THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

CREATE CHANGE

Learning Private Neural Language Modeling with Attentive Aggregation

Shaoxiong Ji

MPhil Candidate

The University of Queensland

Co-authors: Shirui Pan[†], Guodong Long[‡], Xue Li[*], Jing Jiang[‡], Zi Huang[*]

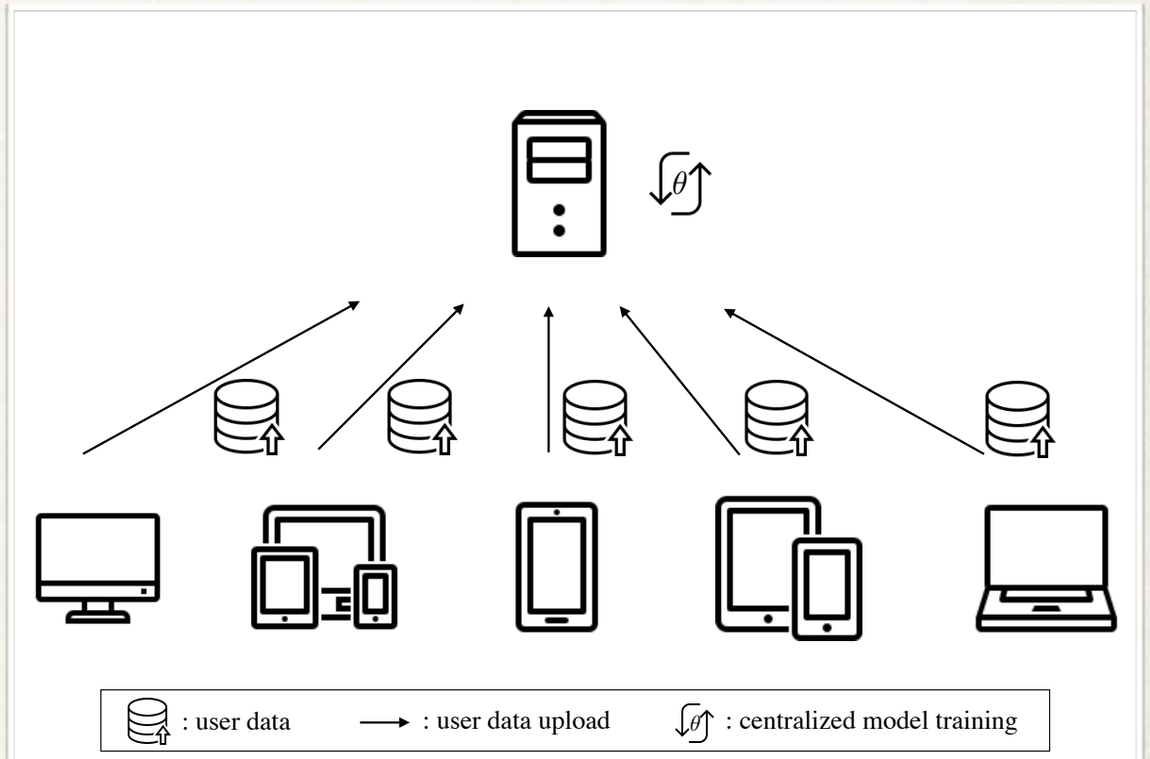
[*] School of ITEE, Faculty of EAIT, The University of Queensland, Australia

[†] Faculty of Information Technology, Monash University, Australia

[‡] Centre for Artificial Intelligence, FEIT, University of Technology Sydney, Australia

