Mixture of Experts Made Personalized: Federated Prompt Learning for Vision-Language Models

Jun Luo[†], Chen Chen[‡], and Shandong Wu[†]

[†] University of Pittsburgh, Pittsburgh, PA, USA
[‡] University of Central Florida, Orlando, FL, USA

Background and motivation

- Vision-Language Models (VLMs) like CLIP with their robust representation learning capabilities, show promise for addressing data heterogeneity in federated learning.

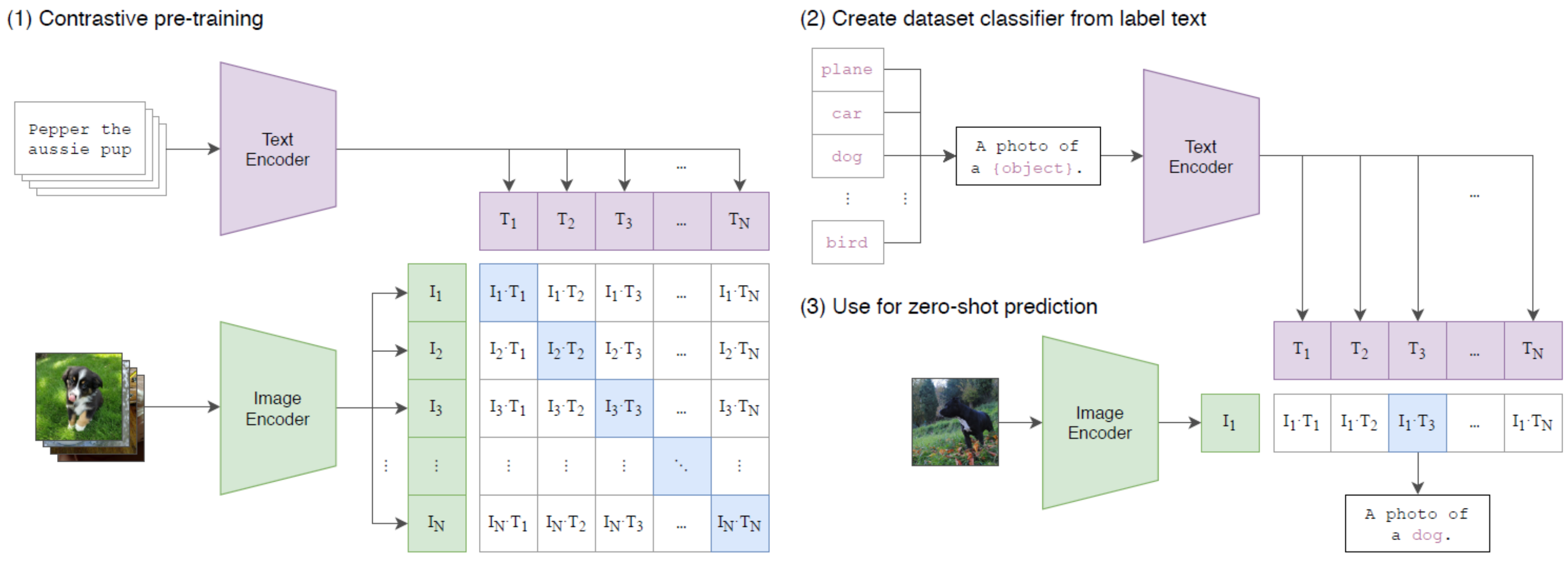
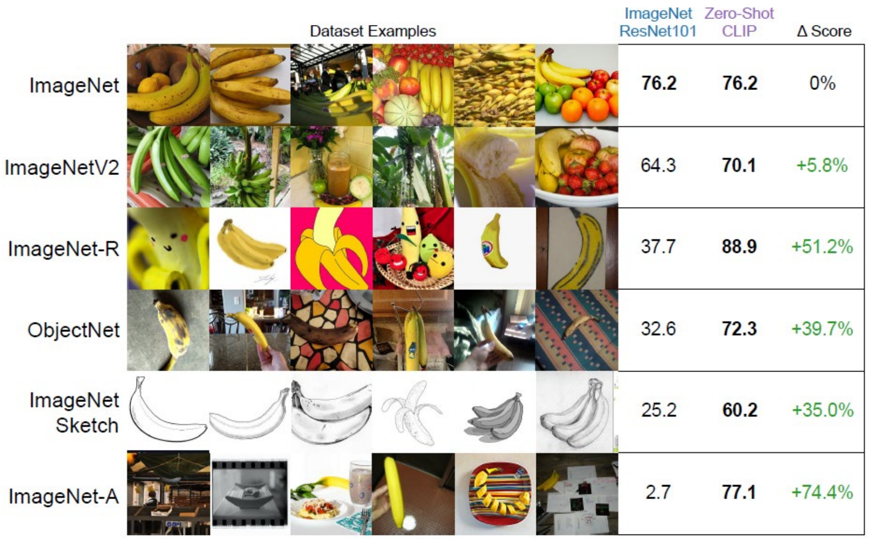


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.



[Figures from CLIP paper]

Background and motivation

- Vision-Language Models (VLMs) like CLIP with their robust representation learning capabilities, show promise for addressing data heterogeneity in federated learning.
- Traditional fine-tuning of VLMs in federated settings is challenging due to high communication overhead, leading researchers to explore prompt learning as a more efficient adaptation technique.

