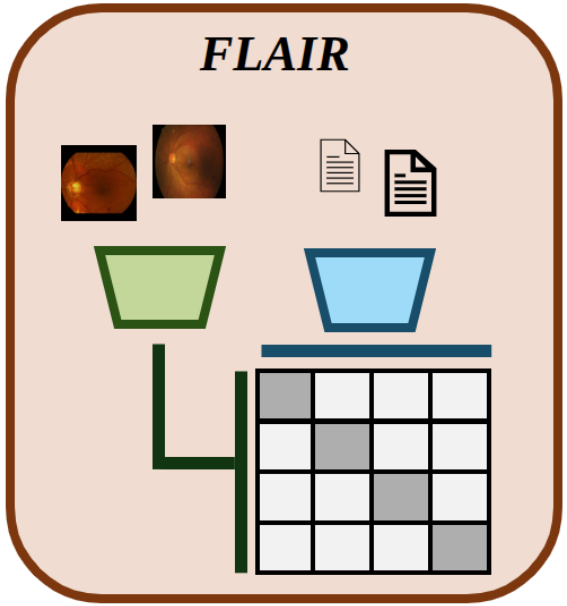


Radford et *al*. Learning transferable visual models from natural language supervision, **ICML'21**

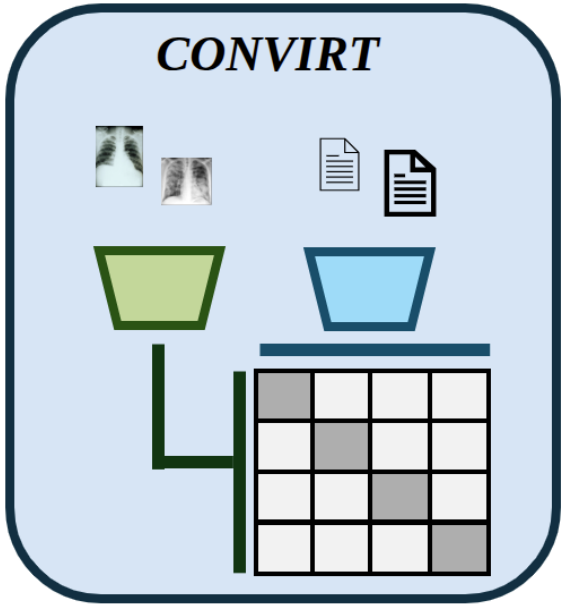
Medical Vision-Language Models



Histology

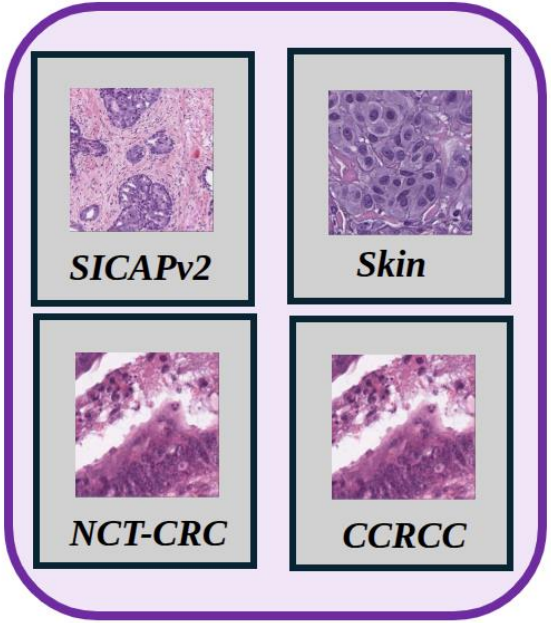


Ophtalmology

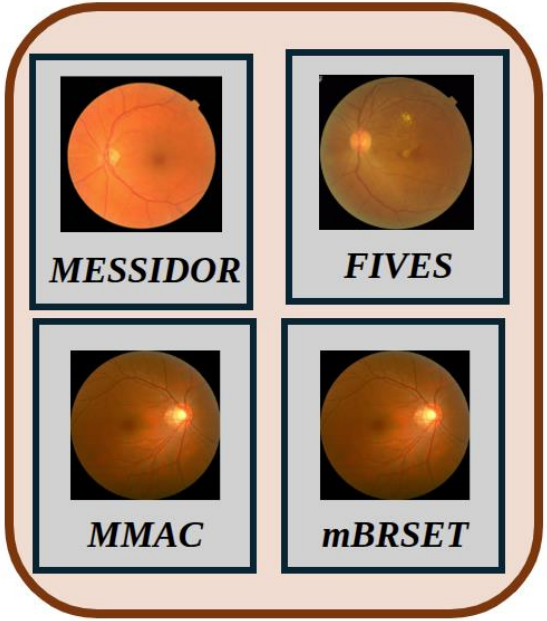


Chest X-ray

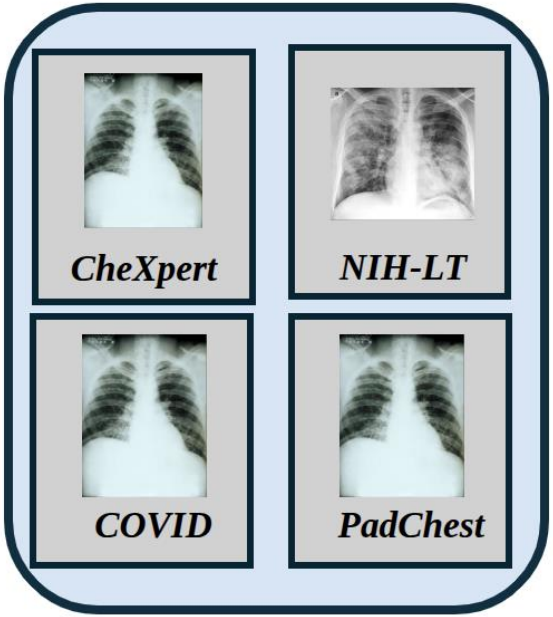
Medical Vision-Language Models



Histology



Ophthalmology



Chest X-ray

Text-informed linear probes.