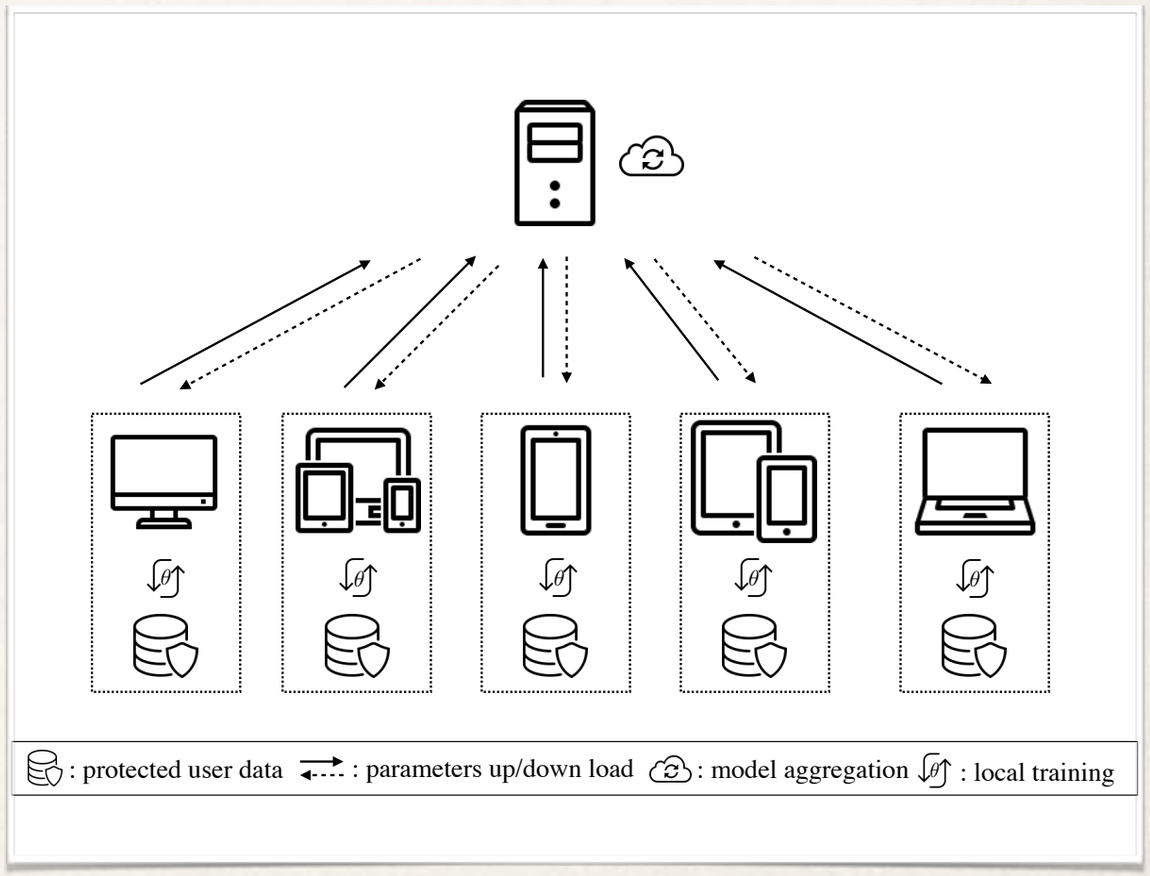
Intro: Federated Learning

- Centralized machine learning
  - Collect data from clients;
  - train a centralized model;
  - make predictions for clients



Intro: Federated Learning

- Federated learning: training a shared global model, from a federation of participating devices which maintain control of their own data, with the facilitation of a central server*.
- Real-world mobile applications
  - suggesting mobile keyboards;
  - retrieving important notifications;
  - detecting the spam messages



* Jakub Konečny. Federated Learning: Privacy-Preserving Collaborative Machine Learning without Centralized Training Data. Trends in Optimization Seminar, University of Washington in Seattle. Jan 30, 2018

Federated Averaging

- Federated Optimization

$$M \leftarrow \sum_{k=1}^{K} m_k$$

- Federated Averaging

$$\theta_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} \theta_{t+1}^{k}$$

- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282, 2017.