Caltech101

Prompt	Accuracy
a [CLASS].	82.68
a photo of [CLASS].	80.81
a photo of a [CLASS].	86.29
$[V]_1 [V]_2 \ldots [V]_M$ [CLASS].	**91.83**

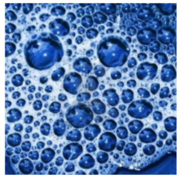
(a)

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
$[V]_1 [V]_2 \ldots [V]_M$ [CLASS].	**94.51**

(b)

Describable Textures (DTD)



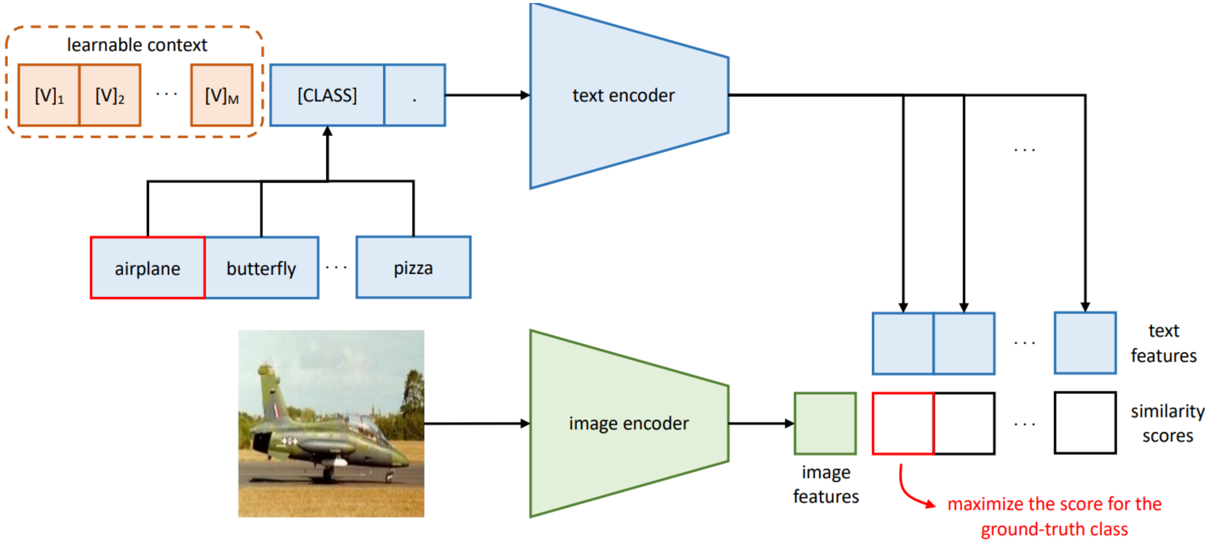
Prompt	Accuracy
a photo of a [CLASS].	39.83
a photo of a [CLASS] texture.	40.25
[CLASS] texture.	42.32
$[V]_1 [V]_2 \ldots [V]_M$ [CLASS].	**63.58**

(c)

EuroSAT

Prompt	Accuracy
a photo of a [CLASS].	24.17
a satellite photo of [CLASS].	37.46
a centered satellite photo of [CLASS].	37.56
$[V]_1 [V]_2 \ldots [V]_M$ [CLASS].	**83.53**

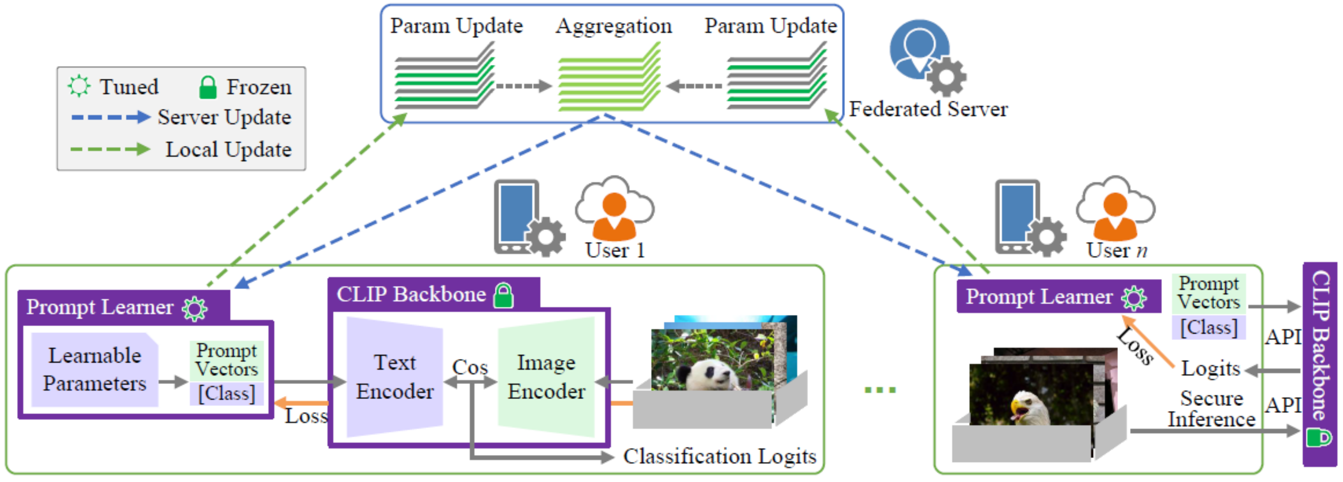
(d)



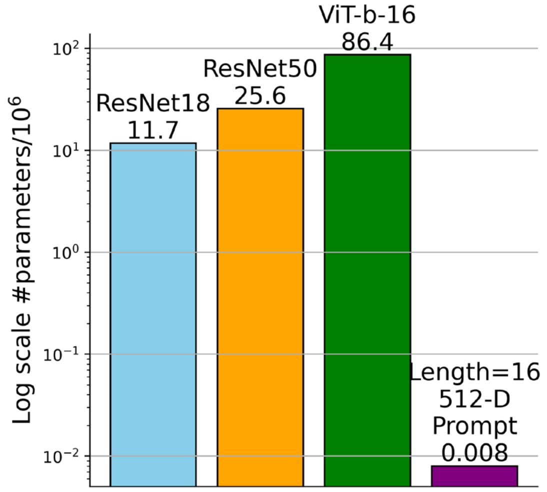
[Figures from CoOp paper]

