FedNLP: Federated Learning for Natural Language Processing

Xiang Ren, USC CS

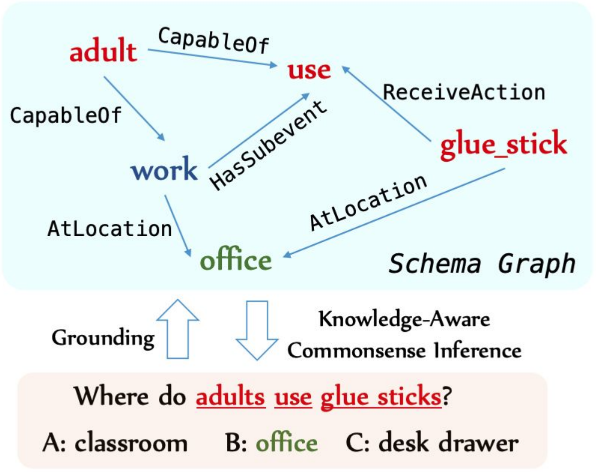
Mahdi Soltanolkotabi, USC ECE

USC Viterbi School of Engineering

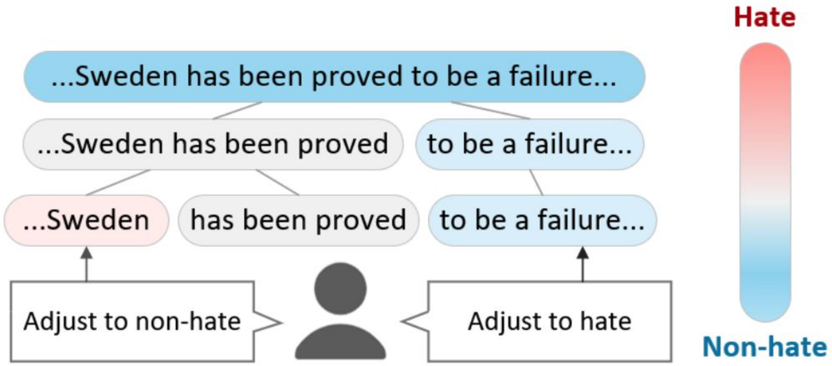
amazon

Prof. Xiang Ren (PI)

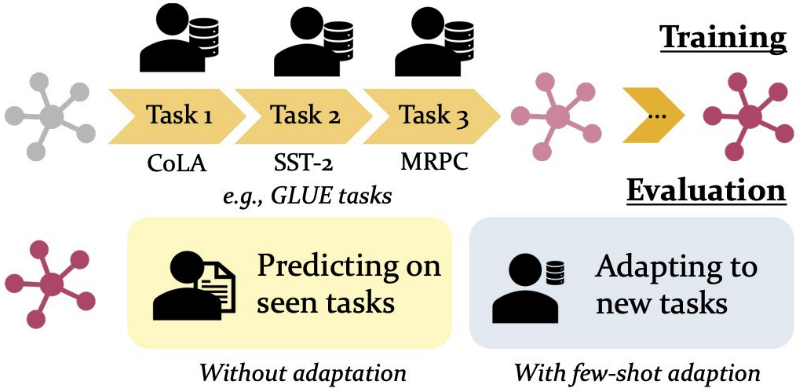
- **Expertise**: natural language processing, explainable AI, continual learning
- Make NLP systems trustworthy, cheaper to build, easier to maintain
- Create benchmark datasets for a range of NLP tasks



Commonsense reasoning



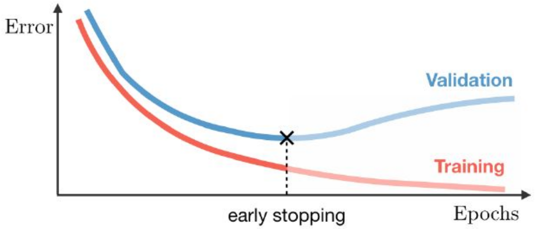
Human-in-the-loop learning, Explanability & Robustness



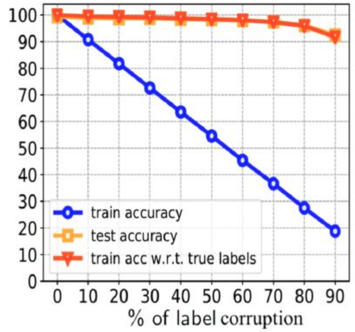
Cross-task Generalization

Prof. Mahdi Soltanolkotabi

- **Expertise:** Foundations of AI & ML, (non)convex optimization, high-dimensional statistics and probability
- Developed some of the first guarantees for deep learning



Optimization
(how neural nets fit the training data)



Robustness to label noise

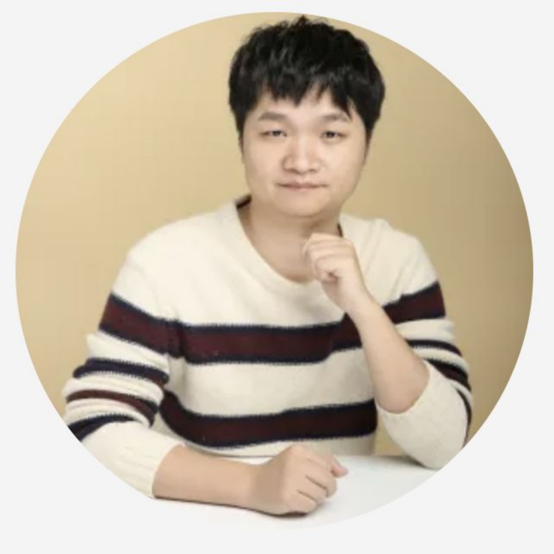
Generalization
(how neural networks predict)

- Extensive contributions to distributed/federated learning

Bill Yuchen Lin, PhD candidate @ USC

- **(Bill) Yuchen Lin** is a Ph.D. candidate in USC CS, working with Prof. Xiang Ren at the *Intelligence and Knowledge Discovery Research Lab*
- Develop intelligent systems that demonstrate a deep understanding of the world with **common-sense knowledge and reasoning ability**
- **Research interests**: information extraction, knowledge graphs, logical reasoning, graph neural networks, explanations, robustness

Chaoyang He, PhD candidate @ USC

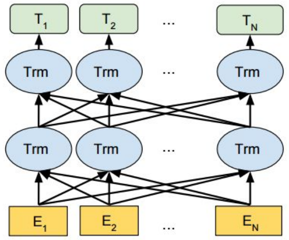


- Chaoyang He is a Ph.D. Candidate in the CS department at the University of Southern California. He is advised by Professor Salman Avestimehr, Professor Mahdi Soltanolkotabi, and Professor Murali Annavaram (USC). He also works closely with researchers/engineers at Google, Facebook, Amazon, and Tencent. Previously, He was an R&D Team Manager and Staff Software Engineer at Tencent (2014-2018), a Team Leader and Senior Software Engineer at Baidu (2012-2014), and a Software Engineer at Huawei (2011-2012).
- His research focuses on **distributed/federated machine learning algorithms, systems, and applications (https://FedML.ai, https://DistML.ai)**.
- Chaoyang He has received a number of awards in academia and industry, including Amazon ML Fellowship (2021-2022), Qualcomm Innovation Fellowship (2021-2022), Tencent Outstanding Staff Award (2015-2016), WeChat Special Award for Innovation (2016), Baidu LBS Group Star Awards (2013), and Huawei Golden Network Award (2012).

A Surprisingly "Simple" Recipe for Modern NLP

Model

+



Labeled Data

+

SWAG

SQuAD
The Stanford Question Answering Dataset

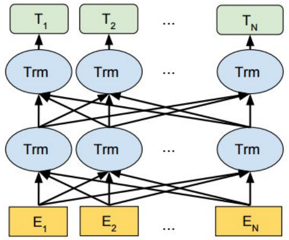
GLUE

Computing Power

A Surprisingly “Simple” Recipe for Modern NLP

Model

+



pip install transformers
~~from transformers import BertModel~~
from transformers import RobertaModel

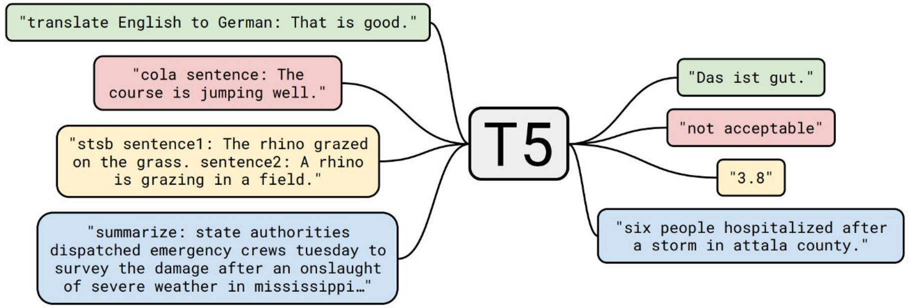
Labeled Data

+

SWAG

SQuAD
The Stanford Question Answering Dataset

GLUE



Computing Power

aws ec2 run-instances \
~~--instance-type p3.2xlarge~~
--instance-type p3.16xlarge