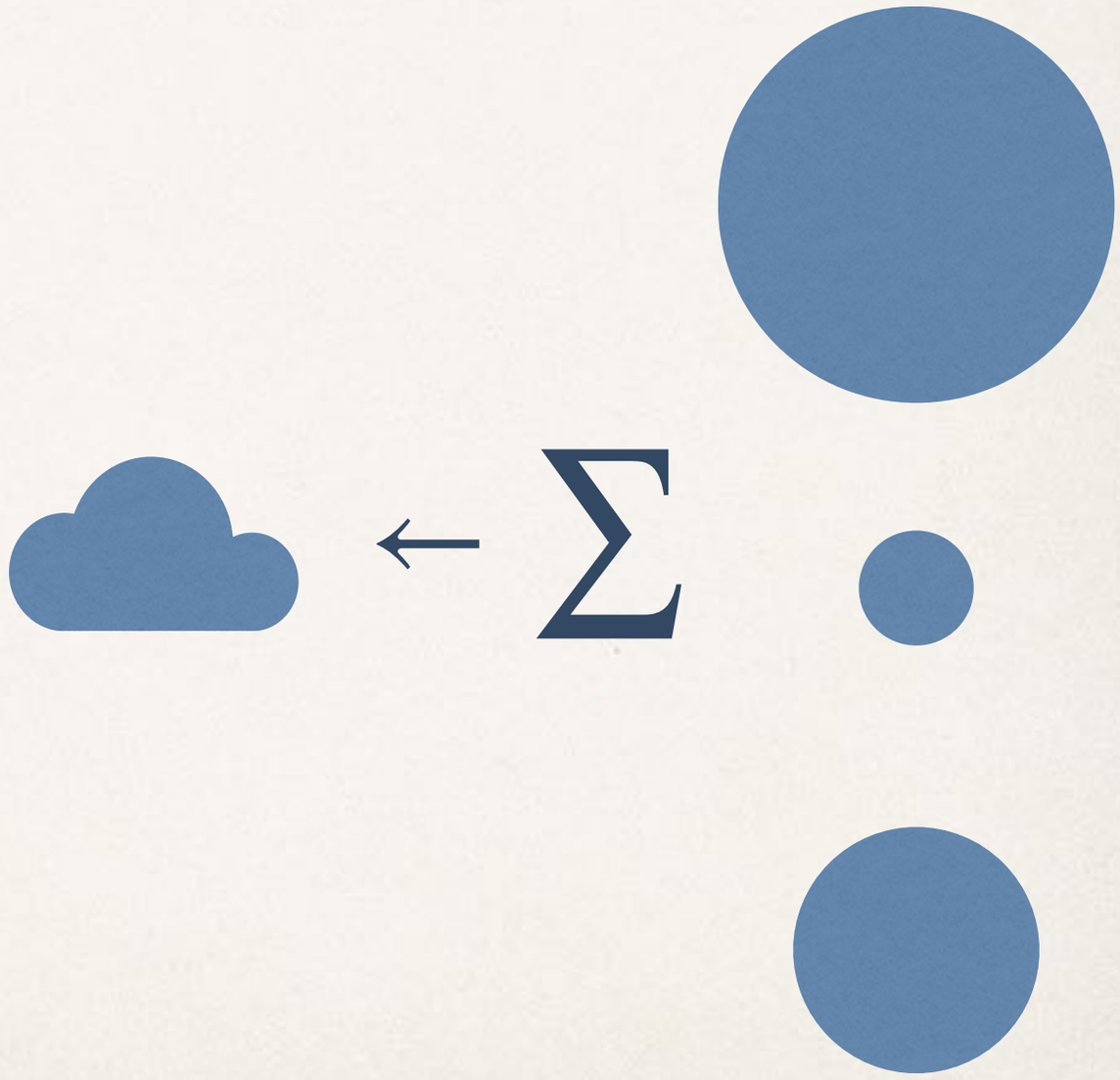
Private Language Modeling

- Personalized Keyboard Suggestion
  - next word prediction
  - language preference and patterns
- Private Neural Language Modeling
  - protect private content and sensitive information

Motivation

- Simple averaging ignores the importance of different distributed clients.

- Model importance and knowledge transferring should be considered.



Proposed Method

- FedAtt: Attentive Federated Aggregation

$$\theta_{t+1} \leftarrow \theta_t - \epsilon \sum_{k=1}^{m} \alpha_k (\theta_t - \theta_{t+1}^k),$$

- Attentive weighted expected distance

$$\arg\min_{\theta_{t+1}} \sum_{k=1}^{m} [\frac{1}{2} \alpha_k L(\theta_t, \theta_{t+1}^k)^2],$$

