Learning from distributed and heterogeneous data

Li Ju

Division of Scientific Computing
Department of Information Technology
Uppsala University

August 16, 2025

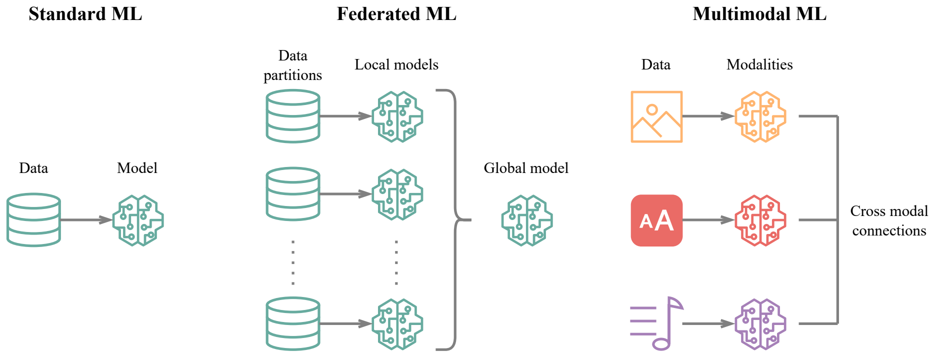
Machine Learning (ML) "...the development and study of statistical algorithms that can learn from data and generalize to unseen data...", from Wikipedia

Data are by nature distributed:

- generated from diverse sources (social media, IoT devices...)
- infeasible to be collected together (cost, legal restrictions, different formats...)

Data are also inherently heterogeneous, including:

- Heterogeneity across data partitions.
- Inherent heterogeneity across data of different formats.



Algorithms need to adapt to the distributed and heterogeneous nature of data.

My work focuses on two aspects:

- Federated learning: Learning from distributed data.
- Vision language models: Learning from data of different formats.

But what is FL?

Classification problem

We are interested in a classifier $\hat{y} = f(\hat{x}; \theta), \theta \in \Theta$.

Given a dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$, with $y = f(x; \theta)$ and $\ell(\cdot, \cdot)$, we aim to solve the optimisation problem

$$\theta^\star = \arg\min_{\theta \in \Theta} \frac{1}{N} \sum_{n=1}^{N} \ell(f(x_n; \theta), y_n).$$

Generally solved with stochastic first-order methods.

But what is FL?

Federated Learning

The dataset is $\{\mathcal{D}_k\}_{k=1}^K$, where $\mathcal{D}_k = \{(x_n, y_n)\}_{n=1}^{N_k}, \forall k \in [K]$. Then the optimisation problem is in the form of

$$\theta^\star = \arg\min_{\theta\in\Theta} \frac{1}{K}\sum_{k=1}^{K} R_k(\theta),$$

where $R_k(\theta) = \sum_{n=1}^{N_k} \ell(f(x_n;\theta), y_n)$.

How to solve this problem efficiently, w.r.t. the distributed data access pattern? Baseline algorithm: `FedAvg`.

Generalised framework

Algorithm 1 `FedOpt`[1]

Require: Initialize parameters θ^0
  for round t in $\{1, \dots T\}$ **do**
    for client k in $\{1, \dots K\}$ **parallel do**
      $\theta_k^t = \text{ClientOpt}(\theta^{t-1})$ ▷ Client-side
      $\Delta_k^t \theta := \theta_k^t - \theta^{t-1}$
    end for
    $\Delta^t \theta = \text{Aggre}(\{\Delta_k^t \theta, 0 \leq k < K\})$ ▷ Server-side
    $\theta^{t+1} = \text{ServerOpt}(\Delta^t \theta)$
  end for

- `FedAvg`: *SGD* + Averaging + *GD*.
- `FedAdam`: *SGD* + Averaging + *Adam*.

[1] Reddi et al., "Adaptive federated optimization".

Federated learning for predicting compound mechanism of action based on image-data from cell painting[2]

[2] Ju, Hellander, and Spjuth, "Federated learning for predicting compound mechanism of action based on image-data from cell painting".

Questions of interest

An image classification problem:

- Fluorescence image X: $H \times W \times \#\text{channels}$.
- MoA Y: Categorical variable.
- Model: a classifier $\hat{y} = f(\hat{x}; \theta)$.