Background and motivation

- Vision-Language Models (VLMs) like CLIP with their robust representation learning capabilities, show promise for addressing data heterogeneity in federated learning.
- Traditional fine-tuning of VLMs in federated settings is challenging due to high communication overhead, leading researchers to explore prompt learning as a more efficient adaptation technique.
- Existing federated prompt learning works
  - Habitually fall into traditional FL paradigm where clients are restricted to downloading only a single globally aggregated model – not fully leveraging the prompt's lightweight nature
  - Struggling to handle extreme data heterogeneity, lacking personalization strategies to handle



[Figure from CoOp paper]

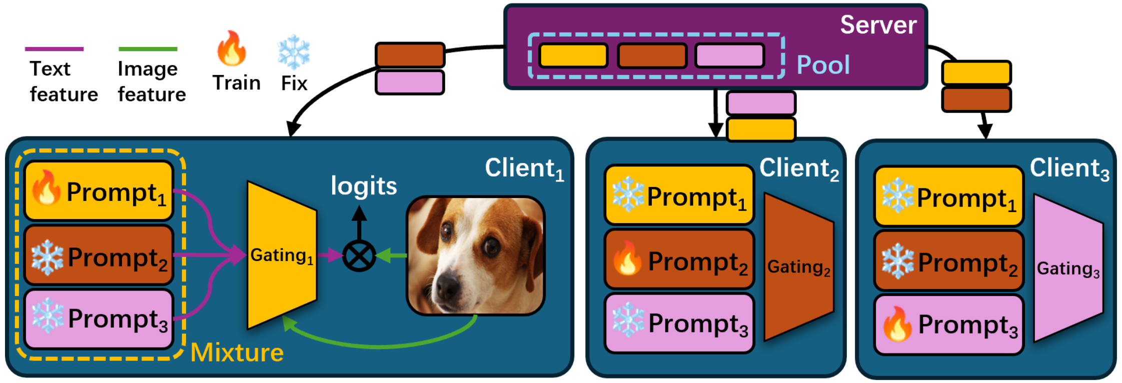


Background and motivation

Research question: *How can we devise a personalized federated learning framework, tailored for prompt learning in CLIP-like VLMs, while fully exploiting the lightweight nature of the prompts?*

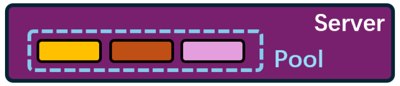
Background and motivation

Research question: *How can we devise a personalized federated learning framework, tailored for prompt learning in CLIP-like VLMs, while fully exploiting the lightweight nature of the prompts?*



- **P**ersonalized **Fed**erated **M**ixture **o**f **A**daptive **P**rompts (**pFedMoAP**)
  - Allows download of multiple pre-aggregated prompts
  - Uses a Mixture of Experts approach to treat locally updated prompts as specialized experts
  - Implements a client-specific, attention-based gating network to generate enhanced text features

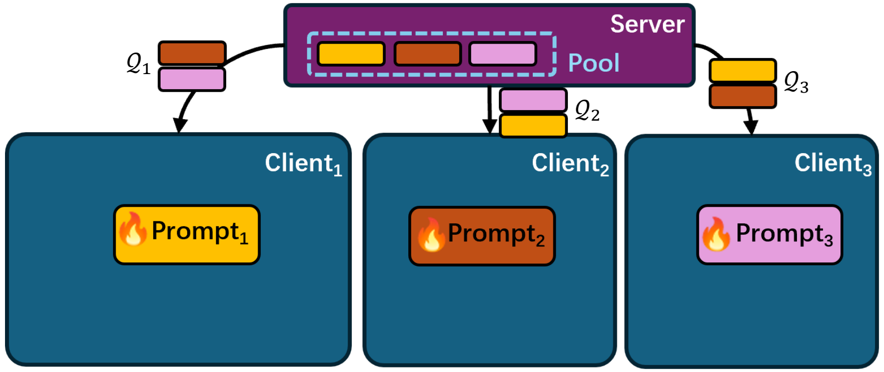
pFedMoAP – Method



- Workflow
  - Server maintains a pool of prompts $\mathcal{P}_t = \mathcal{P}_{t-1} - \{\boldsymbol{P}_i^{t-1}\}_{i\in\mathcal{P}_{t-1}\cap\mathcal{S}_t} + \{\boldsymbol{P}_j^{t}\}_{j\in\mathcal{S}_t}$

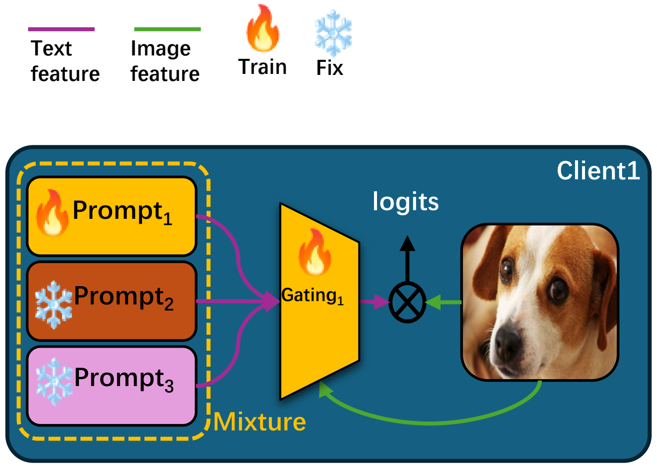
pFedMoAP – Method

- Workflow
  - Server maintains a pool of prompts $\mathcal{P}_t = \mathcal{P}_{t-1} - \{\boldsymbol{P}_i^{t-1}\}_{i \in \mathcal{P}_{t-1} \cap \mathcal{S}_t} + \{\boldsymbol{P}_j^t\}_{j \in \mathcal{S}_t}$
  - Each client $i \in S_t$ download K pre-aggregated (non-local) prompt
    - K-Nearest Neighbors (KNN) since most likely to have similar distribution
    - $\mathcal{Q}_i = \{NL_j\}_{j=1}^K$: set of clients assigned to client i, with prompts P_{NL_j}(NL= abbr. for non-local)