Few-shot learning objective:

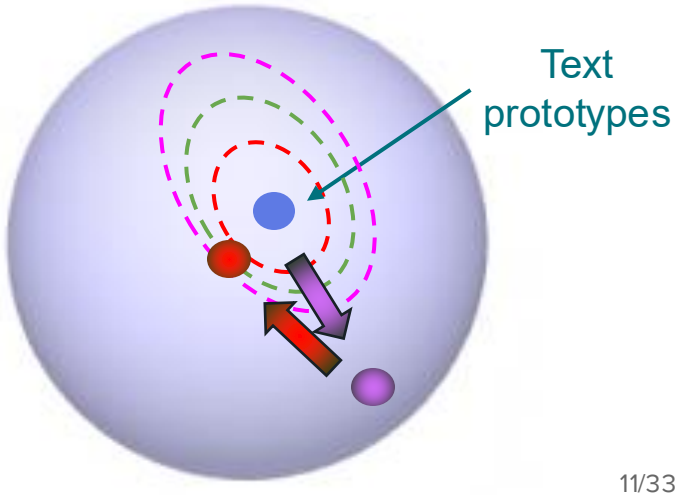
$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathcal{H}(\mathbf{y}_i, \mathbf{p}_i(\mathbf{W})) + \sum_{c \in \mathcal{Y}} \lambda_c^{\mathrm{T}} ||\mathbf{w}_c - \mathbf{t}_c||_2^2$$

Cross-entropy loss:

$$\mathcal{H}(\mathbf{y}_i, \mathbf{p}_i(\mathbf{W})) = -\sum_{c \in \mathcal{Y}} y_{i,c} \ln(p_{i,c}(\mathbf{W}))$$

SoftMax similarity:

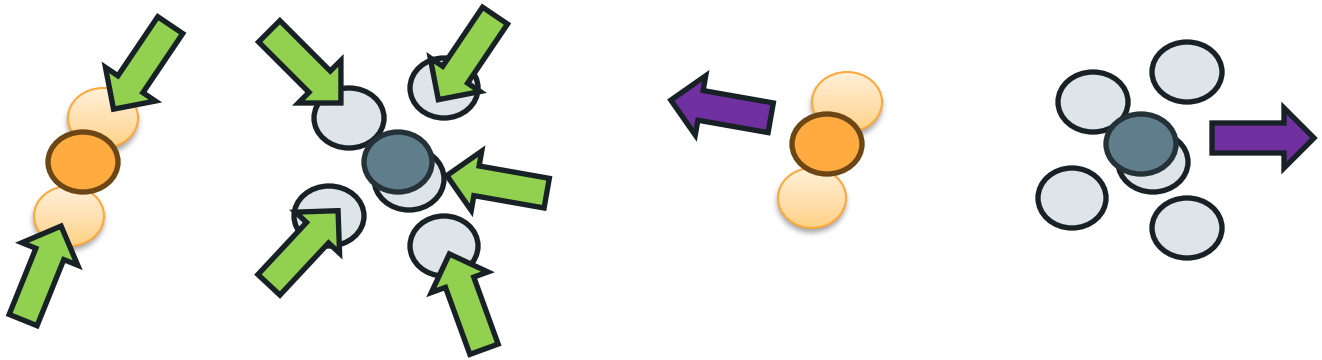
$$p_{i,c}(\mathbf{W}) = \frac{\exp(\mathbf{v}_i^\top \mathbf{w}_c / \tau)}{\sum_{j \in \mathcal{Y}} \exp(\mathbf{v}^\top \mathbf{w}_j / \tau)}$$



Text-informed linear probes.

A closer look to cross-entropy loss:

$$\mathcal{H}(\mathbf{y}_i, \mathbf{p}_i(\mathbf{W})) = \underbrace{-\sum_{c\in\mathcal{Y}} y_{i,c}\,(\mathbf{v}_i^\top \mathbf{w}_c/\tau)}_{\mathcal{H}^t(\mathbf{y}_i,\mathbf{W})} + \underbrace{\sum_{c\in\mathcal{Y}} y_{i,c}\,\ln(\sum_{j\in\mathcal{Y}} \exp(\mathbf{v}_i^\top \mathbf{w}_j/\tau))}_{\mathcal{H}^{\text{cont}}(\mathbf{W})}$$



Silva-Rodriguez *et al*. Few-Shot, Now for Real: Medical VLMs Adaptation without Balanced Sets or Validation, **MICCAI'25**

Text-informed linear probes.

Constrained linear probe approximation with closed-form solution:

$$\min_{\mathbf{W}} \mathcal{L}_{\text{FEW-SHOT}}(\mathbf{W}) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathcal{H}^t(\mathbf{y}_i, \mathbf{W}) + \sum_{c \in \mathcal{Y}} \lambda_c^{\mathrm{T}} ||\mathbf{w}_c - \mathbf{t}_c||_2^2$$

Text-informed linear probes.