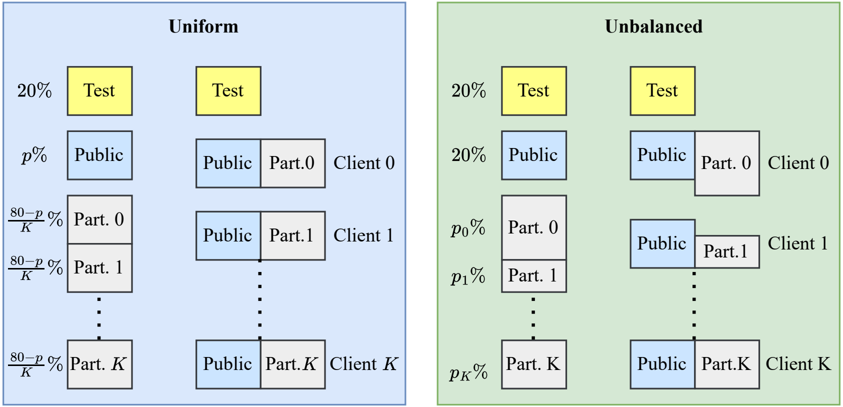
In pharmaceutical industry, collaborative ML without sharing data is necessary. FL is the option!

In the context of MoA prediction, we are interested in

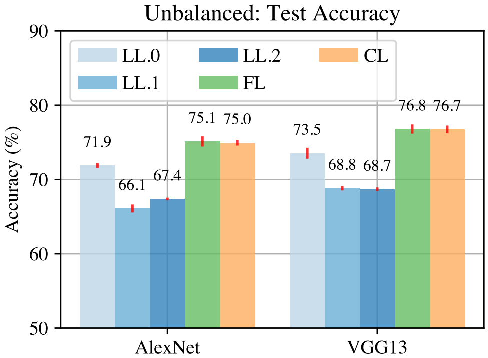
- the effectiveness of FL.
- how data heterogeneity affects the performance.

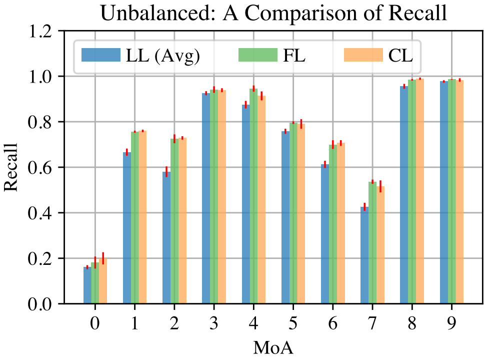
Scenarios

We simulate three scenarios, Uniform, Unbalanced (in sizes), and Non-IID (specialisation in certain MoAs).