pFedMoAP – Method

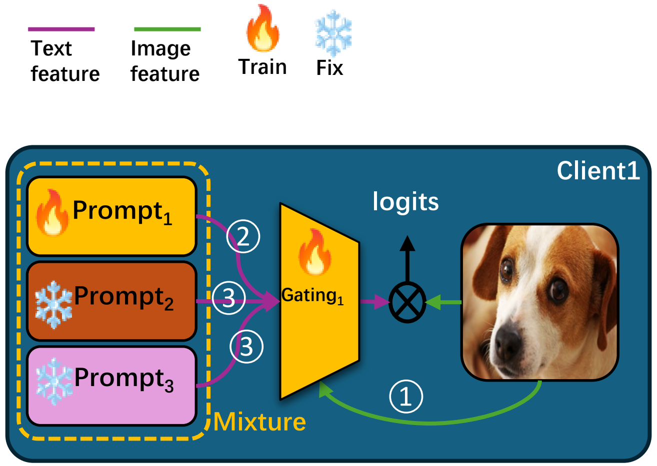
- Workflow
  - Server maintains a pool of prompts $\mathcal{P}_t = \mathcal{P}_{t-1} - \{\boldsymbol{P}_i^{t-1}\}_{i \in \mathcal{P}_{t-1} \cap \mathcal{S}_t} + \{\boldsymbol{P}_j^t\}_{j \in \mathcal{S}_t}$
  - Each client $i \in S_t$ download K pre-aggregated (non-local) prompt
    - K-Nearest Neighbors (KNN) since most likely to have similar distribution
    - $\mathcal{Q}_i = \{NL_j\}_{j=1}^K$: set of clients assigned to client i, with prompts P_{NL_j} (NL= abbr. for non-local)
  - Before local training, for once, client compute (fixed) text feature from non-local prompts $\forall c \in [C],\ \boldsymbol{T}_{NL}^{(c)} \triangleq \{\boldsymbol{T}_{NL_j}^{(c)} | \boldsymbol{T}_{NL_j}^{(c)} = g(\boldsymbol{P}_{NL_j}^{(c)}),\ \forall NL_j \in \mathcal{Q}_i\}$



pFedMoAP – Method

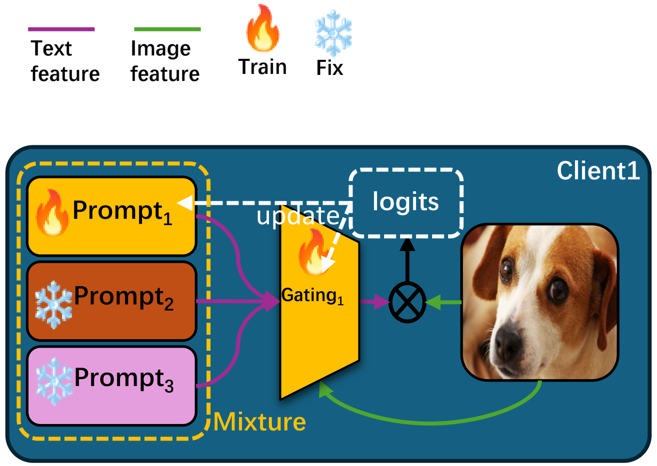
- Workflow
  - Server maintains a pool of prompts $\mathcal{P}_t = \mathcal{P}_{t-1} - \{P_i^{t-1}\}_{i \in \mathcal{P}_{t-1} \cap \mathcal{S}_t} + \{P_j^t\}_{j \in \mathcal{S}_t}$
  - Each client $i \in S_t$ download K pre-aggregated (non-local) prompt
    - K-Nearest Neighbors (KNN) since most likely to have similar distribution
    - $\mathcal{Q}_i = \{NL_j\}_{j=1}^K$: set of clients assigned to client i, with prompts P_{NL_j}(NL= abbr. for non-local)
  - Before local training, for once, client compute (fixed) text feature from non-local prompts $\forall c \in [C],\ \boldsymbol{T}_{NL}^{(c)} \triangleq \{\boldsymbol{T}_{NL_j}^{(c)} | \boldsymbol{T}_{NL_j}^{(c)} = g(\boldsymbol{P}_{NL_j}^{(c)}),\ \forall NL_j \in \mathcal{Q}_i\}$

  - Gating (detailed in following slides)
    - Input type 1: image feature $\boldsymbol{I}_k = f(\boldsymbol{x}_k)$
    - Input type 2: text feature from local prompt $\boldsymbol{T}_L^{(c)} = g(\boldsymbol{P}_i^{(c)})$
    - Input type 3: text features from non-local prompts $\boldsymbol{T}_{NL}^{(c)}$
    - Output: MoE text feature $\forall c \in [C],\ \boldsymbol{T}_{MoE}^{(c)} \triangleq G(\boldsymbol{I}_k, \boldsymbol{T}_L^{(c)}, \boldsymbol{T}_{NL}^{(c)} | \boldsymbol{\theta}_i)$



