Enabling Scalable and Trustworthy Collaborative Learning Systems

Huancheng Chen
Advisor: Haris Vikalo
Department of Electrical and Computer Engineering
University of Texas at Austin

1 Background

Modern distributed networks of devices such as mobile phones, wearables and smart home devices produce a vast amount of data used by machine learning models for various prediction/inference tasks. As the computational capabilities of these devices grow, along with concerns about privacy, there is a growing interest in locally storing data and moving network computation to the edge. This has motivated *federated learning (FL)* [1], a popular learning paradigm that enables privacy-preserving collaborative training of machine learning (ML) models across a number of devices by avoiding the need to collect private data stored at those devices. The canonical problem in federated learning involves the task of learning a *universal, global* statistical model under coordination of a central server. Typically, the goal is to minimize the objective function

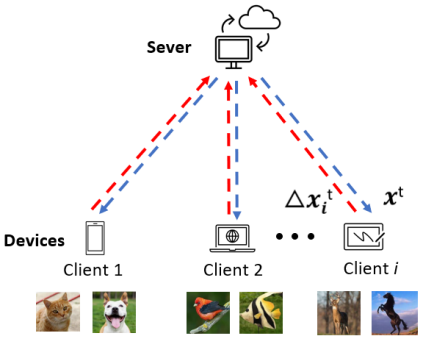


Figure 1: Example of FL workflow.

$$\min_{\mathbf{x}} F(\mathbf{x}) \triangleq \sum_{k=1}^{N} p_k F_k(\mathbf{x}), \tag{1}$$

where N is the total number of devices (referred to as *clients* in the remainder of this report), $\mathbf{x}$ denotes parameters of the global model, $F_k(\mathbf{x})$ is the loss (empirical risk) of the model on k-th client's data $\mathcal{B}_k$, and p_k denotes the weight assigned to client k ($\sum_{k=1}^{N} p_k = 1$). At each global round, only the intermediate model updates are communicated to the central server for aggregation while each clent' data remains private.

A large number of ML tasks in computer visions (CV) and natural language processing (NLP) have been adapted to the federated learning framework, demonstrating the capability of FL to produce highly accurate models by aggregating knowledge from diverse sources. However, several main challenges including **(1) statistical heterogeneity**, **(2) expensive communication**, **(3) systems heterogeneity** and **(4) safety** in distributed learning, adversely affect the performance of FL systems and make it difficult to deploy FL frameworks in realistic settings. The main focus of my research is to develop novel FL algorithms addressing the above four challenges and build scalable and trustworthy collaborate learning systems. In the following, I will shortly introduce my recent work on these topics and outline future research agenda.

2 Research Projects

2.1 Learning Accurate Global Model in Federated Learning with Statistical Heterogeneity

An early FL method, FedAvg [1], performs well in the settings where the devices train on independent and identically distributed (IID) data. However, compared to the IID scenario, training on non-IID data under statistical heterogeneity is detrimental to the convergence speed, variance and accuracy of the learned model. Figure 2 illustrates *objective drift* in non-IID FL manifested through large

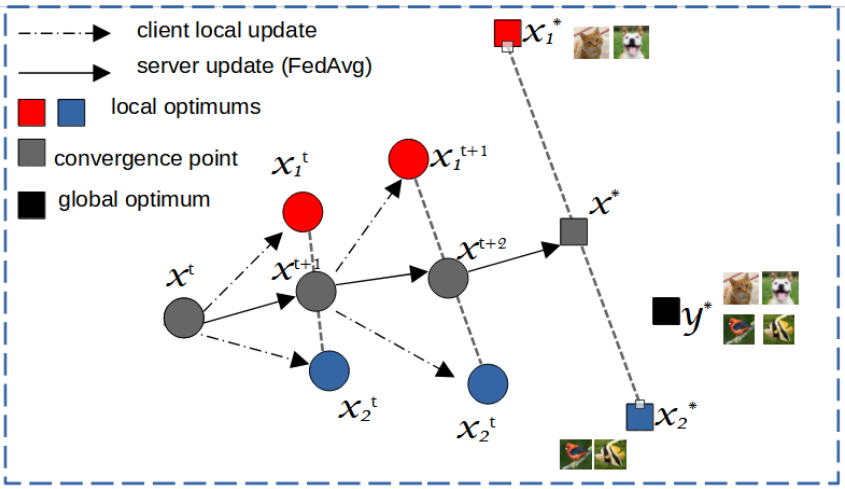


Figure 2: *objective drift*: the local optimal model of client 1 may be very distant from the local optimal model of client 2. Nevertheless, the server that deploys FedAvg still forms the global model by simply averaging the collected local models. Therefore, the global model converges toward the average of the two local optimal models $\mathbf{x}^*$ instead of the true global optimal model $\mathbf{y}^*$.

differences between local models trained on substantially different data distributions. Essentially, objective drift is caused by overfitting the local models due to class imbalance. We studied two strategies aiming to tackle statistical heterogeneity in FL: (1) data augmentation by synthesizing artificial data; (2) adding regularization terms to mitigate objective drift in local training.

