FL WITH MATCHED AVG

FEDERATED LEARNING WITH
MATCHED AVERAGING

REFERENCE

- Wang, Hongyi, et al.
"Federated learning with matched averaging."
arXiv preprint arXiv:2002.06440 (2020).

ABSTRACT

ABSTRACT

- Federated matched averaging (FedMA)
  - 최신 신경망을 위한 FL 알고리즘
  - CNNs, LSTMs

ABSTRACT

- FedMA는 글로벌 모델을 layer-wise하게 형성
- Hidden 요소들
  - Conv. 레이어의 채널
  - LSTM의 히든 스테이트 등을
- Matching
- Averaging

ABSTRACT

- SOTA(state-of-the-art)의 성능
- 커뮤니케이션 비용을 줄임

INTRODUCTION

INTRODUCTION

- 전통적인 FL 패러다임은 두 스테이지로 구성:
  - (i) 클라이언트가 모델을 로컬 데이터셋으로 독립적으로 학습
  - (ii) 데이터센터가 학습된 모델을 수집해 통합, 공유되는 글로벌 모델을 형성

INTRODUCTION

- 잘 알려진 표준 통합 방법으로는 FedAvg가 있음
  - 로컬 모델의 파라미터들을 element-wise하게 평균냄
  - 클라이언트의 데이터 수에 따른 가중평균을 내기도 함

INTRODUCTION

- FedProx
  - 이종적인 데이터 환경
  - 클라이언트 비용 함수에 proximal 항을 활용
  - 로컬 업데이트의 영향을 제한해
  - 글로벌 모델에 가깝게 유지하도록 함
- 클라이언트 드리프트

INTRODUCTION

- FedProx
  - 로컬 함수 F를 최소화하는 것이 아닌,
  - proximal 항을 포함한 다음을 최소화:

$$\min_{w} h_k(w;\ w^t) = F_k(w) + \frac{\mu}{2}\|w - w^t\|^2$$

INTRODUCTION

- Agnostic Federated Learning (AFL)
  - FedAvg의 변형
  - 중앙화된 분포가
  - 클라이언트 분포의 혼합에 의해 형성된
  - 목표 분포에 최적화

INTRODUCTION

- Agnostic Federated Learning (AFL)
- AFL에서는 중앙화된 모델이 특정 분포에 최적화되는 것이 아님
  - $\overline{\mathcal{U}} = \sum_{k=1}^{p} \frac{m_k}{m} \mathcal{D}_k$
  - 이 특정 분포가 목표와 부합하지 않을 수 있다는 높은 risk가 있기에
- 중앙화된 모델이 클라이언트 분포의 혼합으로부터 만들어지는
  - 어느 가능한 분포에나 최적화될 수 있도록 함

INTRODUCTION

- FedAvg에서 coordinate-wise averaging
    - 여러 문제가 생길 수 있음
    - 이는 신경망 파라미터의 치환 불변성 때문

INTRODUCTION

- 치환 분별성(permutation invariance)
  - 주어진 신경망 파라미터의 순서만 바꿔서 여러 변형을 만들 수 있음

INTRODUCTION

- 확률적(Probability) Federated Neural Matching (PFNM)
  - Averaging 하기 전
  - 클라이언트 신경망의 뉴런들을
  - 매칭함으로써 문제를 다룸