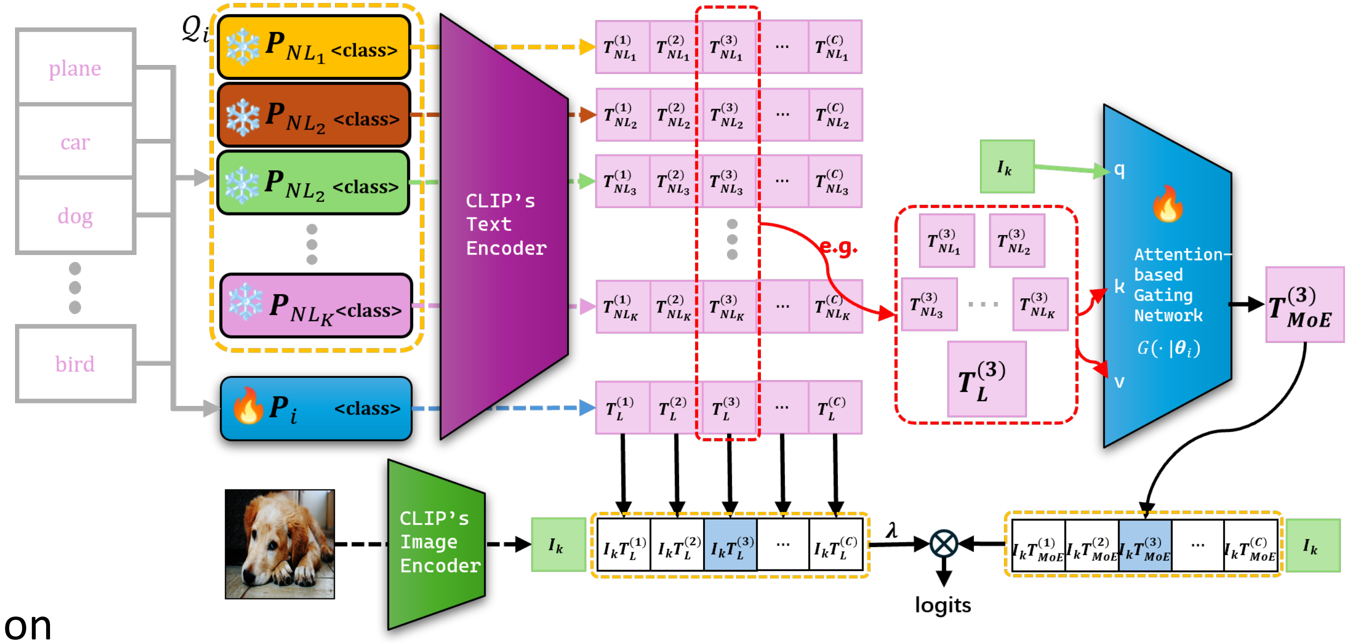
pFedMoAP – Method

- Workflow
  - Server maintains a pool of prompts $\mathcal{P}_t = \mathcal{P}_{t-1} - \{P_i^{t-1}\}_{i \in \mathcal{P}_{t-1} \cap \mathcal{S}_t} + \{P_j^t\}_{j \in \mathcal{S}_t}$
  - Each client $i \in S_t$ download K pre-aggregated (non-local) prompt
    - K-Nearest Neighbors (KNN) since most likely to have similar distribution
    - $\mathcal{Q}_i = \{NL_j\}_{j=1}^K$: set of clients assigned to client i, with prompts P_{NL_j} (NL= abbr. for non-local)
  - Before local training, for once, client compute (fixed) text feature from non-local prompts $\forall c \in [C],\ \boldsymbol{T}_{NL}^{(c)} \triangleq \{\boldsymbol{T}_{NL_j}^{(c)} | \boldsymbol{T}_{NL_j}^{(c)} = g(\boldsymbol{P}_{NL_j}^{(c)}),\ \forall NL_j \in \mathcal{Q}_i\}$
  - Gating (detailed in following slides)
    - Input type ①: image feature $\boldsymbol{I}_k = f(\boldsymbol{x}_k)$
    - Input type ②: text feature from local prompt $\boldsymbol{T}_L^{(c)} = g(\boldsymbol{P}_i^{(c)})$
    - Input type ③: text features from non-local prompts $\boldsymbol{T}_{NL}^{(c)}$
    - Output: MoE text feature $\forall c \in [C],\ \boldsymbol{T}_{MoE}^{(c)} \triangleq G(\boldsymbol{I}_k, \boldsymbol{T}_L^{(c)}, \boldsymbol{T}_{NL}^{(c)} | \theta_i)$
  - Final step: compute logits, manually address local prompt since it is the only locally learnable prompt $\forall c \in [C],\ \mathrm{logit}^{(c)} = \mathrm{sim}(\boldsymbol{I}_k, \boldsymbol{T}_{MoE}^{(c)}) + \lambda \cdot \mathrm{sim}(\boldsymbol{I}_k, \boldsymbol{T}_L^{(c)})$