1. Considering the relation between the server model and client models and their weights

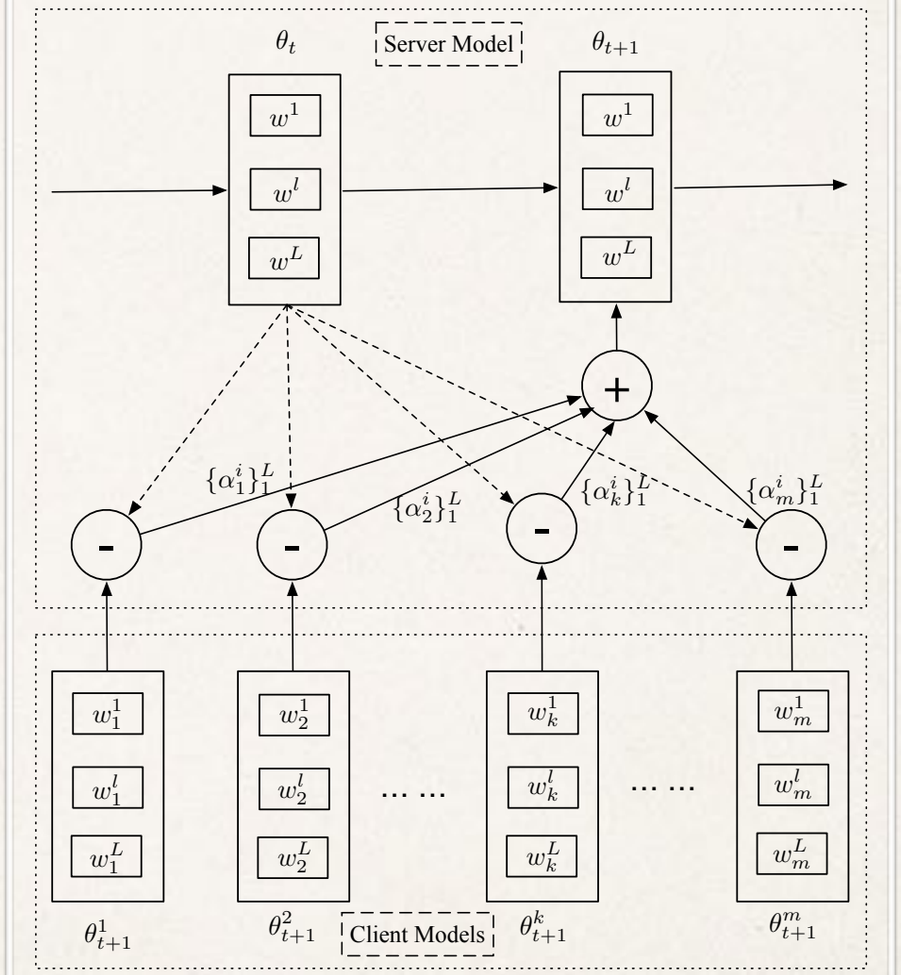
2. Optimizing the distance between the server model and client models in parameter space

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

Attentive Federated Aggregation

Algorithm 3 Attentive Federated Optimization

k is the ordinal of clients; l is the ordinal of neural layers; ϵ is the stepsize of server optimization
Input: server parameters θ_t at t, client parameters $\theta_{t+1}^1, ..., \theta_{t+1}^m$ at $t+1$.
Output: aggregated server parameters θ_{t+1}.
procedure ATTENTIVE OPTIMIZATION(θ_t, θ_{t+1}^k)
    Initialize $\alpha = \{\alpha_1, \ldots, \alpha_k, \ldots, \alpha_m\}$ ▷ attention for each clients
    for each layer $l = 1,\ 2, ...$ **do**
        for each user k **do**
            $s_k^l = \|w^l - w_k^l\|_p$
            $\alpha_k^l = softmax(s_k^l) = \frac{e^{s_k^l}}{\sum_{k=1}^m e^{s_k^l}}$
    $\alpha_k = \{\alpha_k^0, \alpha_k^1, \ldots, \alpha_k^l, \ldots\}$
    $\theta_{t+1} \leftarrow \theta_t - \epsilon \sum_{k=1}^m \alpha_k(\theta_t - \theta_{t+1}^k)$
    return θ_{t+1}



The illustration of our proposed layer-wise attentive federated aggregation

GRU-based Client Learner

- Client side: model-agnostic
- GRU-based language modeling

$$
\begin{aligned}
z_t &= \sigma(w_z \cdot [h_{t-1},\ x_t]), \\
r_t &= \sigma(w_r \cdot [h_{t-1},\ x_t]), \\
\tilde{h_t} &= tanh(w \cdot [r_t * h_{t-1},\ x_t]), \\
h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h_t},
\end{aligned}
$$

- K. Cho, B. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014.