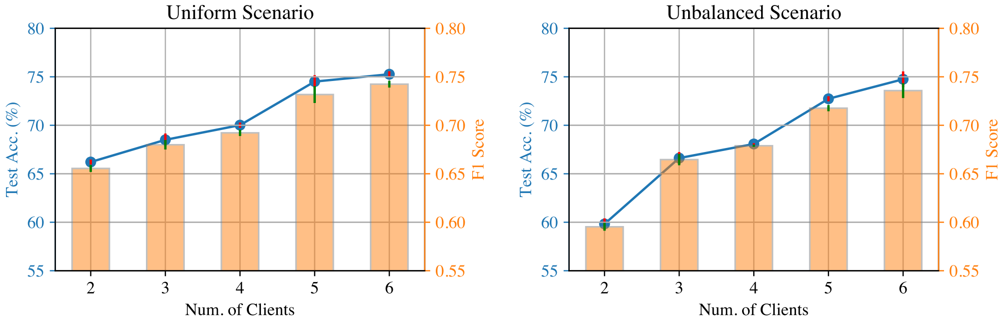
CL $\approx$ FL $>$ LL



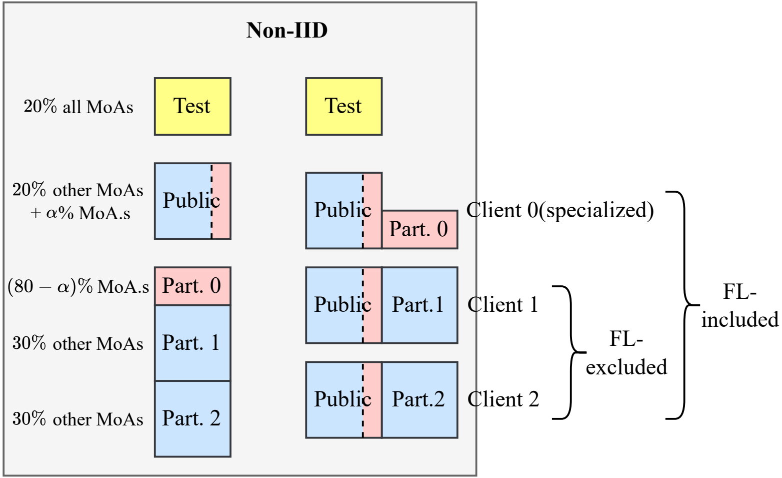


This encourages collaboration across pharm entities using FL, instead of training local models.

The more participants, the better performance

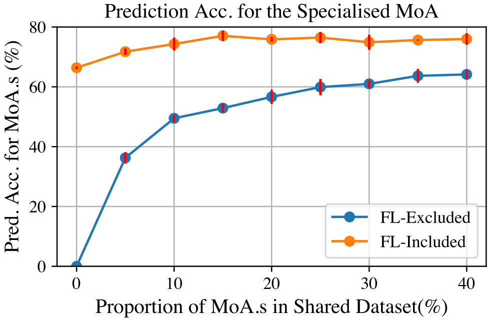


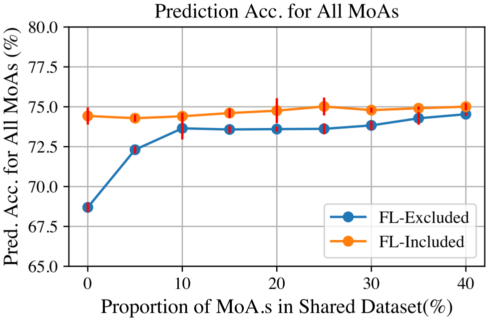
This encourages existing participants to keep engaging in FL throughout the life cycle of a model.



We compare the performance of the federated models with the specialised client included and excluded.

Specialised participant brings benefits





Including the specialised client in federated learning

- significantly improves the prediction accuracy for the specialised MoA.
- slightly improves the average prediction accuracy for all MoAs.

This encourages both specialised and general clients to join federated learning.

We conclude that

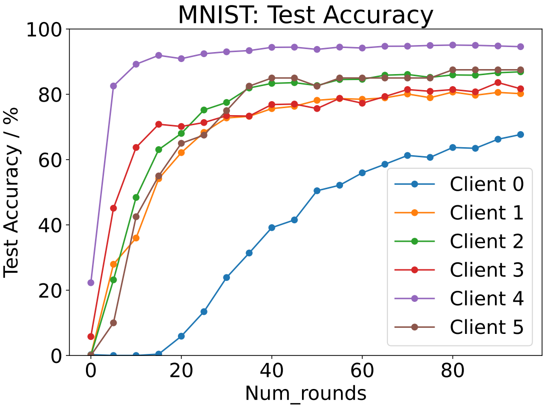
- Federated learning does bring benefits for MoA prediction.
- Our studies provide motivations for different (potential) participants.
- Theoretical studies for data heterogeneity are too pessimistic in the context of MoA prediction.

Accelerating fair federated learning: Adaptive federated adam[3]

[3] Ju, Zhang, et al., "Accelerating Fair Federated Learning: Adaptive Federated Adam".

Fairness problem?

If clients own their own local test sets (instead of a global test set):



Fairness problem: the discrepancy in model performance across clients in FL.

Q-Fair FL

Standard FL

$$\theta^\star = \arg\min_\theta \sum_{k=1}^{K} R_k(\theta)$$

Q-Fair FL

$$\theta^\star = \arg\min_\theta \sum_{k=1}^{K} R_k^{q+1}(\theta)$$

where $q \geq 0$ is a hyperparameter. A commonly used approach in resource allocation, with q-fairness guarantee.

The update rule and the gradient are given by:

$$\theta^{t+1} := \theta^t + \eta_t \cdot \nabla_\theta \sum_{k=1}^{K} R_k^{q+1}(\theta^t)$$

$$\nabla_\theta \sum_{k=1}^{K} R_k^{q+1}(\theta^t) = (q+1) \sum_{k=1}^{K} R_k^{q}(\theta^t) \cdot \nabla R_k(\theta^t)$$

Diminishing gradient scales require adaptive η_t to make progress!