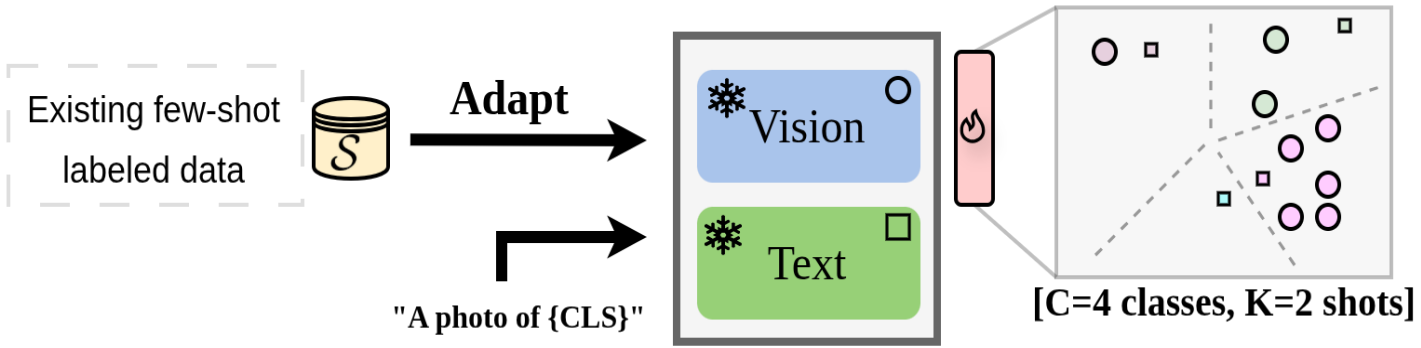
Constrained linear probe approximation with closed-form solution:

$$\min_{\mathbf{W}} \mathcal{L}_{\text{FEW-SHOT}}(\mathbf{W}) = -\frac{1}{|\mathcal{S}|}\sum_{c\in\mathcal{Y}} y_{i,c}\,(\mathbf{v}_i^\top \mathbf{w}_c/\tau) + \sum_{c\in\mathcal{Y}} \lambda_c^{\mathrm{T}} ||\mathbf{w}_c - \mathbf{t}_c||_2^2$$

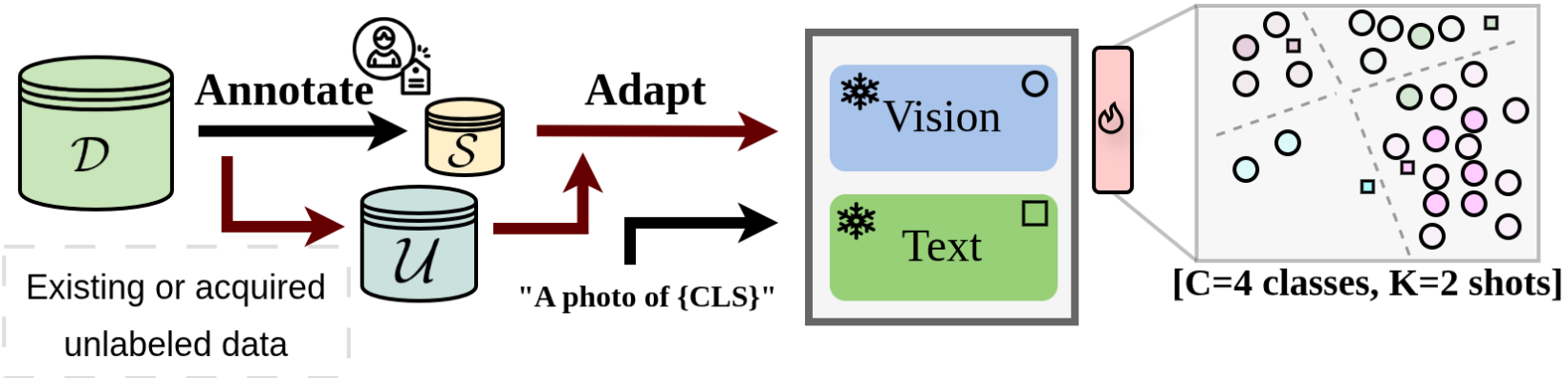
$$\mathbf{w}_c^t = \arg\min_{\mathbf{w}_c} \frac{\partial \mathcal{L}_{\text{FEW-SHOT}}(\mathbf{W})}{\partial \mathbf{w}_c} = \frac{1}{2\lambda_c^{\mathrm{T}}\,|\mathcal{S}|\,\tau}\sum_{i\in\mathcal{S}} y_{i,c}\mathbf{v}_i + \mathbf{t}_c$$

Visual **Textual**

Semi-supervision: making the most of unlabelled data



Semi-supervision: making the most of unlabelled data



Learning from unlabelled data

Combining a few-shot support set with unlabeled data:

$$\mathcal{D} = \{i \in \mathbb{N} : 1 \leq i \leq N + M\} = \mathcal{S} \cup \mathcal{U}$$

$$\{(\mathbf{v}_i, \mathbf{y}_i)\}_{i \in \mathcal{S}} \qquad \{(\mathbf{v}_i)\}_{i \in \mathcal{U}}$$

Learning from unlabelled data

Incorporating supervisory signals trough unlabeled data:

$$\mathcal{L}_{\mathrm{U}}(\mathbf{W}, \mathbf{z}) = \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} \mathcal{H}^t(\mathbf{z}_i, \mathbf{W}) \text{ s.t. } \hat{\mathbf{m}} = \mathbf{m}$$

pseudo-labels
(treated as learnable variables)

$$\mathbf{z}_i = (z_{i,k})_{(1 \leq c \leq C)} \in \triangle_C, i \in \mathcal{U}$$