In our work [2], we proposed framework FedDMPS where each client utilizes Variational Auto-Encoder (VAE) to generate synthetic data to enrich local dataset and thus ameliorate the detrimental effects of non-iid data distributions. In particular, each client trains a local VAE model (composed of encoder, classifier and decoder) with its local private (potentially class-imbalanced) data. The encoder of the local VAE model is able to extract data representations of raw data in the latent space and compute class-wise data representations by average. The server matches the pairs of clients having complementary local datasets and facilitates differentially-private[3] sharing of class-wise data representations; the clients then deploy the decoder of VAE model to reconstruct artificial data based on these shared data representations. More details of FedDMPS's workflow can be found in Figure 3.

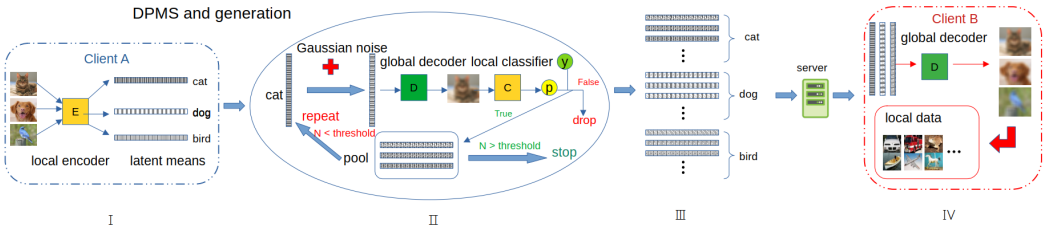


Figure 3: FedDPMS and synthetic data generation. The four parts of the figure depict: (1) finding data representation of raw data via a local encoder; (2) creating noisy means (by adding Gaussian noise to the class-wise means of data representations) and filtering out unusable ones with the help of a local classifier; (3) uploading usable noisy class-wise means to the server; (4) a benefiting client utilizing the decoder to generate synthetic data from the received noisy class-wise means, expanding its local dataset.

Although FedDPMS outperforms state-of-the-art Federated Learning methods on image classification tasks with varied levels of heterogeneity across clients, it does requires additional computation and memory resources (almost double) to train a decoder for each client. In another line of our work,

FedHKD [4], we continue to exploit the shared class-wise data representations as extra information to regularize client's local objective function without introducing significant computation and memory overhead. In particular, we apply the ideas from knowledge distillation [5], improving it in FL settings by removing the need for a public dataset previously required at the server. To be specific, each client k learns a feature extractor $R_k(\cdot)$ and a classifier $G_k(\cdot)$ and uses them to compute the local knowledge – average class-wise features and the corresponding predictions on clients' data – as follows:

$$\mathbf{h}_k^c = \frac{1}{N_k^c} \sum_{\mathbf{s}_i \sim \mathcal{B}_k^c} R_k(\mathbf{s}_i) + \mathcal{N}(0, \sigma_k^2), \forall c \in [C] \tag{2}$$

$$\mathbf{q}_k^c = \frac{1}{N_k^c} \sum_{\mathbf{s}_i \sim \mathcal{B}_k^c} G_k(R_k(\mathbf{s}_i)), \forall c \in [C], \tag{3}$$

where N_k^c denotes the number of samples with label c in client k's local dataset $\mathcal{B}_k$; $\mathcal{B}_k^c$ is a subset of $\mathcal{B}_k$ where all samples have label c; C is the number of classes in the classification task; σ^2 is the predetermined variance of differential privacy (DP) inducing noise [3] that promotes privacy. $\mathbf{h}_k^c$ and $\mathbf{q}_k^c$ are computed and transmitted to the server as the local knowledge of client k, upon which the server aggregates all collected local knowledge into global knowledge defined as

$$\mathcal{H}^c = \sum_{k=1}^{N} p_k \mathbf{h}_k^c, \quad \mathcal{Q}^c = \sum_{k=1}^{N} p_k \mathbf{q}_k^c \tag{4}$$

The entire procedure is shown in Fig. 4.