pFedMoAP – Method

- Attention-based gating network: mechanism



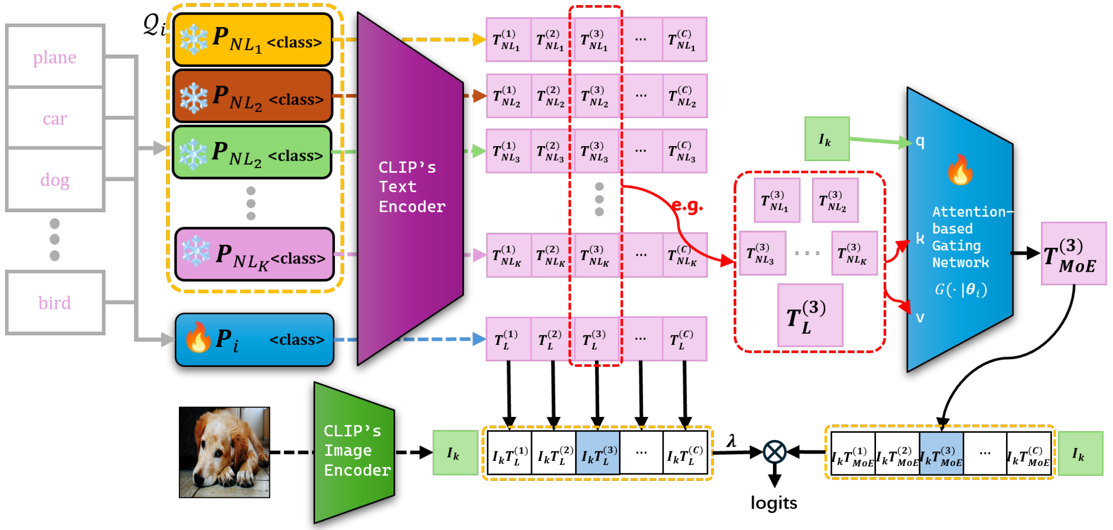
  - Multi-head attention
  - Pooling on features to reduce the size of gating from 1024 to 128
  - Q=Pooling($\boldsymbol{I}_k$), K=V=Pooling{ $\boldsymbol{T}_L^{(c)}, \boldsymbol{T}_{NL_1}^{(c)}, \boldsymbol{T}_{NL_2}^{(c)}, \ldots, \boldsymbol{T}_{NL_K}^{(c)}$ }
  - MoE text feature: $\boldsymbol{T}_{MoE}^{(c)} = G(\boldsymbol{I}_k, \boldsymbol{T}_L^{(c)}, \boldsymbol{T}_{NL}^{(c)}|\boldsymbol{\theta}_i) = \mathrm{MHA}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)W^O \quad \mathrm{head}_q = \mathrm{Attention}(QW_q^Q, KW_q^K, VW_q^V)$

pFedMoAP – Method

- Attention-based gating network: design rationale against traditional projection-based gating network
  - Projection-based gating network $G_{\text{proj}}(\boldsymbol{x}_k) \in \mathbb{R}^{K+1}$

$$MoE(\boldsymbol{x}) = \sum_{i=1}^{N} G(\boldsymbol{x})_i \cdot E_i(\boldsymbol{x})$$

  - Attention-based gating against projection-based gating
    - is more robust to adaptive experts
    - is agnostic to experts' order
    - serves as linear probing with more capacity
    - leverages CLIP's feature alignment with attention mechanism



pFedMoAP – Method

- Algorithm

Algorithm 4 pFedMoAP

Input: N clients, learning rates η_1, η_2, number of rounds T, logit coefficient λ, CLIP image/text encoder $f(\cdot), g(\cdot)$, datasets $\{D_i\}_{i \in [N]}$

Output: Personalized prompts $\boldsymbol{P}_1, \boldsymbol{P}_2, ..., \boldsymbol{P}_N$, gating network weights $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_N$.

ServerExecute:

Server initialize $\boldsymbol{P}_g^0$ and the pool of prompt experts $\mathcal{P}_0$ as an empty set
Clients intialize $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_N$.
for $t \leftarrow 1, 2, ..., T$ **do**
  Select a subset of $|\mathcal{S}_t|$ clients, $\mathcal{S}_t$
  for $i \in \mathcal{S}_t$ **in parallel do**
    if Client i does not have an entry in the server-maintained pool, $\mathcal{P}_t$ **then**
      $\boldsymbol{P}_i^t$ =ClientUpdate($\boldsymbol{P}_g^{t-1}$, standard=True)
    else
      Compute $\mathcal{Q}_i$ by K nearest neighbor, given $\boldsymbol{P}_i^t = \mathcal{P}_{t-1}[i]$.
      $\boldsymbol{P}_{NL} = \{\mathcal{P}_{t-1}[j]\}_{j \in \mathcal{Q}_i}$ // prompt in the pool from selected group of clients
      $\boldsymbol{P}_i^t$ =ClientUpdate($\boldsymbol{P}_g^{t-1}$, $\boldsymbol{P}_{NL}$) // downloaded as non-local experts
      $\mathcal{P}_t[i] = \boldsymbol{P}_i^t$ // cache to pool
    end if
    $\boldsymbol{P}_g^t = \sum_{i \in \mathcal{S}_t} p_i \boldsymbol{P}_i^t$
  end for
end for
return $\boldsymbol{P}_1^T, \boldsymbol{P}_2^T, ..., \boldsymbol{P}_N^T$ and $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_N$

ClientUpdate($\boldsymbol{P}_g^{t-1}$, $\boldsymbol{P}_{NL}$=None, standard=False):

$\boldsymbol{P}_i^t \leftarrow \boldsymbol{P}_g$
if standard **then**
  client does a standard fine-tuning
else
  for $(\boldsymbol{x}_k, y_k) \in D_i$ **do**
    $\boldsymbol{T}_L = g(\boldsymbol{P}_i^t)$
    $\boldsymbol{T}_{NL} = g(\boldsymbol{P}_{NL})$
    $\boldsymbol{I}_k = f(\boldsymbol{x}_k)$
    $\boldsymbol{T}_{MoE} = G(\boldsymbol{I}_k, \boldsymbol{T}_L, \boldsymbol{T}_{NL} | \boldsymbol{\theta}_i)$
    logit $= \text{sim}(\boldsymbol{I}_k, \boldsymbol{T}_{MoE}) + \lambda \cdot \text{sim}(\boldsymbol{I}_k, \boldsymbol{T}_L)$
    $p(\hat{y} = c | \boldsymbol{x}_k) = \text{Softmax}(\text{logit}, \tau)$ // τ is the temperature
    $\mathcal{L}_{ce} = -\sum_c y_k^{(c)} p(\hat{y} = c | \boldsymbol{x}_k)$
  end for
end if
return $\boldsymbol{P}_i^t$

pFedMoAP – Experiments & results

- Datasets

Dataset	Training Set Size	Test Set Size	Number of Classes	Number of Clients	Sample Rate	Data Heterogeneity
Flowers102	4,093	2,463	102	10	100%	Pathological non-IID
OxfordPets	2,944	3,669	37	10	100%	Pathological non-IID
Food101	50,500	30,300	101	10	100%	Pathological non-IID
Caltech101	4,128	2,465	100	10	100%	Pathological non-IID
DTD	2,820	1,692	47	10	100%	Pathological non-IID
Office-Caltech10	2,025	508	10	20	50%	Dir(0.3)
DomainNet	18,278	4,573	10	30	25%	Dir(0.3)
CIFAR10	50,000	10,000	10	100	10%	Dir(0.5)
CIFAR100	50,000	10,000	100	100	10%	Dir(0.5)