Learning from unlabelled data

Incorporating supervisory signals trough unlabeled data:

$$\mathcal{L}_{\mathrm{U}}(\mathbf{W}, \mathbf{z}) = \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} \mathcal{H}^t(\mathbf{z}_i, \mathbf{W}) \text{ s.t. } \hat{\mathbf{m}} = \mathbf{m}$$

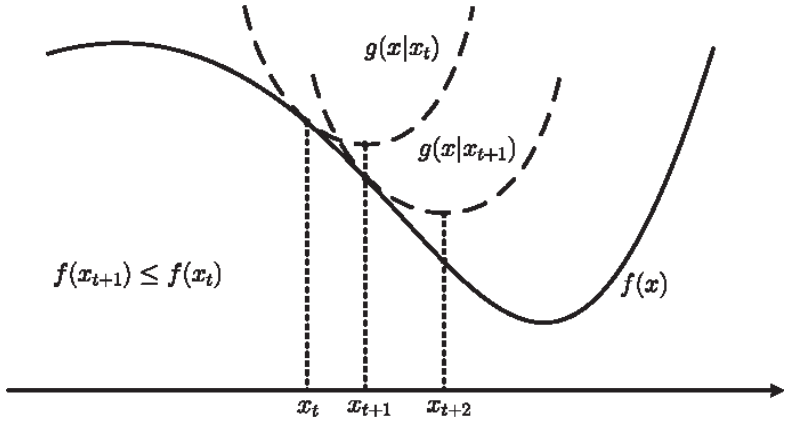
$$\hat{\mathbf{m}} = \sum_{i \in \mathcal{U}} \mathbf{z}_i / \sum_{i \in \mathcal{U}, c \in \mathcal{Y}} z_{i,c} \qquad \mathbf{m} = \sum_{i \in \mathcal{S}} \mathbf{y}_i / N$$

SS-Text-U: Objective & block-wise optimization

$$\min_{\mathbf{W},\mathbf{z}} \mathcal{L}_{\text{SEMI}}(\mathbf{W},\mathbf{z}) = \mathcal{L}_{\text{FEW-SHOT}}(\mathbf{W}) + \lambda^{\text{U}} \mathcal{L}_{\text{U}}(\mathbf{W},\mathbf{z})$$

Block-wise optimizer.
For t in T:

1. **Freeze W, argmin Z wrt Z-block.**
2. **Freeze Z, argmin W wrt W-block.**



Lange et al. Optimization transfer using surrogate objective functions (2000).
Razaviyayn et al. A unified convergence analysis of block successive minimization methods for non-smooth optimization (2013).
Wright et al. Coordinate descent algorithms (2015).
Figure from Sun *et al*. Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning (2017).

SS-Text-U: Objective & block-wise optimization

1. Z-block updates.

$$\min_{\mathbf{z}} \ \mathcal{L}_{\mathrm{U}}(\mathbf{W}^{t-1}, \mathbf{z}) = -\frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} \sum_{c \in \mathcal{Y}} z_{i,c} (\mathbf{v}_i^\top \mathbf{w}_c^{t-1} / \tau) \ \text{s.t.} \ \hat{\mathbf{m}} = \mathbf{m}$$

$$(\mathbf{w}_c^0 = \mathbf{t}_c)$$

SS-Text-U: Objective & block-wise optimization

1. Z-block updates.

$$\min_{\mathbf{z}} \ \mathcal{L}_{\mathrm{U}}(\mathbf{W}^{t-1}, \mathbf{z}) = -\frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} \sum_{c \in \mathcal{Y}} z_{i,c} (\mathbf{v}_i^\top \mathbf{w}_c^{t-1} / \tau) \ \text{s.t.} \ \hat{\mathbf{m}} = \mathbf{m}$$

$(\mathbf{w}_c^0 = \mathbf{t}_c)$

(in the matrix form)

$$\mathbf{S} = ((\mathbf{v}_i^\top \mathbf{w}_c^{t-1} / \tau)_{c \in \mathcal{Y}})_{i \in \mathcal{U}} \qquad \mathbf{Q} = (\mathbf{z}_i)_{i \in \mathcal{U}}$$

SS-Text-U: Objective & block-wise optimization

1. Z-block updates.

$$\min_{\mathbf{z}} \ \mathcal{L}_{\mathrm{U}}(\mathbf{W}^{t-1}, \mathbf{z}) = -\frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} \sum_{c \in \mathcal{Y}} z_{i,c} (\mathbf{v}_i^\top \mathbf{w}_c^{t-1} / \tau) \ \text{s.t.} \ \hat{\mathbf{m}} = \mathbf{m}$$

$(\mathbf{w}_c^0 = \mathbf{t}_c)$

(in the matrix form)

$$\mathbf{S} = ((\mathbf{v}_i^\top \mathbf{w}_c^{t-1} / \tau)_{c \in \mathcal{Y}})_{i \in \mathcal{U}} \qquad \mathbf{Q} = (\mathbf{z}_i)_{i \in \mathcal{U}}$$

The optimization can be addressed from an O*ptimal Transport* perspective:

$$\max_{\mathbf{Q} \in \mathcal{Q}} \ tr(\mathbf{Q}^\top \mathbf{S}), \text{s.t.} \ \hat{\mathbf{m}} = \mathbf{m}$$