INTRODUCTION

- 또한, PFNM은 베이지안 non-parametric 방법으로
  - 글로벌 모델 크기를 조정하고
  - 데이터의 이종성을 반영

INTRODUCTION

- PFNM이 FedAvg 대비
  - 우수한 성능
  - 효율적인 커뮤니케이션 비용

INTRODUCTION

- 그러나 PFNM은 단순한 신경망에서만 동작
    - 가령, 전연결 feedforward 네트워크

CONTRIBUTION

- PFNM을 CNN과 LSTM에 적용
  - 그러나 가중치 평균 방법 대비 매우 적은 수준의 향상만 있었음

CONTRIBUTION

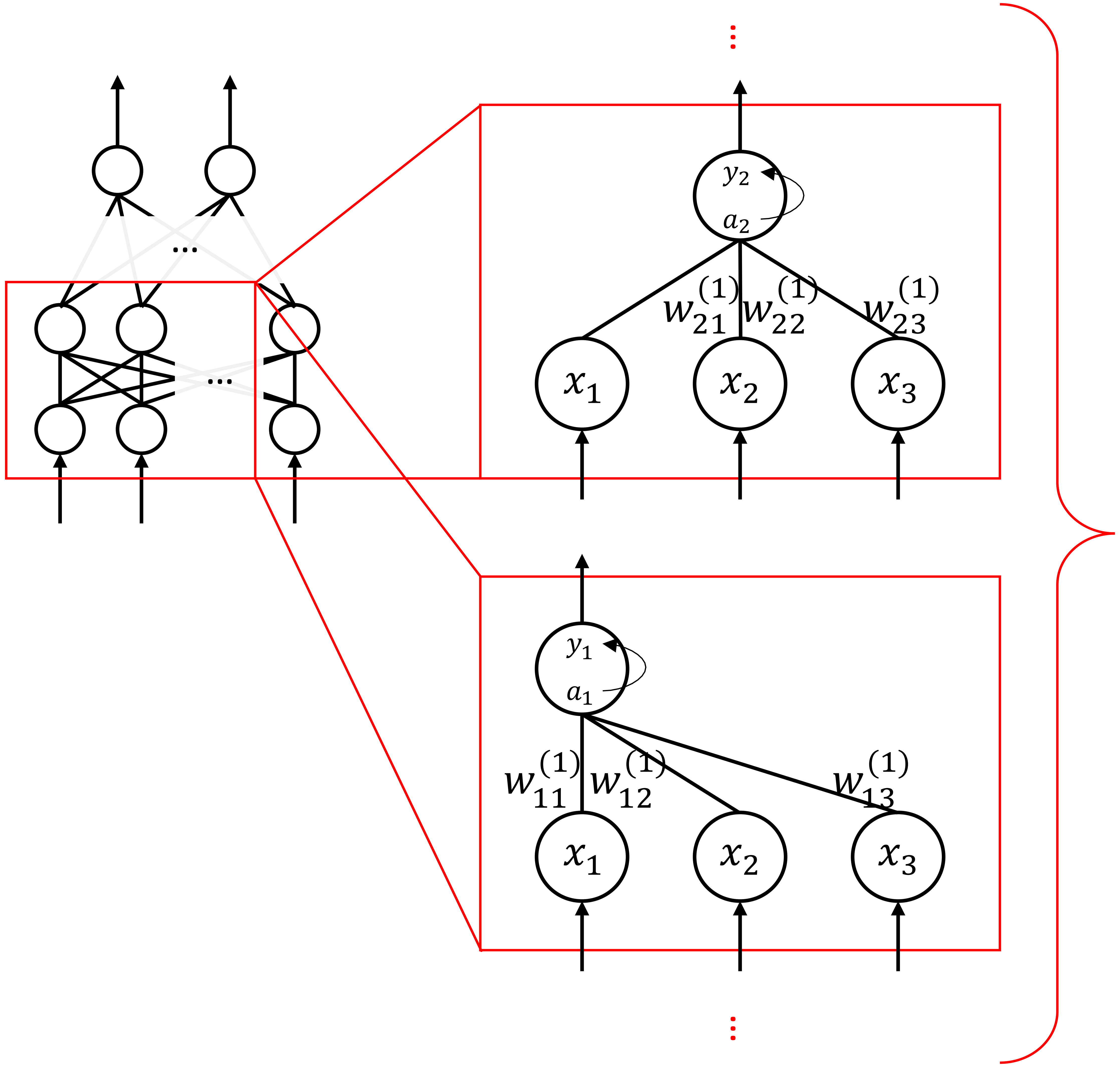
- Federated Matched Averaging (FedMA) 방법을 제안
    - 최신 CNN이나 LSTM을 위한
    - 새로운 layer-wise 연합 학습 알고리즘
- 커뮤니케이션 비용을 줄이고, SOTA의 성능을 보임