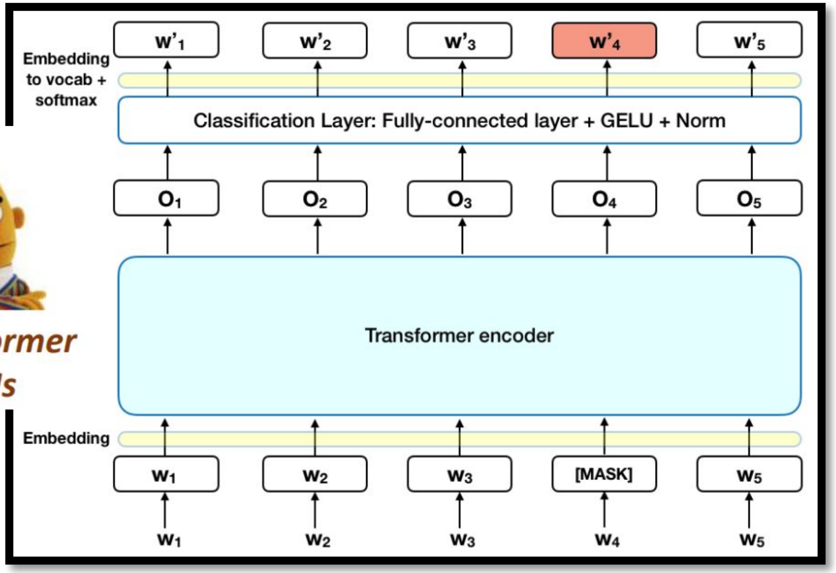
The “pre-train then fine-tune” paradigm for NLP

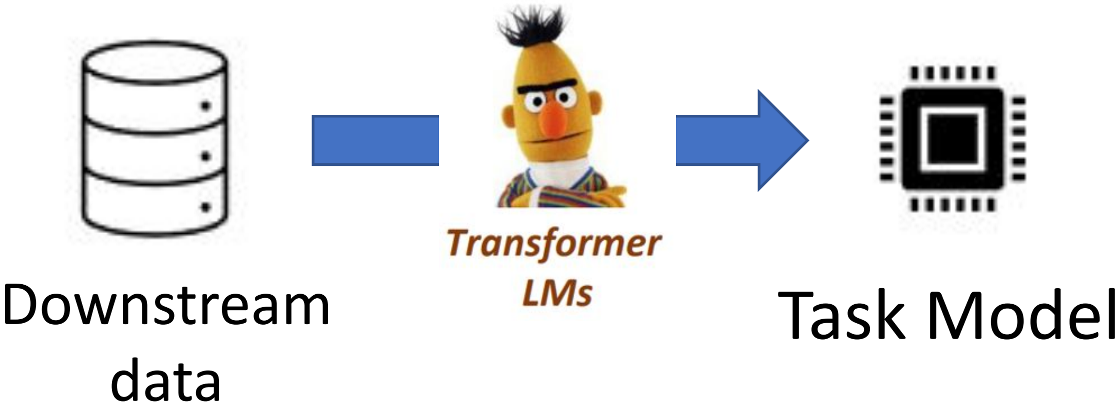
Randomly masked A quick [MASK] fox jumps over the [MASK] dog