SS-Text-U: Objective & block-wise optimization

Iterative solution via the Sinkhorn-Knopp algorithm.

$$\mathbf{Q}^* = \mathrm{Diag}(\mathbf{r}^{(T_{\mathrm{OT}})})\mathbf{Q}^{(0)}\mathrm{Diag}(\mathbf{c}^{(T_{\mathrm{OT}})})$$

$$\mathbf{r}^{(t_{\mathrm{OT}})} = \mathbf{m}/(\mathbf{Q}^{(0)}\mathbf{c}^{(t_{\mathrm{OT}}-1)}) \qquad \mathbf{c}^{(t_{\mathrm{OT}})} = \mathbf{u}_{(|\mathcal{U}|)}/(\mathbf{Q}^{(0)}\mathbf{r}^{(t_{\mathrm{OT}})})$$

$\mathbf{Q}^{(0)} = \exp(\mathbf{S})/\sum(\exp(\mathbf{S}))$ (batch-level normalized logits for initialization)

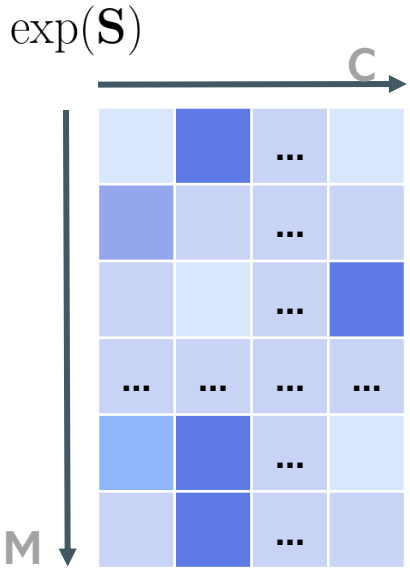
$\mathbf{c}^{(0)} = \mathbf{1}_{(|\mathcal{U}|)}$ (normalization initialization – initially, all samples weight the same)

$\mathbf{u}_{(|\mathcal{U}|)}$ (target uniform sample-marginal distribution)

Cuturi *et al*. Sinkhorn distances: Lightspeed computation of optimal transport. **NeurIPS'13**

SS-Text-U: Objective & block-wise optimization

Sinkhorn-Knopp algorithm.



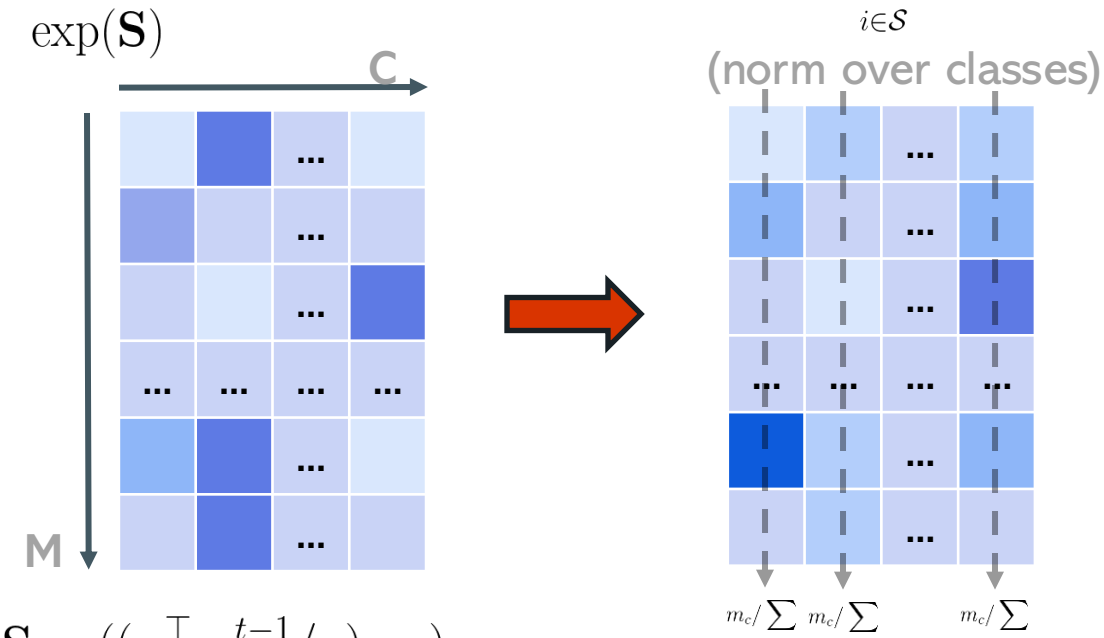
$$\mathbf{S} = ((\mathbf{v}_i^\top \mathbf{w}_c^{t-1}/\tau)_{c\in\mathcal{Y}})_{i\in\mathcal{U}}$$

Cuturi *et al*. Sinkhorn distances: Lightspeed computation of optimal transport. **NeurIPS'13**

SS-Text-U: Objective & block-wise optimization

Sinkhorn-Knopp algorithm.