FEDMA

PERMUTATION INVARIANCE OF FULLY CONNECTED ARCHITECTURES

- 전연결(FC) 구조에서의 치환 불변성
- FC NN을 다음과 같이 수식화 가능:
- $\hat{y} = \sigma(xW_1)W_2$
  - 단순화를 위해 Bias 생략
  - σ 는 non-linearity, entry-wise하게 적용

PERMUTATION INVARIANCE OF FULLY CONNECTED ARCHITECTURES



$$\sigma\left(\underset{x}{\begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix}} \underset{W_1}{\begin{pmatrix} w_{11}^{(1)} & w_{21}^{(1)} \\ w_{21}^{(1)} & w_{22}^{(1)} \\ w_{31}^{(1)} & w_{23}^{(1)} \end{pmatrix}}\right) = \underset{\hat{y}}{\begin{pmatrix} y_1 & y_2 \end{pmatrix}}$$

PERMUTATION INVARIANCE OF FULLY CONNECTED ARCHITECTURES

- $\hat{y} = \sigma(xW_1)W_2$ 를 다음과 같이 확장 가능:
- $\hat{y} = \Sigma W_{2,i\cdot}\sigma(<x, W_{\cdot i}>), i = 1$ to L
  - <, > : 내적
  - $i\cdot$: 행렬의 행
  - $\cdot i$: 행렬의 열
  - L : hidden unit의 수

PERMUTATION INVARIANCE OF FULLY CONNECTED ARCHITECTURES

- $\hat{y} = \Sigma W_{2,i\cdot}\sigma(< x, W_{\cdot i} >), i = 1$ to L
- $$\sigma\left(\boxed{\begin{pmatrix}x_1 & x_2 & x_3\end{pmatrix}}\begin{pmatrix} w^{(1)}_{11} & w^{(1)}_{21}\\ w^{(1)}_{21} & w^{(1)}_{22}\\ w^{(1)}_{31} & w^{(1)}_{23}\end{pmatrix}\right) = (\boxed{y_1} \quad y_2)$$
  $$\sigma(\langle x, W_{1,\cdot i}\rangle) = y_1$$
- $$\underbrace{\begin{pmatrix} w^{(2)}_{11} & w^{(2)}_{21}\\ w^{(2)}_{12} & w^{(2)}_{22}\end{pmatrix}}_{W_2}\begin{matrix}\times y_1\\ \times y_2\end{matrix} \rightarrow \begin{pmatrix} w^{(2)}_{11}y_1 & w^{(2)}_{21}y_1\\ w^{(2)}_{12}y_2 & w^{(2)}_{22}y_2\end{pmatrix} \Sigma = (y_1 \quad y_2)W_2$$

PERMUTATION INVARIANCE OF FULLY CONNECTED ARCHITECTURES

- $\hat{y} = \Sigma\sigma(< x, W_{.i} >), i = 1$ to L
- 덧셈(Σ)은 치환 불변한 연산자
- 임의의 $L \times L$ 치환행렬(permutation matrix) Π
  에 대해 다음과 같이 쓸 수 있음:
- $\hat{y} = \sigma(xW_1\Pi)\Pi^T W_2$
  - 가령, $\Pi = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \Pi\Pi^T = I$

PERMUTATION INVARIANCE OF FULLY CONNECTED ARCHITECTURES

- $\hat{y} = \sigma(xW_1\Pi)\Pi^T W_2$
- 최적의 가중치가 $\{W_1, W_2\}$ 이라 가정,
- 두 이종적 데이터셋 $X_j, X_{j'}$ 에 대한 훈련된 가중치는
  - $\{W_1\Pi_j, \Pi_j^T W_2\}$ 와 $\{W_1\Pi_{j'}, \Pi_{j'}^T W_2\}$
- 높은 확률로 $\Pi_j \neq \Pi_{j'}$ 이므로
- 어떠한 Π에 대해서도 $(W_1\Pi_j + W_1\Pi_{j'})/2 \neq W_1\Pi$

PERMUTATION INVARIANCE OF FULLY CONNECTED ARCHITECTURES

- 따라서 의미있게 평균을 내기 위해서는 치환을 풀어야 함:
- $(W_1\Pi_j\Pi_j^T + W_1\Pi_{j'}\Pi_{j'}^T)/2 = W_1$