Predict A quick brown fox jumps over the lazy dog

WIKIPEDIA
The Free Encyclopedia

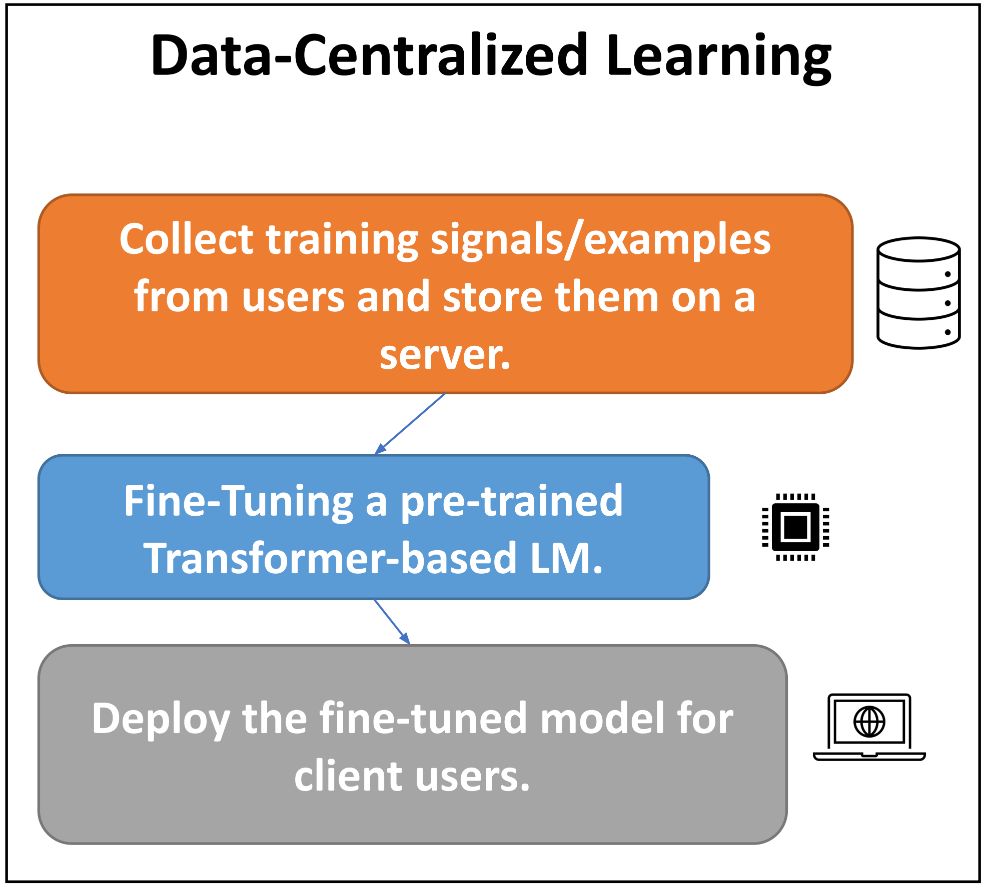
Transformer LMs





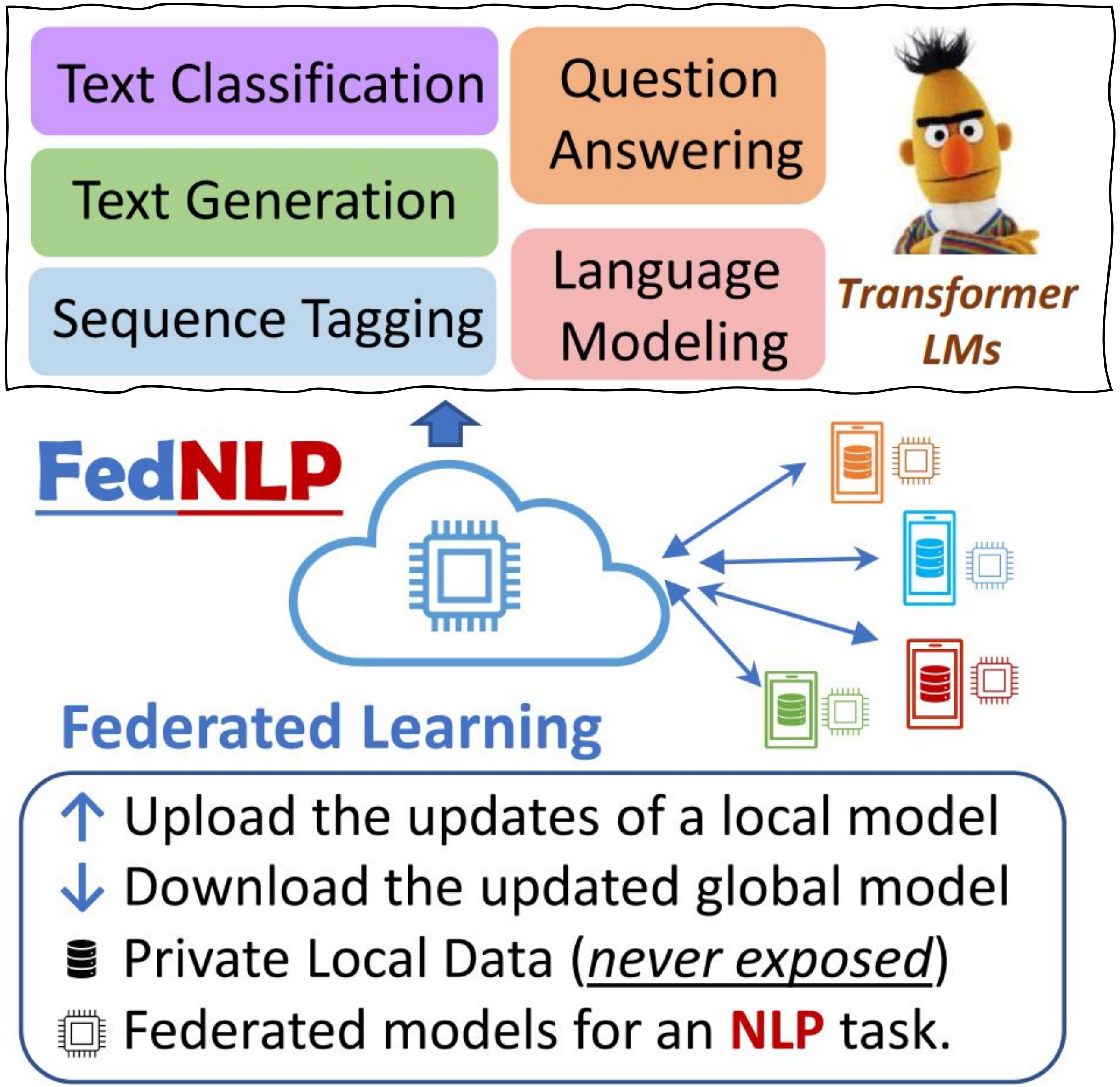
Fine-tuning a pretrained transformer

What happen in workplace..

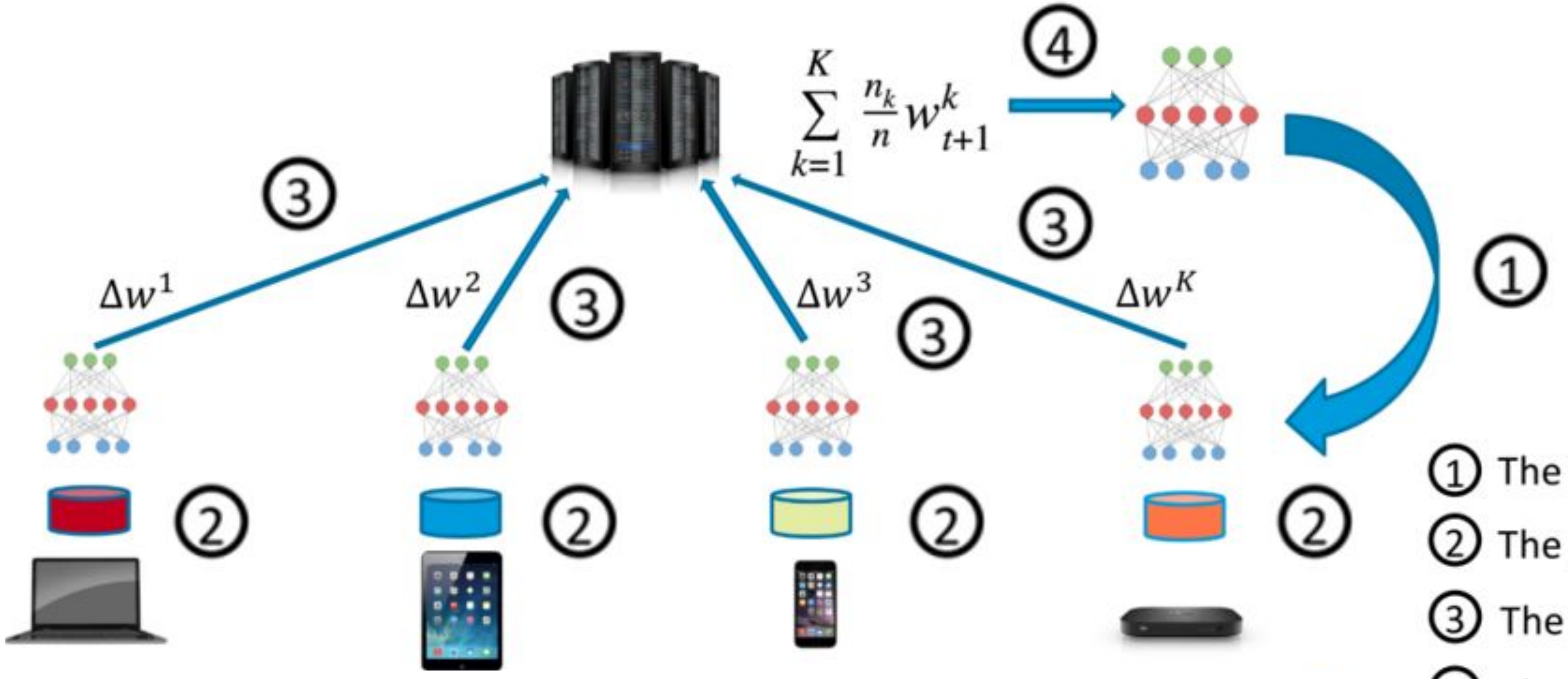


Disadvantages of Data-Centralized Learning

- User Privacy Concerns
- Data Sharing Regulation Laws
- High Cost of Transferring Raw Data
- Expensive Computation of Centralized Training


Text Classification
Text Generation
Sequence Tagging
Question Answering
Language Modeling
Transformer LMs
FedNLP
Federated Learning
Upload the updates of a local model
Download the updated global model
Private Local Data (never exposed)
Federated models for an NLP task.

Federated Learning (FL)



1 The server sends the global model to the parties
2 The parties update models with local data
3 The parties send local models to the server
4 The server updates the global model

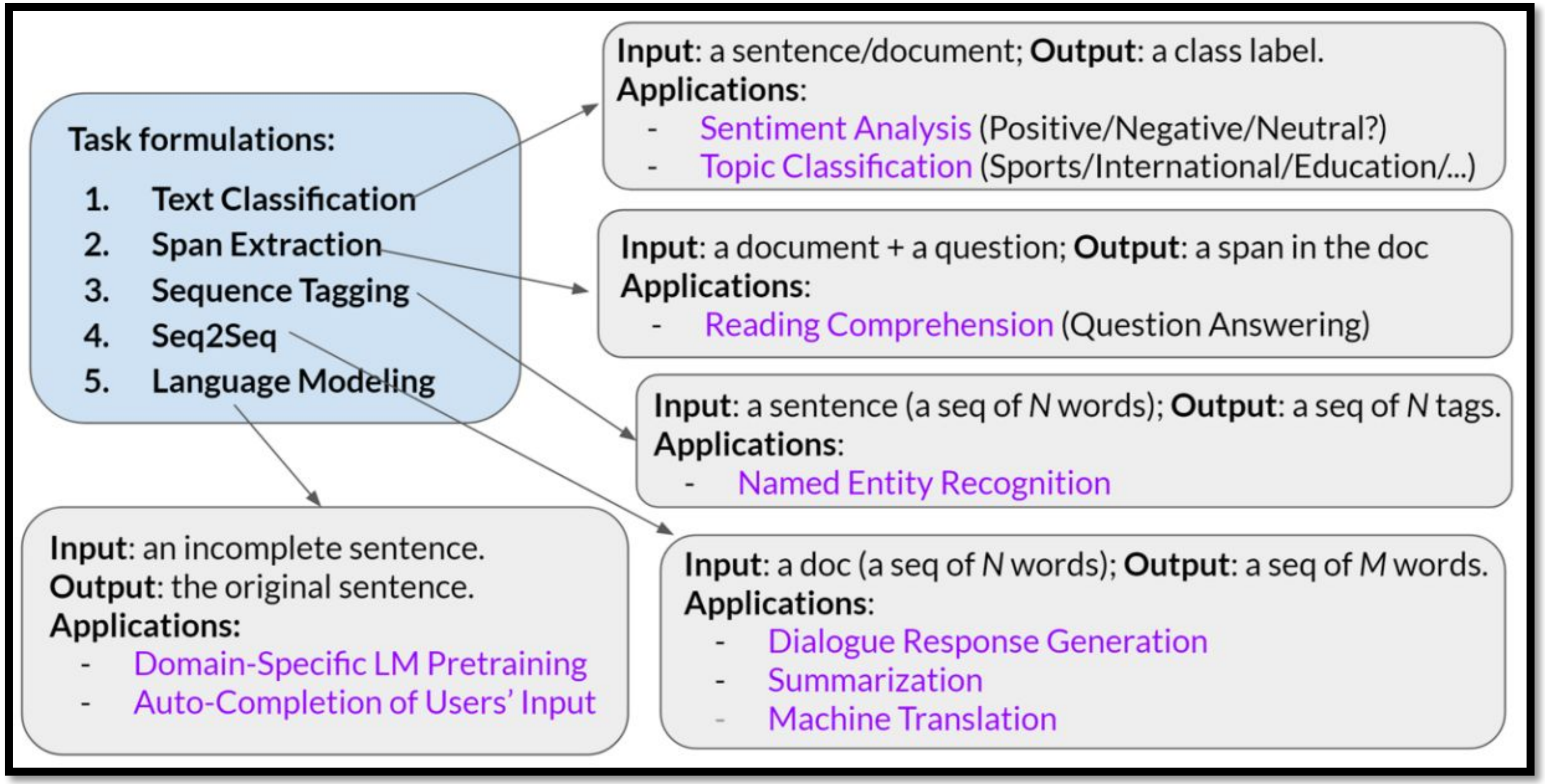
Taking *FedAvg* as an example FL method

the local data is never exposed to others!