Differential Privacy

- Differential privacy
- Randomized mechanism

$$\theta_{t+1} = \theta_t - \frac{1}{m}(\sum_{k=1}^{K} \Delta\theta_{t+1}^k + \mathcal{N}(0, \sigma^2))$$

- Attentive Federated Aggregation with Randomization

$$\theta_{t+1} \leftarrow \theta_t - \epsilon \sum_{k=1}^{m} \alpha_k(\theta_t - \theta_{t+1}^k + \beta\mathcal{N}(0, \sigma^2))$$

Experiments

- Datasets: Penn Treebank, WikiText-2, Reddit Comments
- Data Partitioning: shuffle and sampling
- Settings:
  - 1) FedSGD: Federated stochastic gradient descent takes all the clients for federated aggregation and each client performs one epoch of gradient descent.
  - 2) FedAvg: Federated averaging samples a fraction of users for each iteration and each client can take several steps of gradient descent.
  - 3) FedAtt: Our proposed FedAtt takes a similar setting as FedAvg, but uses an improved attentive aggregation algorithm.

Code is available at https://github.com/shaoxiongji/fed-att

Experiments

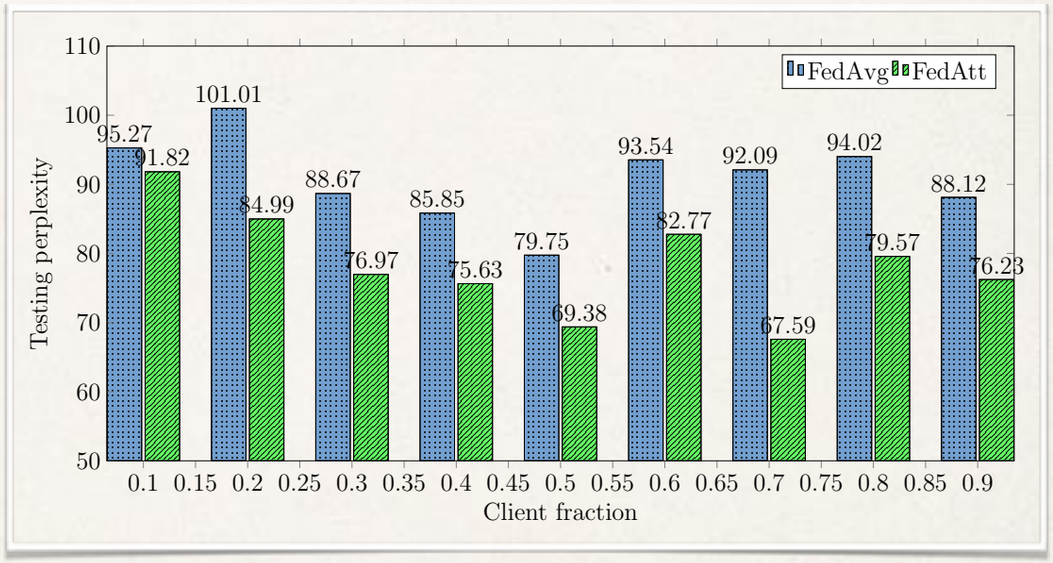
Results

Testing perplexity of 50 communication rounds for federated training using small-scaled GRU network as the client model

Frac.	Methods	WikiText-2	PTB	Reddit
1	FedSGD	112.45	155.27	128.61
0.1	FedAvg	95.27	138.13	126.49
	FedAtt	91.82	115.43	120.25
0.5	FedAvg	79.75	128.24	101.64
	FedAtt	69.38	123.00	99.04