① CLIP datasets, pathological label shift
② Domain adaptation datasets, feature + label shift
③ CIFAR 10/100, Practical label shift

- Compared methods
  - Local methods
    - Zero-shot CLIP
    - CoOp (prompt learning)
  - Federated prompt learning + FL/PFL
    - PromptFL
    - PromptFL + FedProx
    - PromptFL + FT
    - PromptFL + FedAMP
    - PromptFL + FedPer
  - Personalization designed for federated prompt learning
    - pFedPrompt
    - FedOTP

pFedMoAP – Experiments & results

- Main results with different data shift and heterogeneity

Label shift

	Flowers102	OxfordPets	Food101	Caltech101	DTD
ZS-CLIP [71]	62.17±0.12	84.47±0.01	75.27±0.05	85.14±0.24	40.21±0.12
CoOp [100]	70.14±0.76	83.21±1.30	70.43±2.42	87.37±0.44	44.23±0.63
PromptFL [31]	72.80±1.14	90.79±0.61	77.31±1.64	89.70±1.99	54.11±0.22
PromptFL+FT [12]	72.31±0.91	91.23±0.50	77.16±1.56	89.70±0.25	53.74±1.36
PromptFL+FedPer [5]	72.11±1.35	89.50±1.62	71.29±1.87	86.72±1.45	50.23±0.82
PromptFL+FedProx [50]	66.40±0.29	89.24±0.41	76.24±1.94	89.41±0.55	44.26±1.11
PromptFL+FedAMP [37]	69.10±0.13	80.21±0.44	74.48±1.71	87.31±1.60	47.16±0.92
pFedPrompt [30]	86.46±0.15	91.84±0.41	92.26±1.34	96.54±1.31	77.14±0.09
FedOTP [48]	96.23±0.44	98.82±0.11	92.73±0.15	97.02±0.36	87.64±0.70
pFedMoAP (λ=0.0)	97.61±0.11	94.83±0.65	86.71±0.15	95.71±0.37	85.64±0.34
pFedMoAP (λ=0.5)	**98.41±0.04**	**99.06±0.09**	**93.39±0.09**	**97.95±0.07**	**89.13±0.54**

	CIFAR10	CIFAR10
ZS CLIP (Radford et al., 2021)	53.46±0.21	32.68±0.00
CoOp (Zhou et al., 2022b)	80.84±0.39	48.74±0.17
PromptFL (Guo et al., 2023b)	73.29±0.37	45.00±0.62
Prompt+FedProx (Li et al., 2020b)	73.32±0.34	45.63±0.75
pFedMoAP	83.46±0.53	53.42±0.22

Feature shift

	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Average
ZS CLIP	9.18±0.62	10.03±0.16	9.93±0.51	10.25±0.40	9.90±1.30	9.54±1.13	9.81±0.30
CoOp	43.84±3.51	45.72±0.85	29.94±0.46	36.83±1.17	31.64±0.49	33.97±0.78	36.99±0.79
PromptFL	27.63±16.41	27.69±18.07	21.62±8.34	23.45±13.49	20.62±11.03	25.90±8.10	24.48±12.52
Prompt+FedProx	22.23±15.42	21.75±17.00	18.58±8.15	19.40±12.59	17.17±10.25	22.49±8.44	20.27±11.83
pFedMoAP	47.49±0.64	46.73±0.71	32.74±0.84	37.16±0.34	31.02±0.59	37.67±0.72	38.80±0.11

	Amazon	Caltech	DSLR	Webcam	Average
ZS-CLIP (Radford et al., 2021)	9.83±1.63	10.67±0.89	10.89±1.40	6.20±3.84	9.40±0.77
CoOp (Zhou et al., 2022b)	30.29±3.64	35.88±1.30	29.89±5.15	33.43±2.25	32.37±1.81
PromptFL (Guo et al., 2023b)	21.08±9.60	23.72±12.21	22.94±7.96	25.88±7.72	23.41±9.06
Prompt+FedProx (Li et al., 2020b)	18.64±8.58	19.56±11.59	20.89±7.38	22.96±7.56	20.51±8.48
pFedMoAP	35.47±1.37	37.45±1.33	45.11±3.14	35.22±1.04	38.31±1.21

pFedMoAP – Experiments & results

- Differential privacy and visualization of MoE feature contributions

Table 6: Performance under (ϵ, δ)-differential privacy on CLIP datasets under pathological non-IID setting.