Federated Learning for NLP: Promises and Challenges

Background: Current FL research mainly focus on testing methods on *toy* datasets or computer vision (image classification).

Our goal: Provide a universal platform for **benchmarking** and **developing** FL methods for **various** NLP tasks.



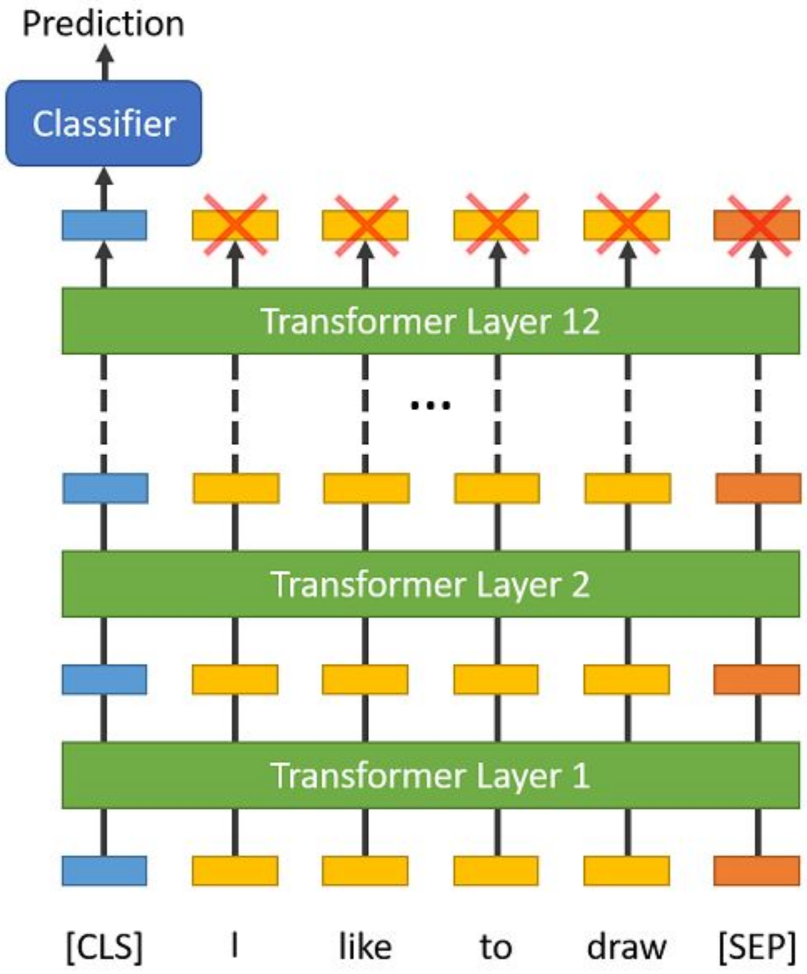
Different NLP tasks have distinct task formulations.

Interface between Transformer LMs and FL

- Most existing FL studies are customized for computer vision tasks/datasets, with specialized model architectures

- Modern NLP are primarily based on pre-training & fine-tuning Transformer LMs.

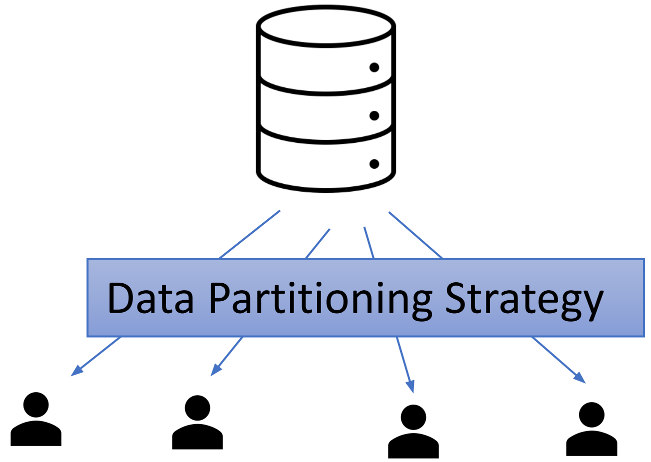
- No existing FL framework connecting Transformer LMs with FL methods.



Creating non-IID datasets for FL in NLP is an open problem

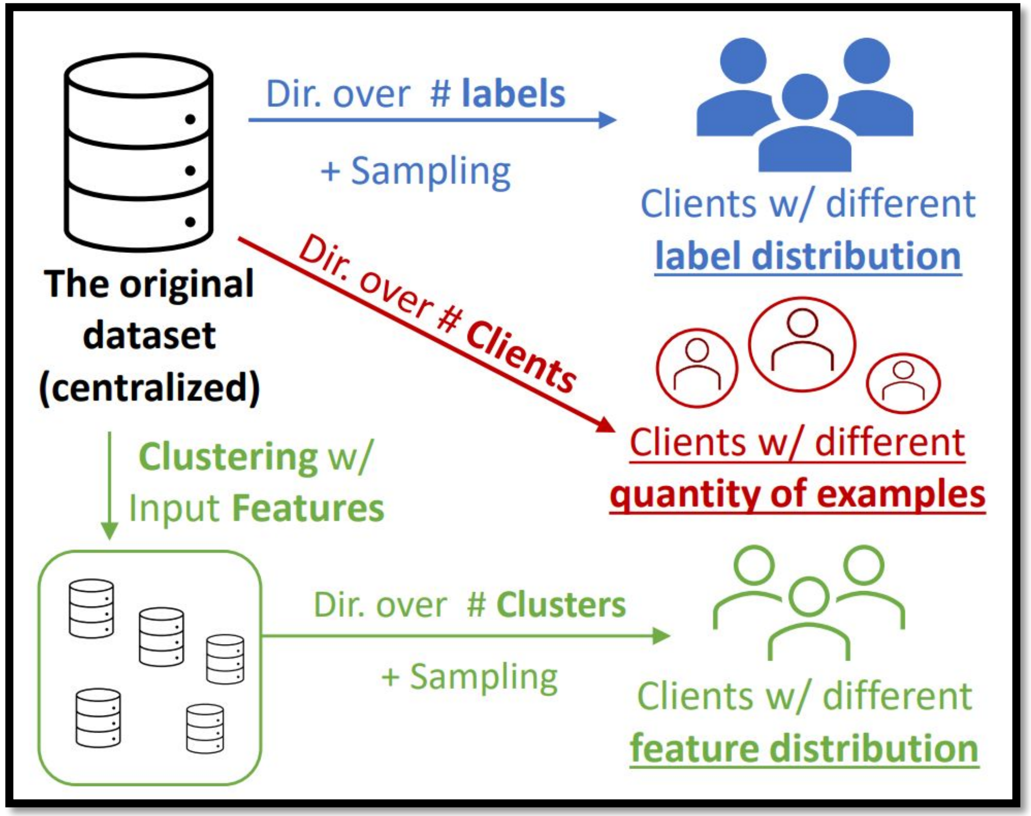
- Current NLP datasets are mainly collected for *centralized* learning, and thus does not have a *natural, non-IID* partition
- Existing FL datasets are mainly for computer vision tasks such as object detection
- Ideal yet not accessible: private user data in large companies; but we cannot make them public for community use