Tian[4] proposed an adaptive method, which is

- Effective
- But slow (2-5 times slower compared to `FedAvg`)
- And not compatible with `FedOpt`.

[4] Li et al., "Fair resource allocation in federated learning".

We want FL to be both fair and fast.

Problems include:

- The diminishing gradient scales
  - Reformulation is required.
- Poor use of `FedOpt`.
  - Study of the server-side optimiser for better convergence.

We propose a new formulation

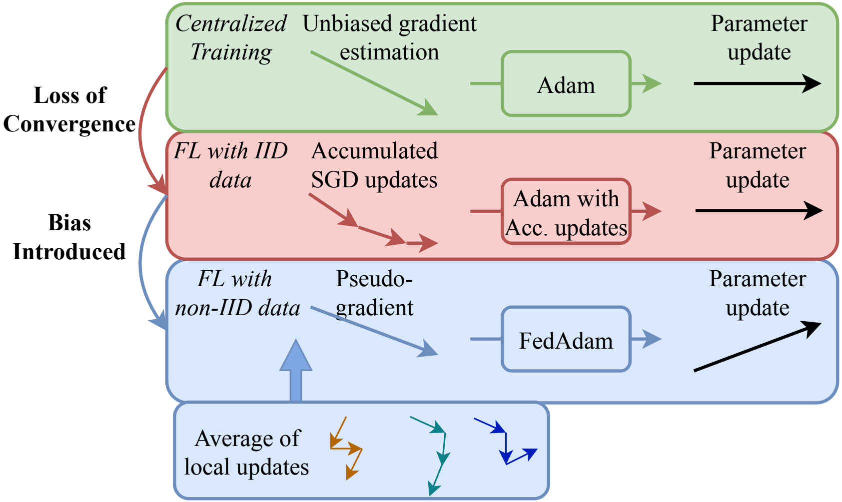
$$\theta^\star = \arg\min_\theta \frac{\sum_{k=1}^K I_k^\alpha(t) \cdot R_k(\theta^t)}{\sum_{k=1}^K I_k^\alpha(t)}$$

where $I_k(t) := R_k(\theta^t)/R_k(\theta^0)$ and $\alpha \geq 0$ is similar to q in Q-fair FL.

Our formulation has two properties:

- Shares the same stationary points with Q-fair FL, thus with the identical fairness guarantee.
- Gets rid of the problem of diminishing gradient scales, thus compatible with `FedOpt`.

To further accelerate the optimisation, we study `Adam` in heterogeneous FL.



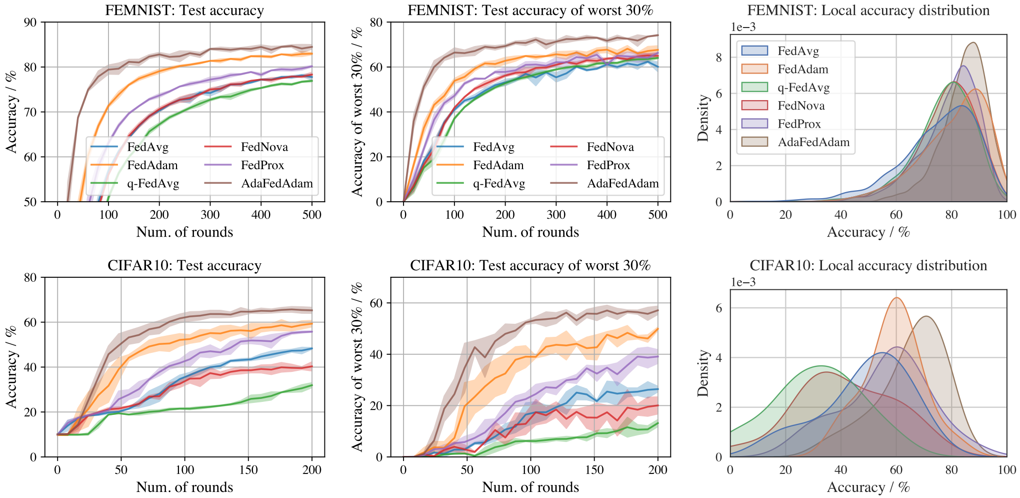
Our method

Tackling the problem of `FedAdam`, we propose our method, Adaptive Federated Adam:

Algorithm 2 `AdaFedAdam`

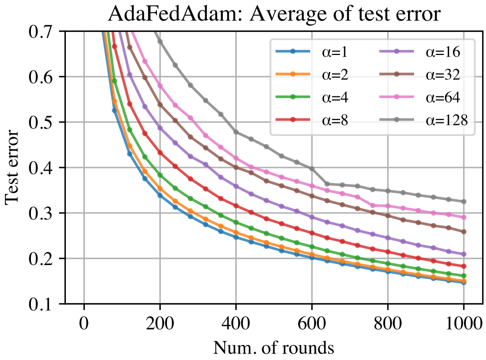
Require: Initialize parameters $\boldsymbol{\theta}^0$
  for round t in $\{1, \ldots T\}$ **do**
    for client k in $\{1, \ldots K\}$ **parallel do**
      $\boldsymbol{\theta}_k^t = \text{ClientOpt}(\boldsymbol{\theta}^{t-1})$ ▷ Client-side
      $\Delta_k^t \boldsymbol{\theta} := \boldsymbol{\theta}_k^t - \boldsymbol{\theta}^{t-1}$
      $\Delta_k^t \boldsymbol{\theta} = \eta_k^t \cdot \boldsymbol{U}_k^t$ s.t. $\|\boldsymbol{U}_k^t\|_2 = \|\nabla_{\boldsymbol{\theta}} R_k(\boldsymbol{\theta}^t)\|_2$ (step size $\times$ direction)
    end for
    $\eta^t, \beta_1^t, \beta_2^t = \text{Aggre. hyperpara.}(\{\eta_k^t) : 0 \leq k < K\}$ ▷ Server-side
    $\Delta^t \boldsymbol{\theta} = \text{Aggre. direction}(\{\boldsymbol{U}_k^t : 0 \leq k < K\})$
    $\boldsymbol{\theta}^{t+1} := \texttt{Adam}(\Delta^t \boldsymbol{\theta}; \eta^t, \beta_1^t, \beta_2^t)$
  end for

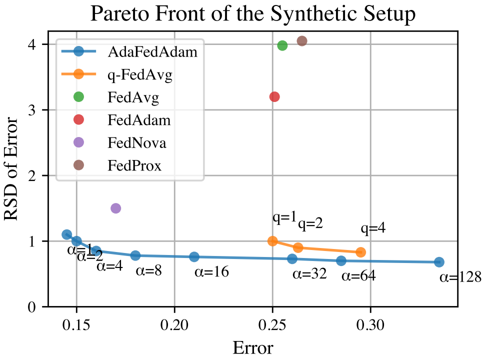
Empirical results: convergence and fairness



Empirical results: the Pareto front

How does the additional hyper-parameter α affect the performance?





Key Properties

Our approach ensures following properties:

- Fairness guarantee: Identical to Q-fair FL.
- Improved convergence rate.
- Fine-tuning free: Adaptivity of hyper-parameters.
- Others: allowance for resource heterogeneity, robustness, compatibility with arbitrary local solvers, etc.

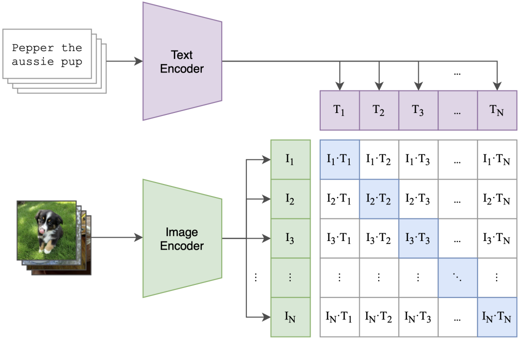
Exploiting the asymmetric uncertainty structure of pre-trained vision-language models on the unit hypersphere[5]

[5] Ju, Andersson, et al., "Exploiting the Asymmetric Uncertainty Structure of Pre-trained VLMs on the Unit Hypersphere".

What is pre-trained VLMs?

"VLMs learn to map relationships between textual and visual data, in which image and text embeddings reside in a joint vector space."