	Flowers102	OxfordPets	Food101	Caltech101	DTD
Without differential privacy (from Tab. 1)					
PromptFL (Guo et al., 2023b)	72.80±1.14	90.79±0.61	77.31±1.64	89.70±1.99	54.11±0.22
PromptFL+FedProx (Li et al., 2020b)	66.40±0.29	89.24±0.41	76.24±1.94	89.41±0.55	44.26±1.11
pFedMoAP(ours)	98.41±0.04	99.06±0.09	93.39±0.09	97.95±0.07	89.13±0.54
With differential privacy ($\epsilon = 50$)					
PromptFL (Guo et al., 2023b)	67.07±0.60	88.05±0.32	77.41±0.60	84.83±0.42	38.39±1.25
PromptFL+FedProx (Li et al., 2020b)	66.22±0.63	87.78±0.61	77.27±0.59	84.68±0.64	39.43±1.11
pFedMoAP(ours)	98.34±0.06	99.08±0.02	93.36±0.04	97.90±0.08	89.99±0.49
With differential privacy ($\epsilon = 25$)					
PromptFL (Guo et al., 2023b)	64.25±1.10	86.26±1.07	76.84±0.66	85.00±1.59	38.19±0.66
PromptFL+FedProx (Li et al., 2020b)	62.87±0.99	86.82±0.47	76.21±0.64	84.51±1.52	37.82±0.52
pFedMoAP(ours)	98.36±0.12	99.02±0.04	93.41±0.13	97.99±0.06	89.11±0.28

(a) Caltech101, 10 experts (b) DTD, 10 experts (c) Food101, 10 experts
(d) Flowers102, 10 experts (e) OxfordPets, 10 experts (f) DomainNet, 20 experts

Figure 6: Contribution of the experts based on averaged attention score across all test images. The first five charts are for CLIP datasets, for which there are 10 clients in each dataset. The last chart is for DomainNet with a total of 30 clients.

pFedMoAP – Experiments & results

- Ablation studies

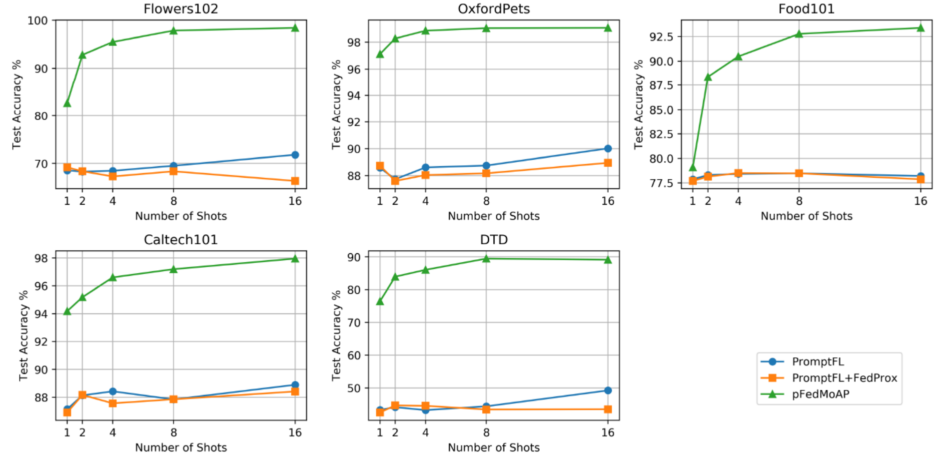


Figure 3: Ablation study on the number of shots.

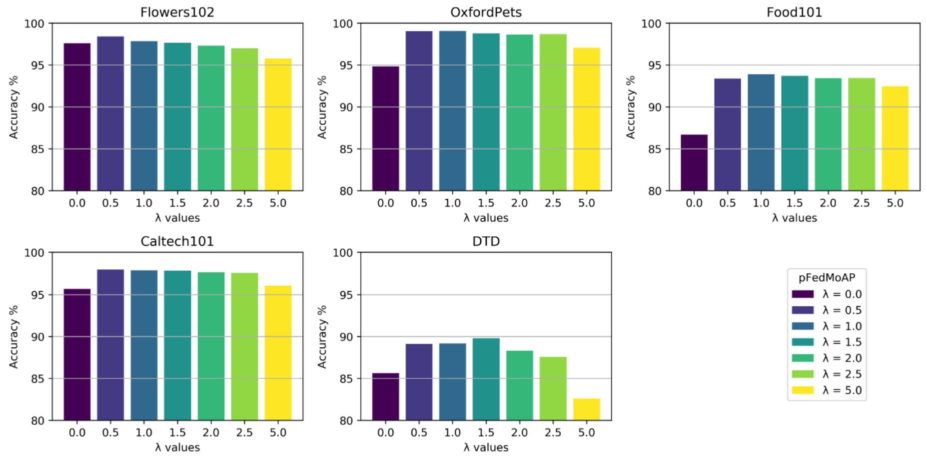


Figure 4: Ablation study on the coefficient for the logits from local prompt, λ.

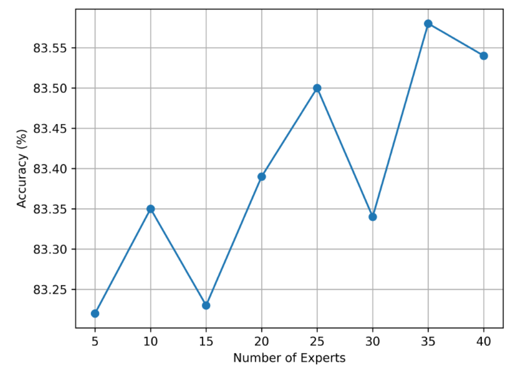
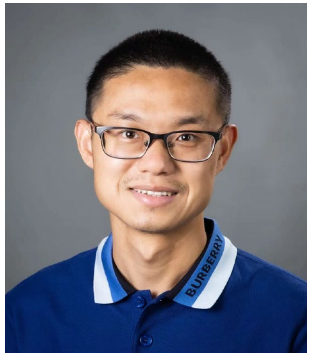


Figure 5: The impact of the number of experts on CIFAR10 with 100 clients

[More experiments in paper]

Acknowledgements

<u>Jun Luo</u>



Dr. Chen Chen

Dr. Shandong Wu

NIH NATIONAL CANCER INSTITUTE

NSF

aws

University of Pittsburgh

XSEDE

- NIH/NIC 1R01CA218405
- NSF CICI: SIVD: 2115082
- NSF/NIH 1R01EB032896
- Bridges-2 by ACCESS program
- NSF ACI-2138259, 2138286, 2138307, 2137603, 2138296

Project

GitHub