Existing Centralized Dataset

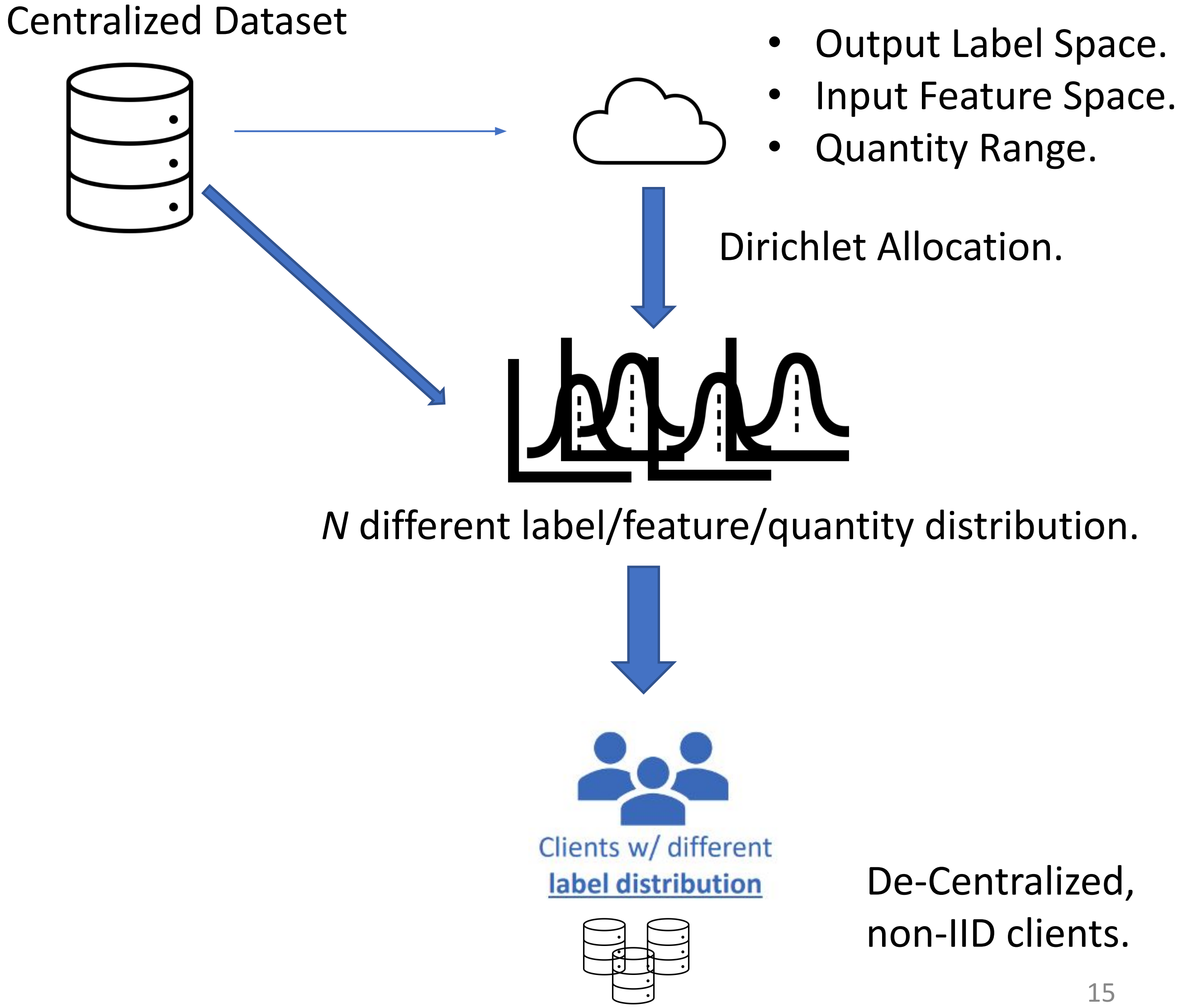


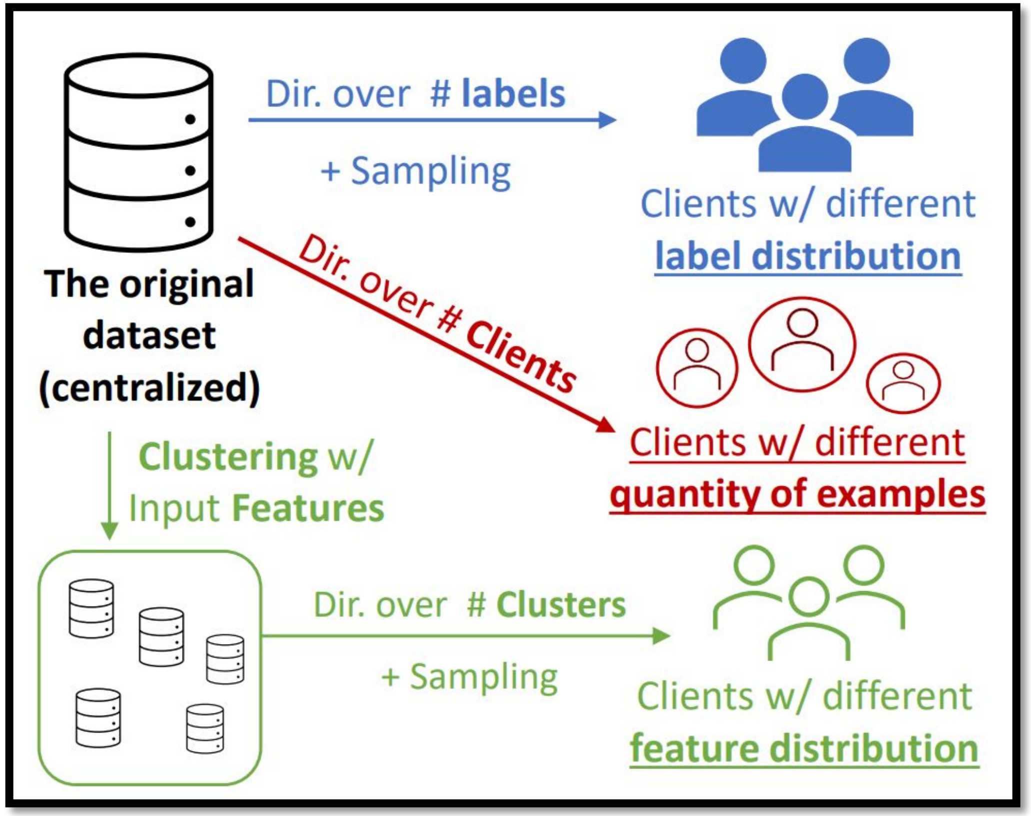
A pool of clients where each client has a relatively unique data distribution.

What happen in workplace..



Three ways to create non-IID data partitions.



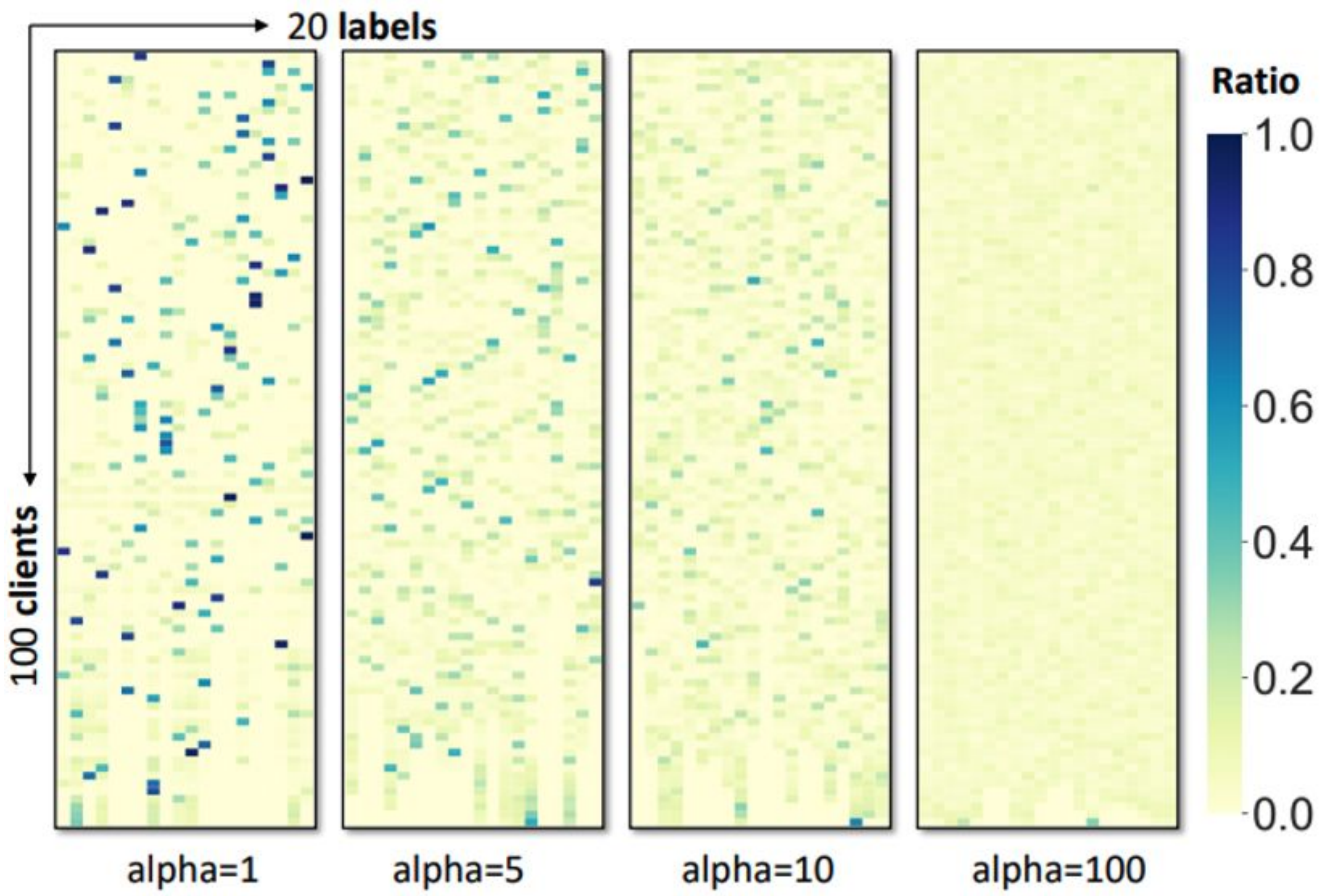


Three ways to create non-IID data partitions.