$$\mathbf{m} = \sum_{i \in \mathcal{S}} \mathbf{y}_i / N$$

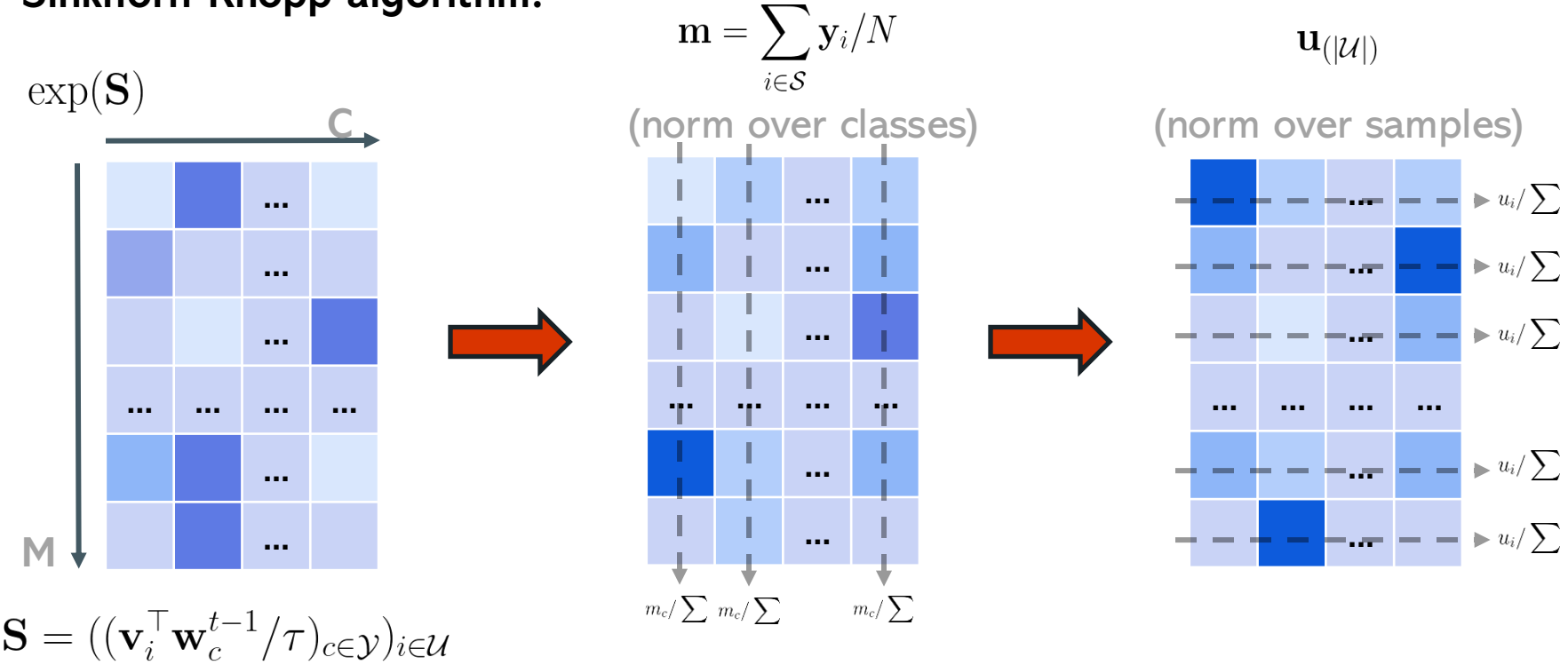


$$\mathbf{S} = ((\mathbf{v}_i^\top \mathbf{w}_c^{t-1}/\tau)_{c \in \mathcal{Y}})_{i \in \mathcal{U}}$$

Cuturi *et al*. Sinkhorn distances: Lightspeed computation of optimal transport. **NeurIPS'13**

SS-Text-U: Objective & block-wise optimization

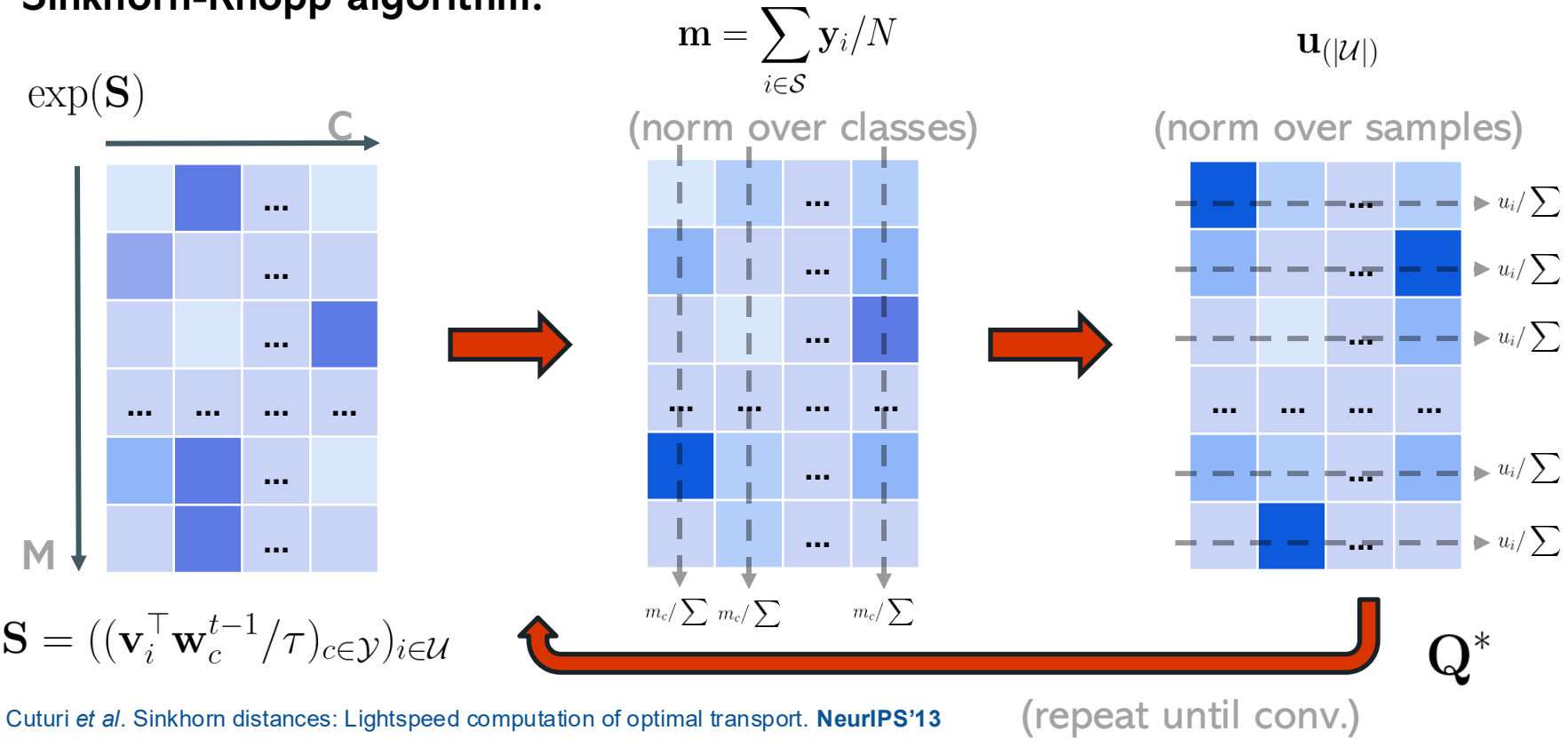
Sinkhorn-Knopp algorithm.