MATCHED AVERAGING FORMULATION

- 다음을 만족하는 최적화 문제를 풀어야 함:
- $$\min_{\{\pi^j_{li}\}} \sum_{i=1}^{L} \sum_{j,l} \min_{\theta_i} \pi^j_{li} c(w_{jl}, \theta_i) \quad s.t. \quad \sum_i \pi^j_{li} = 1 \forall j,l; \sum_l \pi^j_{li} = 1 \forall i,j$$
  - w_{jl} : 데이터셋 j에 대해 훈련된 l번째 뉴런
    - 즉, $W_1 \Pi_j$의 l번째 열
  - θ_i : 글로벌 모델의 i번째 뉴런
  - $c(\cdot,\cdot)$: 뉴런 한 쌍의 유사도 함수 (0에 가까울 수록 유사; 가령, 거리)

MATCHED AVERAGING FORMULATION

- $$\min_{\{\pi^j_{li}\}} \sum_{i=1}^{L} \sum_{j,l} \boxed{\min_{\theta_i} \pi^j_{li} c(w_{jl}, \theta_i)} \quad s.t. \quad \sum_i \pi^j_{li} = 1 \forall j,l; \sum_l \pi^j_{li} = 1 \forall i,j$$
- 데이터셋 j로 훈련한 신경망의 l번째 뉴런을
  글로벌 신경망의 i번째 뉴런과 유사도를 측정해
  임의의 치환행렬을 적용했을 때 크기가 가장 작은 조합을 찾음
  - 치환행렬의 원소 (1 또는 0)

MATCHED AVERAGING FORMULATION

- $$\min_{\{\pi_{li}^{j}\}} \sum_{i=1}^{L} \boxed{\sum_{j,l} \min_{\theta_i} \pi_{li}^{j} c(w_{jl}, \theta_i)} \quad s.t. \quad \sum_{i} \pi_{li}^{j} = 1 \forall j, l; \sum_{l} \pi_{li}^{j} = 1 \forall i, j$$
- 데이터셋 j로 훈련한 신경망의 l번째 뉴런을
  글로벌 신경망의 i번째 뉴런과 유사도를 측정해
  임의의 치환행렬을 적용했을 때 크기가 가장 작은 조합을 찾음
- 모든 로컬 데이터셋들과 모든 로컬 뉴런의 조합에 대해 누계

MATCHED AVERAGING FORMULATION

- $$\min_{\{\pi_{li}^{j}\}} \boxed{\sum_{i=1}^{L} \sum_{j,l} \min_{\theta_i} \pi_{li}^{j} c(w_{jl}, \theta_i)} \quad s.t. \quad \sum_i \pi_{li}^{j} = 1 \forall j,l; \sum_l \pi_{li}^{j} = 1 \forall i,j$$
- 데이터셋 j로 훈련한 신경망의 l번째 뉴런을
  글로벌 신경망의 i번째 뉴런과 유사도를 측정해
  임의의 치환행렬을 적용했을 때 크기가 가장 작은 조합을 찾음
- 모든 로컬 데이터셋들과 모든 로컬 뉴런의 조합에 대해 누계
- 모든 글로벌 신경망의 뉴런들에 대해 누계

MATCHED AVERAGING FORMULATION

- $$\boxed{\min_{\{\pi_{li}^j\}} \sum_{i=1}^{L} \sum_{j,l} \min_{\theta_i} \pi_{li}^j c(w_{jl}, \theta_i)} \quad s.t. \quad \sum_i \pi_{li}^j = 1 \forall j,l; \sum_l \pi_{li}^j = 1 \forall i,j$$
- 데이터셋 j로 훈련한 신경망의 l번째 뉴런을
  글로벌 신경망의 i번째 뉴런과 유사도를 측정해
  임의의 치환행렬을 적용했을 때 크기가 가장 작은 조합을 찾음
- 모든 로컬 데이터셋들과 모든 로컬 뉴런의 조합에 대해 누계
- 모든 글로벌 신경망의 뉴런들에 대해 누계
- 중 가장 작은 값을 갖는 치환행렬을 찾음