Task	Txt.Cls.	Seq.Tag.	QA	Seq2Seq
Dataset	20News	Onto.	MRQA	Giga.
# Training	11.3k	50k	53.9k	10k
# Test	7.5k	5k	3k	2k
# Labels	20	37*	N/A	N/A
Metrics	Acc.	F-1	F-1	ROUGE

We select four typical datasets for each formulation.

A showcase of non-IIDness on 20news dataset using *Dirichlet Allocation Methods*



- **Label distribution**: darker color □ higher probability for a client’s data assigned with a certain label
- Smaller alpha □ more distinct label distributions between the clients (non-IID)
- When alpha=100 □ uniform label distribution for all clients

Our Proposed FedNLP framework

- We support a wide range of FL methods such as FedAvg, FedOpt, FedNova, FedProx, etc.
- We implement the interface between these FL methods to Transformer LMs.
  - We support fine-tuning the full model as well as part of the models (last few layers).
  - All task formulations.

Algorithm 1: FEDOPT (Reddi et al. 2020)): A Generic FedAvg Algorithm

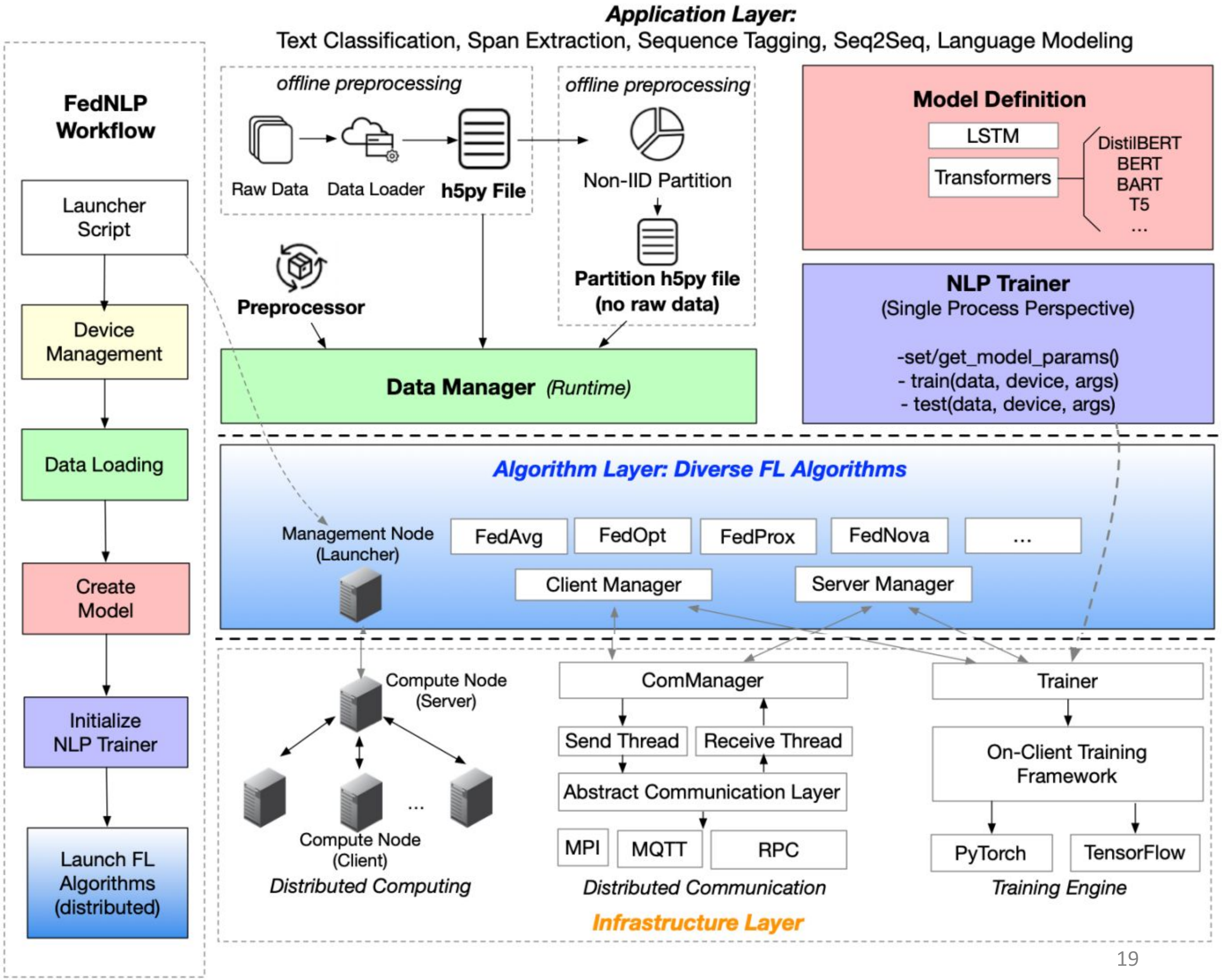
Input: Initial model $\boldsymbol{x}^{(0)}$, CLIENTOPT, SERVEROPT

for $t \in \{0, 1, \dots, T-1\}$ **do**
- Sample a subset $\mathcal{S}^{(t)}$ of clients
- **for client** $i \in \mathcal{S}^{(t)}$ **in parallel do** — On each client.
  - Initialize local model $\boldsymbol{x}_i^{(t,0)} = \boldsymbol{x}^{(t)}$ Download from server.
  - **for** $k = 0, \dots, \tau_i - 1$ **do**
    - Compute local stochastic gradient $g_i(\boldsymbol{x}_i^{(t,k)})$
    - Perform local update $\boldsymbol{x}_i^{(t,k+1)} = \text{CLIENTOPT}(\boldsymbol{x}_i^{(t,k)}, g_i(\boldsymbol{x}_i^{(t,k)}), \eta, t)$
  - Compute local model changes $\Delta_i^{(t)} = \boldsymbol{x}_i^{(t,\tau_i)} - \boldsymbol{x}_i^{(t,0)}$ Upload to server.
- Aggregate local changes $\Delta^{(t)} = \sum_{i \in \mathcal{S}^{(t)}} p_i \Delta_i^{(t)} / \sum_{i \in \mathcal{S}^{(t)}} p_i$
- Update global model $\boldsymbol{x}^{(t+1)} = \text{SERVEROPT}(\boldsymbol{x}^{(t)}, -\Delta^{(t)}, \eta_s, t)$

On the server.

Transformer LMs

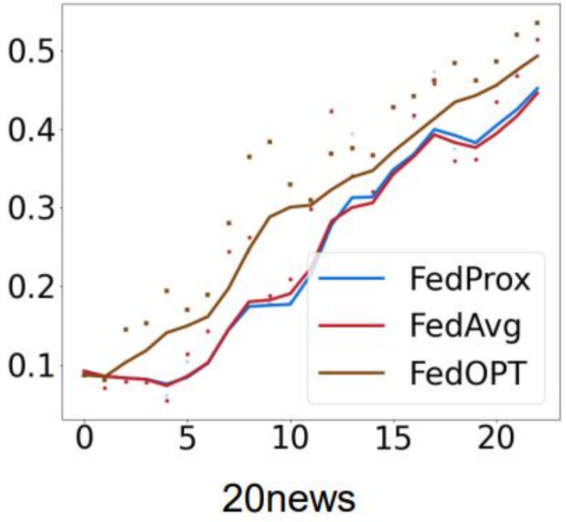
The FedNLP Framework



Q1: How do popular FL methods perform over different NLP tasks?