Cuturi *et al*. Sinkhorn distances: Lightspeed computation of optimal transport. **NeurIPS'13**

SS-Text-U: Objective & block-wise optimization

Sinkhorn-Knopp algorithm.



Cuturi *et al*. Sinkhorn distances: Lightspeed computation of optimal transport. **NeurIPS'13**

SS-Text-U: Objective & block-wise optimization

2. Closed-form solution for the W-block updates.

$$\mathbf{w}_c^t = \arg\min_{\mathbf{w}_c} \frac{\partial \mathcal{L}_{\text{SEMI}}(\mathbf{W}, \mathbf{z}^t)}{\partial \mathbf{w}_c} = \frac{1}{2\lambda_c^{\text{T}} |\mathcal{S}| \tau} \sum_{i \in \mathcal{S}} y_{i,c} \mathbf{v}_i + \frac{\lambda_c^{\text{U}}}{2\lambda_c^{\text{T}} |\mathcal{U}| \tau} \sum_{i \in \mathcal{U}} z_{i,c}^t \mathbf{v}_i + \mathbf{t}_c$$

SS-Text-U: Objective & block-wise optimization

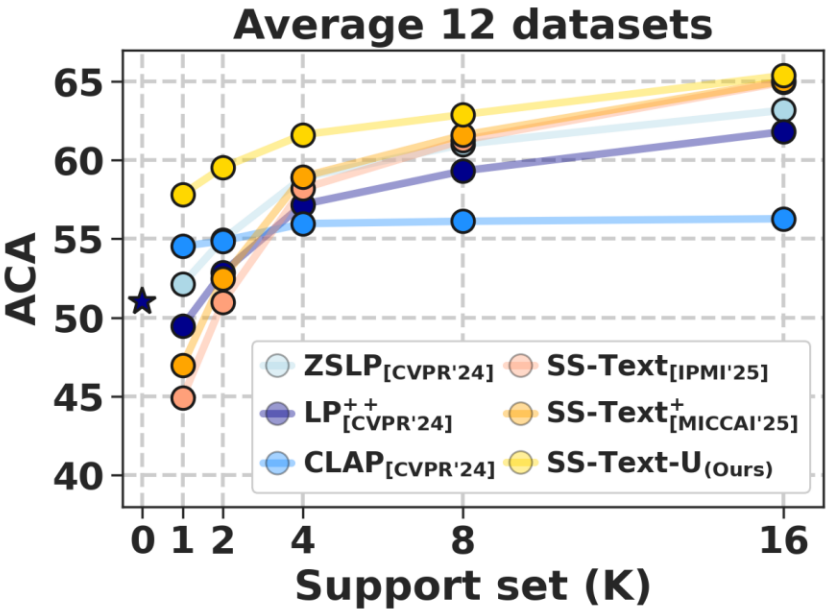
2. Closed-form solution for the W-block updates.