MATCHED AVERAGING FORMULATION

- 만일 $c(\cdot,\cdot)$ 를 squared Euclidean distance로 삼는다면
- k-means 클러스터링과 유사
  - 단, 클러스터 정렬 항인 π_{li}^{j}가 사용됨

MATCHED AVERAGING FORMULATION

- 로컬 신경망과 글로벌 모델이 같은 수의 히든 뉴런을 가진다면
    - Wasserstein 무게중심(barycenter)으로 취급할 수 있음
    - $$\bar{\mu}_k = \arg\min_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \frac{1}{n_k} \sum_{i=1}^{n_k} \mathcal{W}_2^2(\nu, \mu_{k,i})$$

SOLVING MATCHED AVERAGING

- 최적화 문제를 어떻게 해결할 것인가?
  - 헝가리안 정합 알고리즘(Hungarian matching algorithm)을 사용
- $\{\theta_i = \arg\min_{\theta_i} \sum_{j \neq j', l} \pi^j_{li} c(w_{jl}, \theta_i)\}_{i=1}^{L}$

SOLVING MATCHED AVERAGING

- 연합 학습에서 클라이언트들이
  이종 데이터를 가지고 있는 상황을 상정해야만 함
- 각 클라이언트는 개별적 데이터에 대한
  특징 추출기(feature extractor)를 학습할 것
- 이들 특징 추출기들은 오직 부분적으로만 서로 겹칠 것

SOLVING MATCHED AVERAGING

- 이 문제를 다루기 위해 글로벌 모델의 사이즈를
  - 최소 로컬 모델만큼은 돼야하며
  - 최대 로컬 모델들의 결합(concatenation)만큼 커야 함

SOLVING MATCHED AVERAGING

- 적응형 크기를 가진 글로벌 모델에서
- 사이즈가 다르면 헝가리안 알고리즘을 적용할 수 없음
  - 헝가리안 알고리즘을 반복해서 적용
- 데이터 이종성 덕분에 poor한 매칭을 피할 수 있음

SOLVING MATCHED AVERAGING

- 글로벌 모델은 작을 수록 좋음
  - 커뮤니케이션 비용 등으로부터
- 증가 함수 $f(L')$을 추가해 크기 증가에 따른 패널티를 줌
- $$\min_{\{\pi_{li}^{j'}\}_{l,i}} \sum_{i=1}^{L+L_{j'}} \sum_{j=1}^{L_{j'}} \pi_{li}^{j'} C_{li}^{j'} \text{ s.t. } \sum_i \pi_{li}^{j'} = 1 \,\forall\, l; \sum_l \pi_{li}^{j} \in \{0,1\} \,\forall\, i, \text{ where}$$

$$C_{li}^{j'} = \begin{cases} c(w_{j'l}, \theta_i), & i \le L \\ \epsilon + f(i), & L < i \le L + L_{j'}. \end{cases}$$

SOLVING MATCHED AVERAGING

- Π가 더 이상 치환행렬이 아님
  - $L \times L_j$ 의 크기를 가짐
  - 치환행렬에 0으로 패딩한 형태
- 가중치 W_1 역시 그러함
- Dummy 뉴런들에 대해 고려하지 않도록
  - 평균낼 때 그 수만큼 빼줘야 함

SOLVING MATCHED AVERAGING

- 유사도 함수 $c(\cdot, \cdot)$와 역치 ϵ의 크기, 패널티 $f(\cdot)$를 결정해야 함
- Probabiilistic Federated Neural Matching (PFNM) 연구에서 고려됨
  - Beta-Bernoulli process(BBP)에 기반한
  - Bayesian nonparametric 모델의
  - Maximum a posteriori (MAP) 추정을 계산

PROBABILISTIC FEDERATED NEURAL MATCHING

REFERENCE

- Yurochkin, Mikhail, et al.
"Bayesian nonparametric federated learning of neural networks."
arXiv preprint arXiv:1905.12022 (2019).
- TBA

FL WITH MATCHED AVG

FEDERATED LEARNING WITH
MATCHED AVERAGING