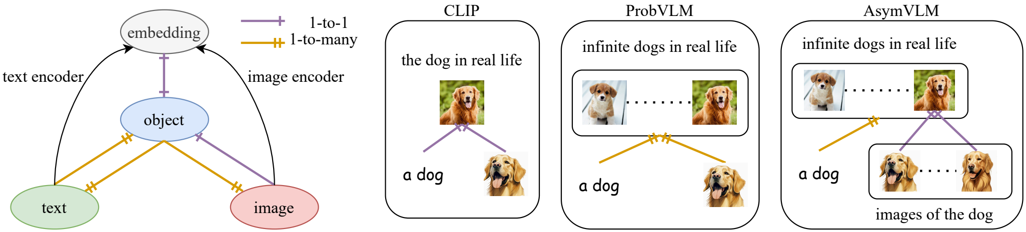
Contrastive Language Image Pre-training (CLIP)[6]



[6] Radford et al., "Learning transferable visual models from natural language supervision".

Rethinking Building VLMs

- CLIP: "Image–text is an one-to-one mapping".
- ProbVLM[7]: "Image–text is a (symmetric) many-to-many mapping".
- AsymVLM: "Image–text is a many-to-many mapping with an asymmetric structure."



[7] Upadhyay et al., "Probvlm: Probabilistic adapter for frozen vison-language models".

Building the method

- Text encoder (text $\rightarrow$ embedding): one-to-many, modelled by probabilistic embeddings.
- Image encoder (image $\rightarrow$ embedding): one-to-one, modelled by deterministic embedding.

Additionally, we need to utlize the pre-trained models (CLIP, BLIP, SigLIP, etc), which has deterministic embeddings on $\mathbb{S}^{d-1}$:

- The method should be post-hoc.
- Probabilistic embeddings should be modelled by directional distributions.

Deriving the Loss

Formally, the embedding of any text $t \in \mathcal{T}$ is modeled by a random variable $\boldsymbol{z}^T$,

$$\boldsymbol{z}^T \sim P(\theta(t)) \text{ where } \theta(t) := g_T \circ f_T(t),$$

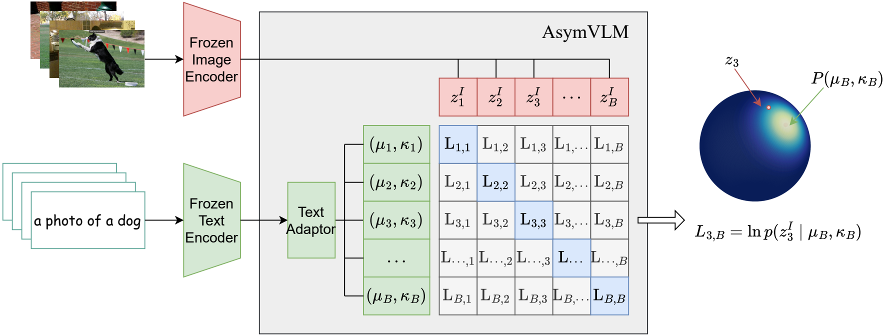
g_T denote the adaptor and f_T denote the pre-trained text encoder.

The embedding of any image $i \in \mathcal{I}$ is given by $z^I = f_I(i)$, where f_I denotes the pre-trained image encoder.

We choose von Mises Fisher distribution (vMF) and Power Spherical distribution (PS) for probabilistic embeddings.

Deriving the Loss

We want to maximize $p(z^I(i) \mid \theta(t))$ if t and i match, and minimize it if they do not:



To maximize the diagonals and minimize the off-diagonals, InfoNCE loss is applied.

Discussion

Unified objectives:

$$\theta = \arg\min_{\theta \in \Theta} -\frac{1}{2B}\sum_{n=1}^{B}\left[\ln \frac{\exp\left(\tau\delta(n,n)\right)}{\sum_{m=1}^{B}\exp\left(\tau \ln \delta(n,m)\right)} + \ln \frac{\exp\left(\tau\delta(n,n)\right)}{\sum_{m=1}^{B}\exp\left(\tau\delta(m,n)\right)}\right].$$