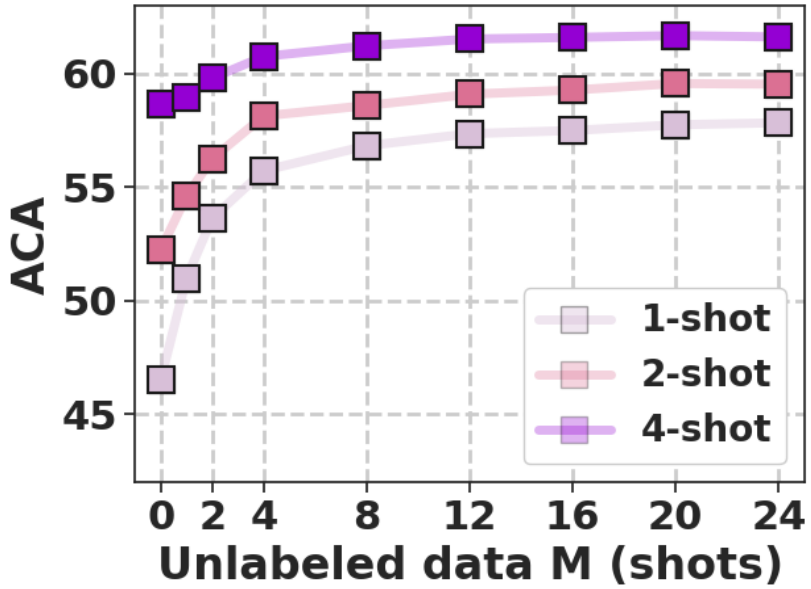
$$\mathbf{w}_c^t = \arg\min_{\mathbf{w}_c} \frac{\partial \mathcal{L}_{\text{SEMI}}(\mathbf{W}, \mathbf{z}^t)}{\partial \mathbf{w}_c} = \frac{1}{2\lambda_c^{\text{T}} |\mathcal{S}| \tau} \sum_{i \in \mathcal{S}} y_{i,c} \mathbf{v}_i + \frac{\lambda_c^{\text{U}}}{2\lambda_c^{\text{T}} |\mathcal{U}| \tau} \sum_{i \in \mathcal{U}} z_{i,c}^t \mathbf{v}_i + \mathbf{t}_c$$

(data-driven hyper-parameters)

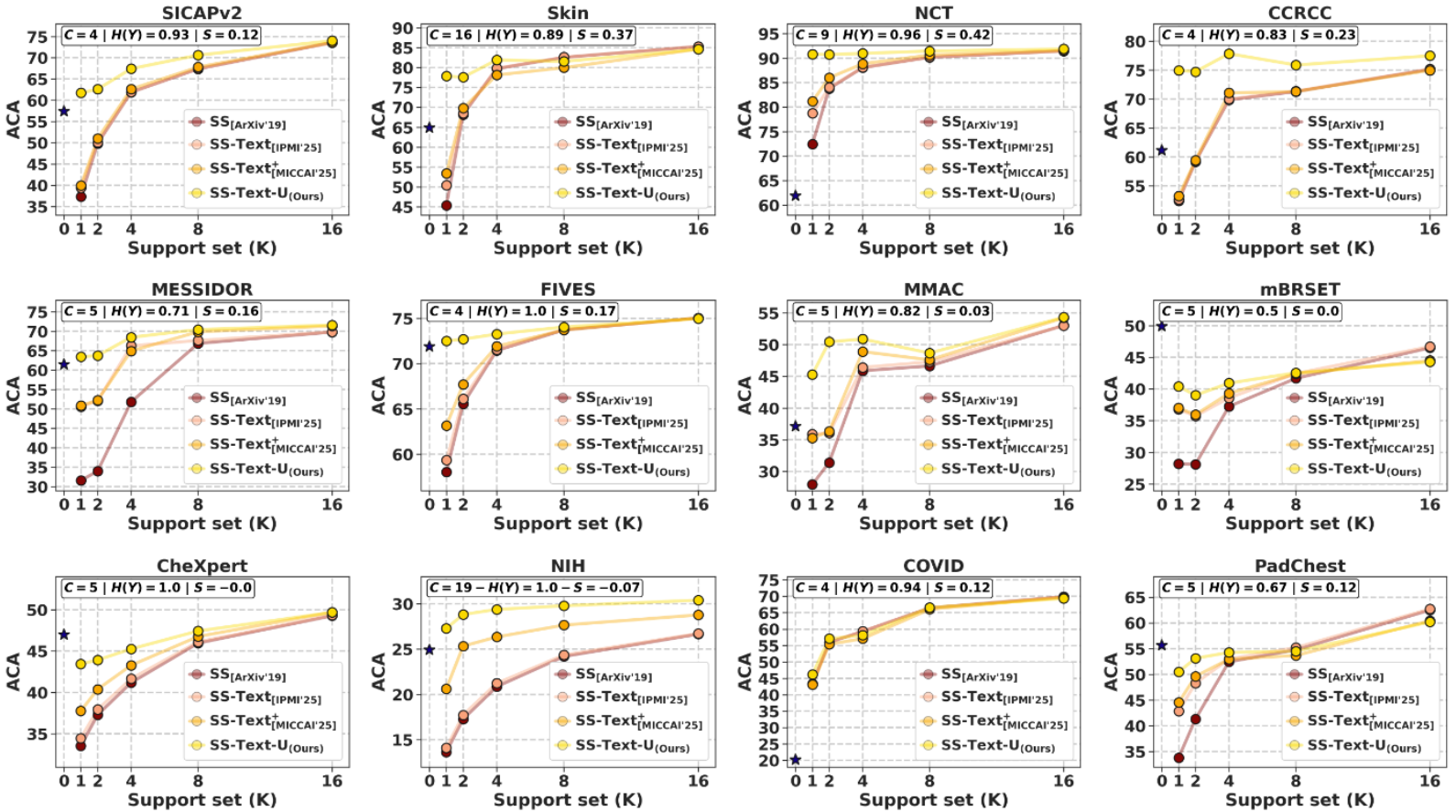
$$\lambda_c^{\text{T}} = (1/K_c) = (1/\sum_{i \in \mathcal{S}} y_{i,c}) \; ; \; \lambda_c^{\text{U}} = 2\lambda_c^{\text{T}}$$

Results & exploratory analysis





Results & exploratory analysis



Results & exploratory analysis