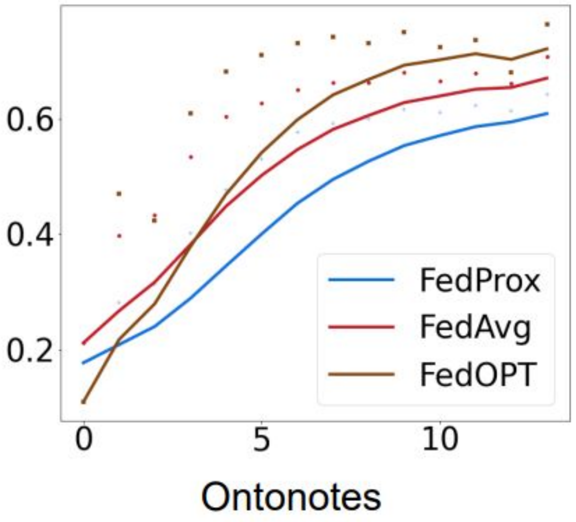
Text Classification

News Classification (acc)



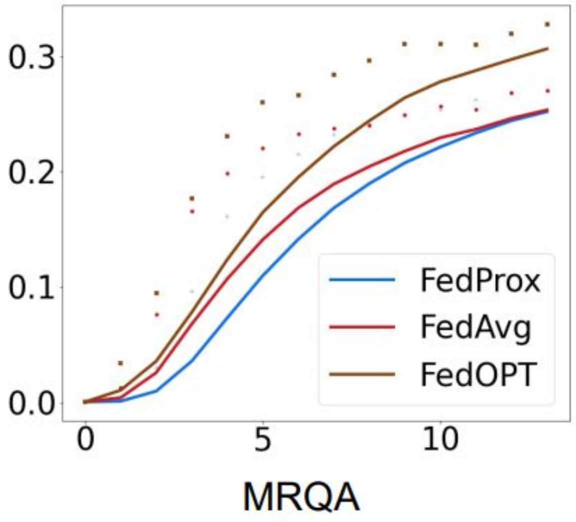
Sequence Tagging

Entity Recognition (span-F1)



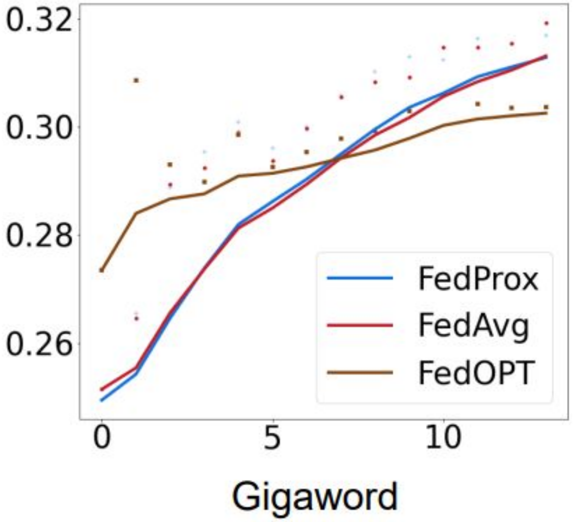
Reading Comprehension

Question Answering (word-F1)



Seq2Seq

Summarization (ROUGE)



Y-axis: evaluation metric of the task. Higher the better.

X-axis: # of iterations in the algorithm

- FedOpt outperforms the other two FL methods in the first three tasks
- FedProx and FedAvg are comparable with each other.

Q2: How do different non-IID partitions of influence FL performance?

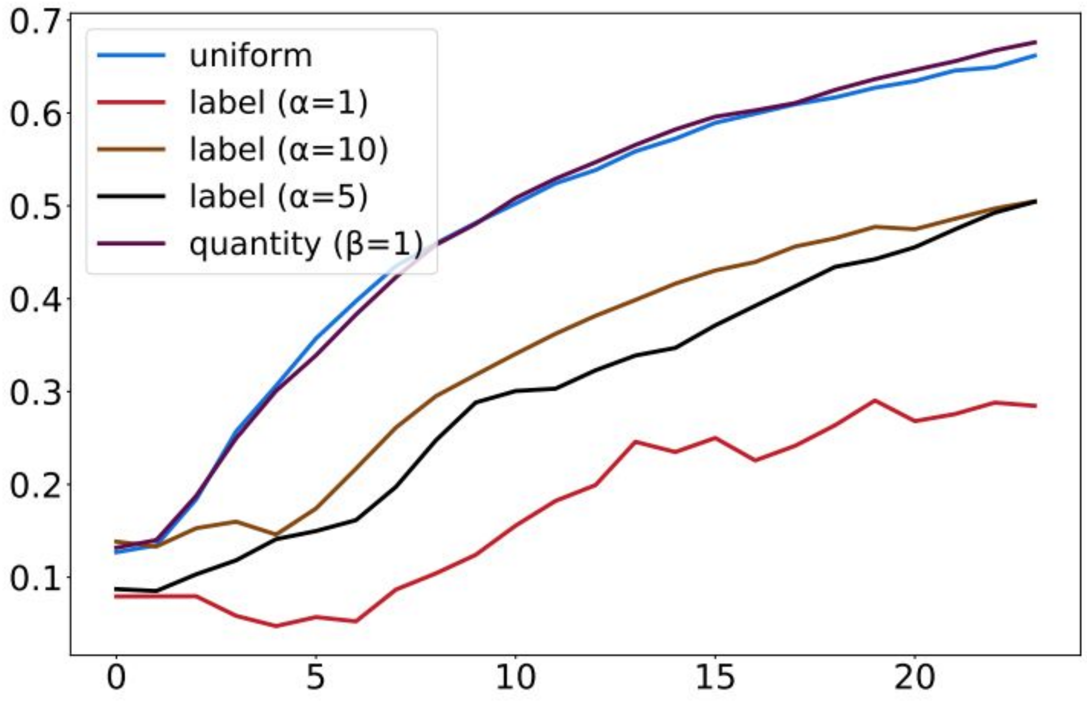
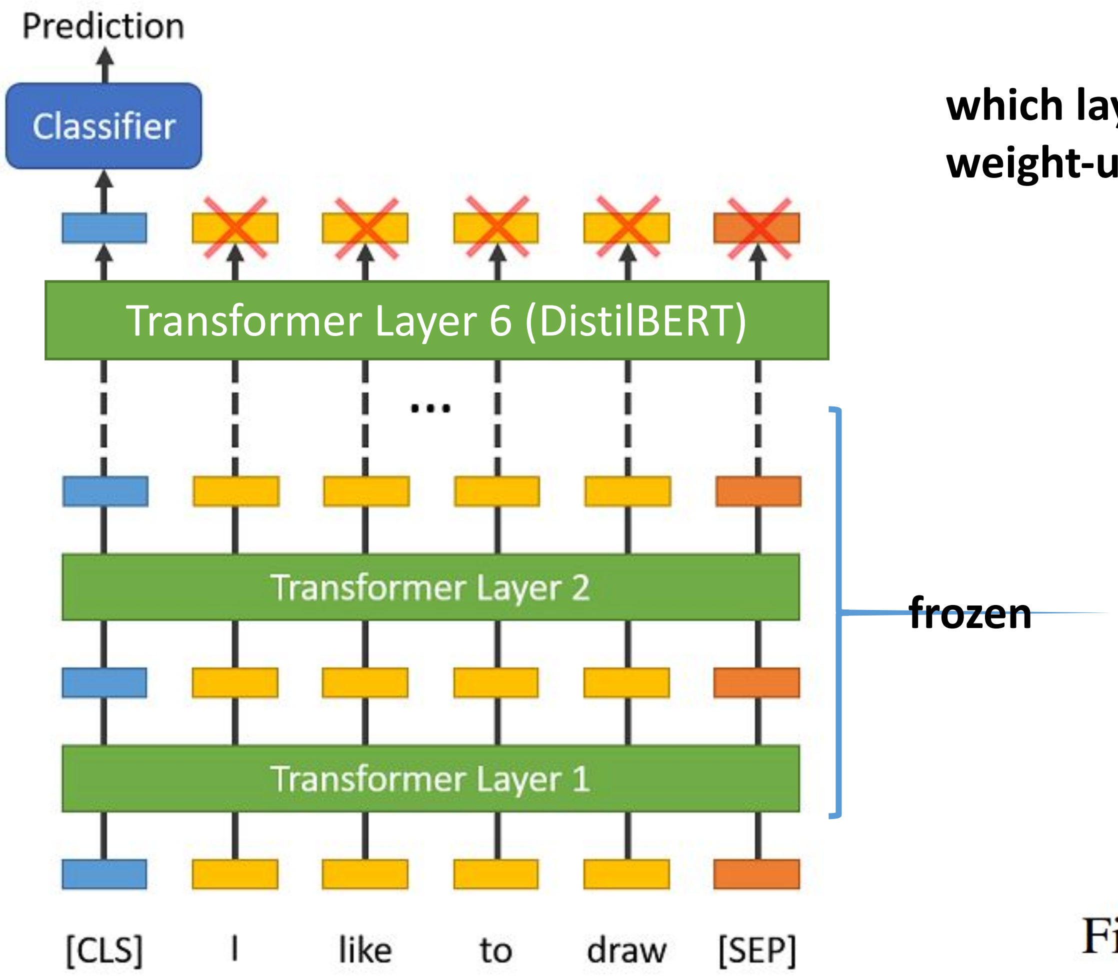


Figure 5: Testing FedOPT with DistilBERT for 20News under different data partition strategies.