Client Fraction

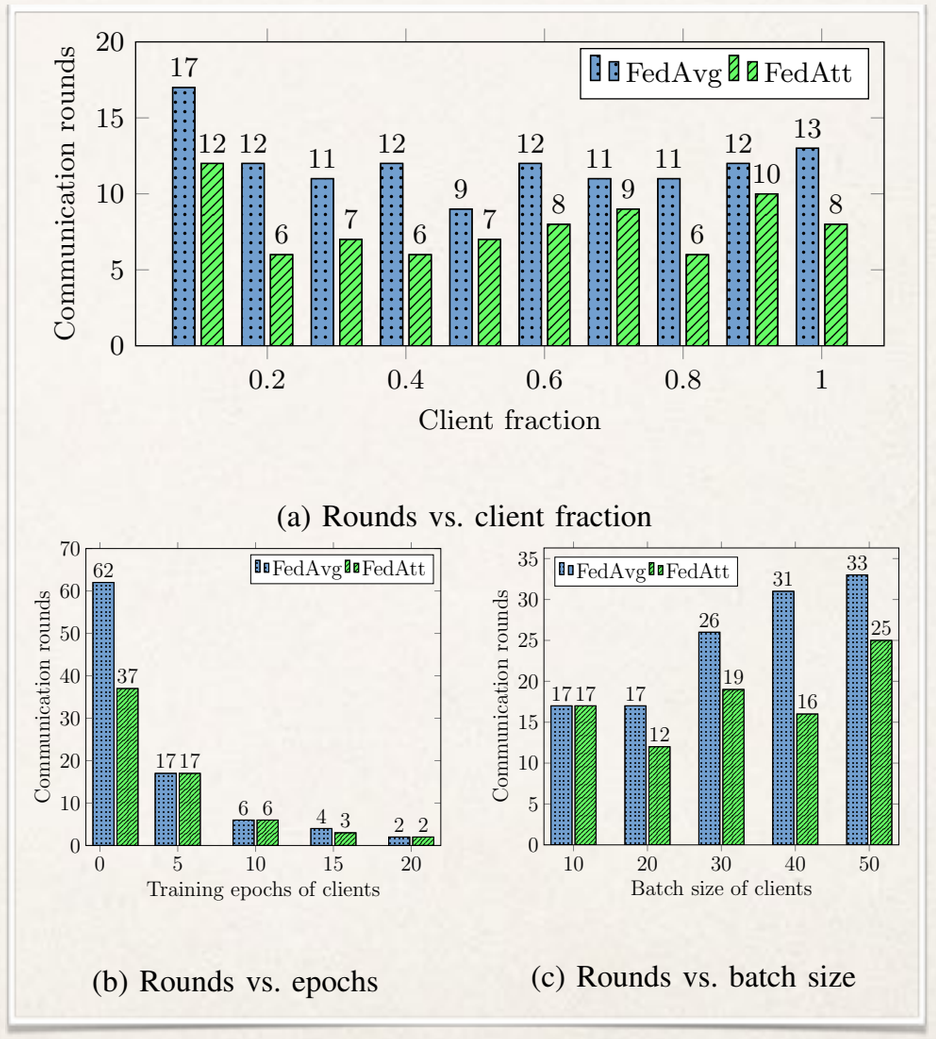
Testing perplexity of 50 communication rounds when a different number of clients is selected for federated aggregation



Experiments

Communication Cost

Effect of the client fraction, epochs, and batch size of clients on communication rounds when the threshold of testing perplexity is set to be 90 and small-scaled GRU-based language model is used



(a) Rounds vs. client fraction

(b) Rounds vs. epochs

(c) Rounds vs. batch size

Experiments

Randomization

Magnitude of randomization vs. testing perplexity using a small-scaled model with tied embedding

Randomization		FedAvg	FedAtt
Nonrandomized	$\beta = 0$	88.21	77.66
Randomized	$\beta = 0.001$	88.17	77.76
	$\beta = 0.005$	88.36	78.59
	$\beta = 0.01$	89.74	79.51
	$\beta = 0.05$	103.17	101.82

Model Scale

Testing perplexity of 50 communication rounds vs. the scale of the model using a tied embedding or untied embedding model

Model		FedAvg	FedAtt
Small	tied	88.21	77.66
	untied	91.25	81.31
Medium	tied	103.07	77.41
	untied	96.67	96.71
Large	tied	77.51	76.37
	untied	82.97	83.40

Conclusion

- This paper introduces the attention mechanism to aggregate multiple distributed models

- A novel layer-wise soft attention to capturing the "attention" among many local models' parameters is proposed.

- Experiments of next word prediction show a comparable performance in terms of perplexity and communication rounds

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

CREATE CHANGE

Q&A

Contacts

shaoxiong.ji@uq.edu.au

https://shaoxiongji.github.io/