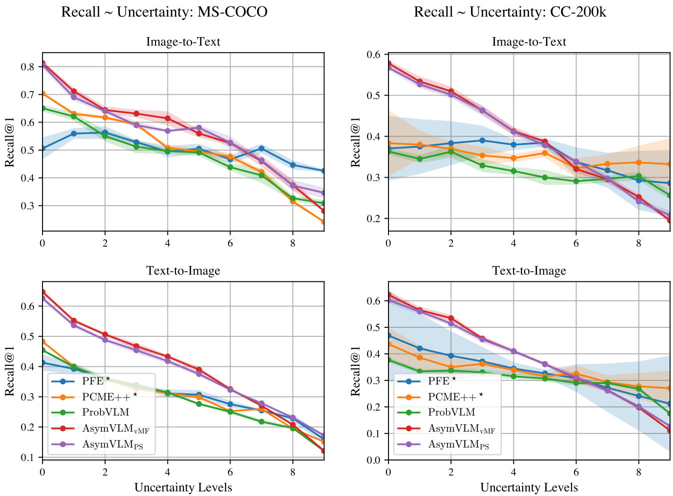
Denoting $\mathrm{CosSim}(r,s) = \mu(t_r)^\top z_s^I$, for any $r, s \in [B]$ we have,

for CLIP: $\delta_{\mathrm{CLIP}}(r,s) = \mathrm{CosSim}(r,s)$,

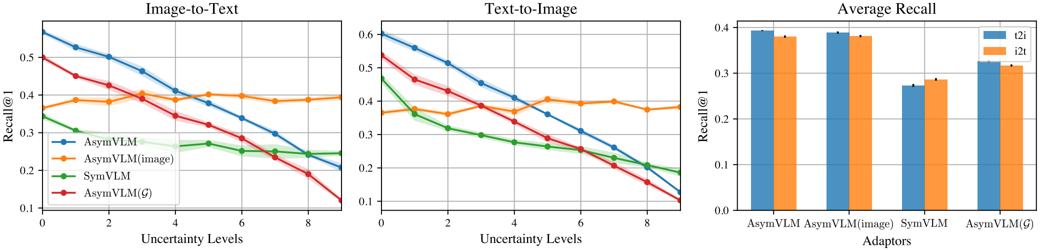
for AsymVLM$_{\mathrm{vMF}}$: $\delta_{\mathrm{vMF}}(r,s) = \kappa(t_r)\cdot \mathrm{CosSim}(r,s) + F_d(\kappa(t_r))$,

for AsymVLM$_{\mathrm{PS}}$: $\delta_{\mathrm{PS}}(r,s) = \kappa(t_r)\ln(1 + \mathrm{CosSim}(r,s) + \ln C_d(\kappa(t_r))$.

Empirical results: Uncertainty evaluation



Empirical results: ablation study



- Asymmetric structure is essential for uncertainty estimates.
- The choice of hyper-spherical (directional) distribution greatly improves the cross-modal retrieval performance.

Key Properties

Our method has following properties:

- Better cross-modal retrieval performance.
- Retrieval with uncertainty (estimated from likelihood).
- Robust fine-tuning.
- Robust zero-shot classification (know unknown).

Future Work

Ongoing works:

- Is logit adjustment a free lunch for heterogeneous federated learning?
- Federated heterogenous rank adaptation for pre-trained large models

Publications

Presented works:

- **Ju L**, Hellander A, Spjuth O. Federated learning for predicting compound mechanism of action based on image-data from cell painting. Artificial Intelligence in the Life Sciences. 2024 Jun 1;5:100098.
- **Ju L**, Zhang T, Toor S, Hellander A. Accelerating fair federated learning: Adaptive federated adam. IEEE Transactions on Machine Learning in Communications and Networking. 2024 Jul 4.
- **Ju L**, Andersson M, Fredriksson S, Glöckner E, Hellander A, Vats E, Singh P. Exploiting the Asymmetric Uncertainty Structure of Pre-trained VLMs on the Unit Hypersphere. arXiv preprint arXiv:2505.11029. 2025 May 16.

Other works:

- **Ju L**, Singh P, Toor S. Proactive autoscaling for edge computing systems with kubernetes. InProceedings of the 14th IEEE/ACM International Conference on Utility and Cloud Computing Companion 2021 Dec 6 (pp. 1-8).
- Li S, Ngai EC, Ye F, **Ju L**, Zhang T, Voigt T. Blades: A unified benchmark suite for byzantine attacks and defenses in federated learning. In2024 IEEE/ACM Ninth International Conference on Internet-of-Things Design and Implementation (IoTDI) 2024 May 13 (pp. 158-169). IEEE.
- Zhang T, **Ju L**, Singh P, Toor S. InfoHier: Hierarchical Information Extraction via Encoding and Embedding. arXiv preprint arXiv:arXiv:2501.08717. 2025 Jan 15.

Thank you for listening!

Questions?