- Non-IID data partition (smaller alpha) creates more challenges for FL methods to perform
- Non-IID data partition (smaller alpha) also makes the FL algorithm less stable
- Uniform and quantity-skew partitions are less challenging to learn

Q3: How does freezing of Transformers influence the FL performance?



which layers are frozen and NOT used for weight-uploading/downloading.

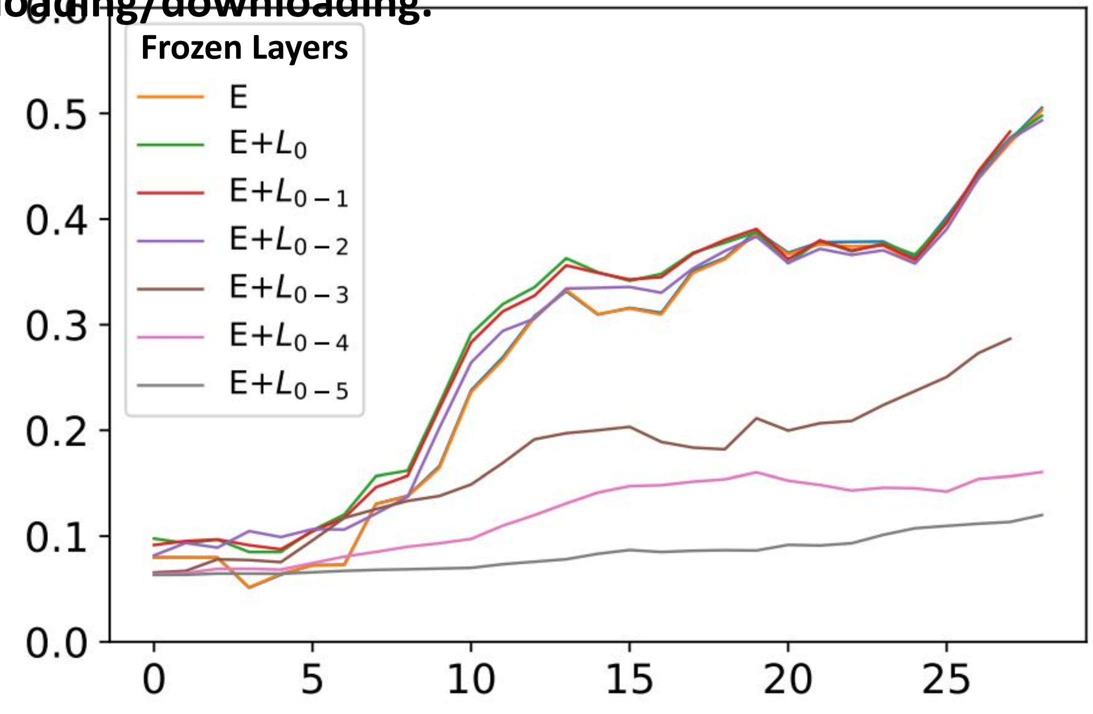


Figure 6: Testing FedOPT with DistilBERT for 20News under different frozen layers.

Q4: Are compact model DistilBERT adequate for FL+NLP?

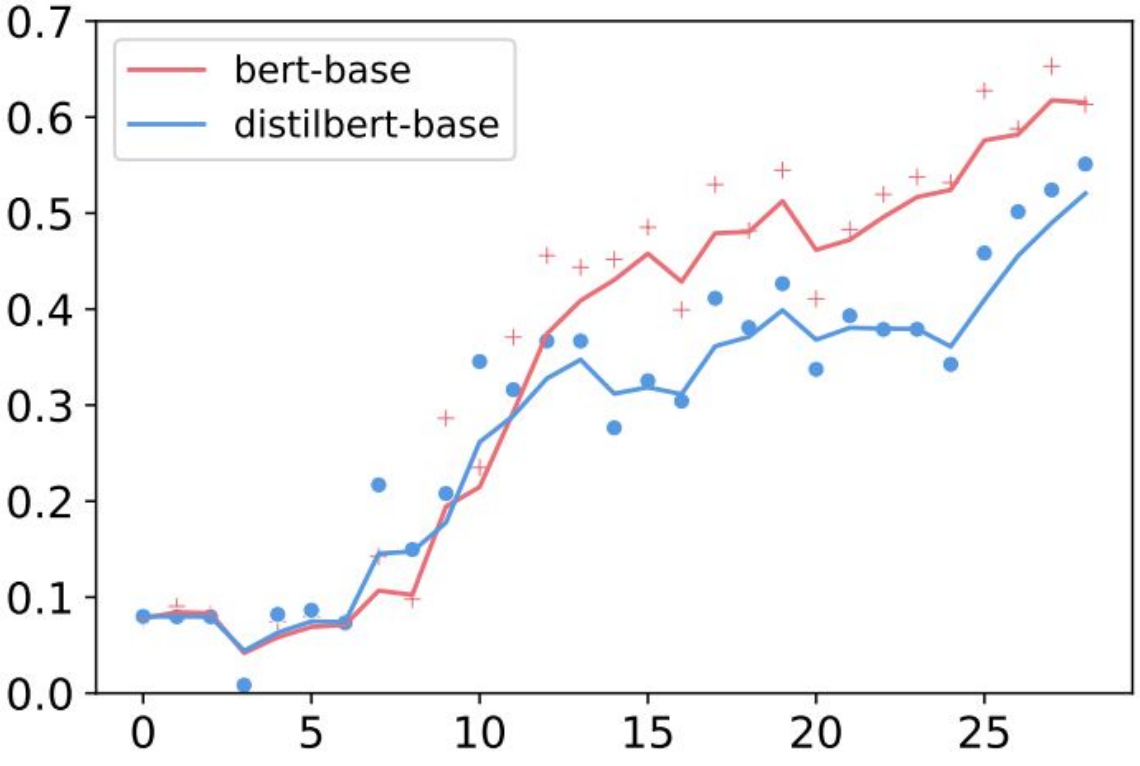
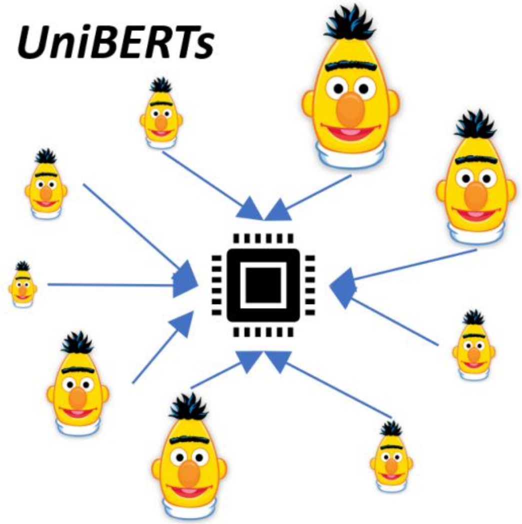


Figure 7: FedOPT for 20News with different LMs.

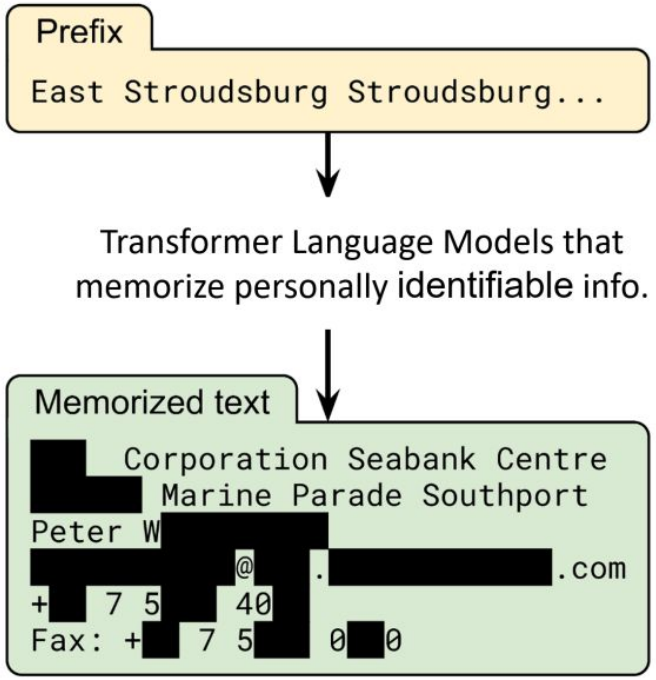
- BERT-base is 2x larger than DistilBERT
- DistilBERT is a more cost-effective choice.
- It's reasonable to do experiments with DistilBERT as the curve is similar to BERT-base.

Future Directions for Extending FL+NLP



Task 1 [heterogenous FedNLP]: Current FL methods focus on the case where all local models are of the same architecture and model size. This is inflexible and can cause problems when users have different devices.

How should we perform FL when clients are using different BERT-style architecture?



Task 2 [privacy]: Concerns about Transformer LMs that can memorize private information.

Can we quantitively measure such data leak? Can we design methods to prevent this?