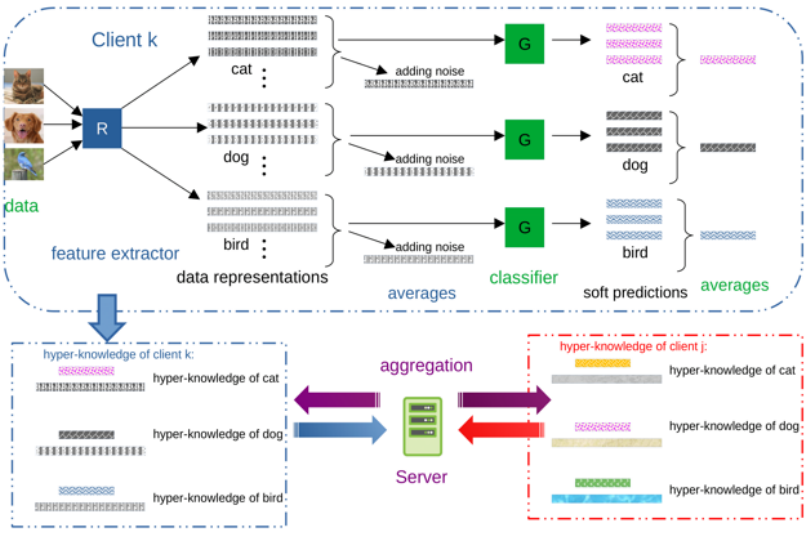


Figure 4: Procedure of computing local knowledge and aggregation.

Following the aggregation at the server, the global knowledge is sent to the clients participating in the next FL round to assist in local training. In particular, given data samples $(\mathbf{x}_i, y_i) \sim \mathcal{B}_k$, the loss function of client k is formed as

$$\mathcal{L}_k = \mathcal{L}_{\text{task}} + \lambda \frac{1}{C} \sum_{c=1}^{C} \left\| G_k(\mathcal{H}^c) - \mathcal{Q}^c \right\|_2 + \gamma \frac{1}{|\mathcal{B}_k|} \sum_{\mathbf{s}_i \sim \mathcal{B}_k} \left\| R_k(\mathbf{s}_i) - \mathcal{H}^{y_i} \right\|_2 . \tag{5}$$

Note that the loss function (5) consists of three terms: the empirical risk of the original task and two regularization terms utilizing global knowledge. Essentially, the second and third terms in the loss function are proximity/distance functions. The second term is to force a local classifier to output similar soft predictions when given global data representations while the third term is to force the features extractor to output similar data representations to the average data representations in the global knowledge when given local data samples.

We analyze convergence of FedHKD and conduct extensive experiments on visual datasets in a variety of scenarios, demonstrating that FedHKD provides significant improvement in both personalized as well as global model performance compared to state-of-the-art FL methods designed for heterogeneous data settings.

2.2 Accelerating Non-IID Federated Learning via Client Selection

Constraints on communication resources might make it unrealistic to require regular contributions to training from all the clients in a large-scale cross-device FL system. Instead, only a fraction of clients participate in any given training round; unfortunately, this further aggravates detrimental effects of statistical heterogeneity. Selecting informative clients in non-IID FL settings is an open problem that has received considerable attention from the research community. However, efficient and effective client selection in FL remains an open challenge, motivating us to develop a novel heterogeneity-guided adaptive client selection method.

If the server were given access to clients' data label distributions, selecting clients would be relatively straightforward. However, privacy concerns typically discourage clients from sharing such information. In our latest work, HiCS-FL [6], we explored the universal property of the gradient of bias of the output layer (see Figure 1) in classification tasks and proposed a novel method to estimate the data label distribution of each client based on the local updates of the output layer's bias. In particular, we assume that the model is trained by minimizing the cross-entropy (CE) loss over one-hot labels – a widely used multi-class classification framework. Relying on the discovered gradient's property, we propose a novel method to estimate the client's data label distribution with theoretical guarantees.

Specifically, we quantify heterogeneity of clients' data by an entropy-like measure $H(\mathcal{D}^{(k)}) \triangleq -\sum_{i=1}^{C} D_i^{(k)} \ln D_i^{(k)}$, where $\mathcal{D}^{(k)}$ denotes label distribution of client k's data. Under the constraint of privacy, the server does not know true $\mathcal{D}^{(k)}$; hence we approximately compute $H(\mathcal{D}^{(k)})$ using the local update $\Delta \mathbf{b}^{(k)}$ from client k,

$$H(\mathcal{D}^{(k)}) \approx \hat{H}(\mathcal{D}^{(k)}) = H(\text{softmax}(\Delta \mathbf{b}^{(k)}, T)), \tag{6}$$

where $\text{softmax}(\Delta \mathbf{b}^{(k)}, T)_i = \exp(\Delta b_i^{(k)}/T)/\sum_{c=1}^{C} \exp(\Delta b_c^{(k)}/T)$, $1 \leq i \leq C$; here T is a scaling hyper-parameter (so-called *temperature*). We provided theoretical analysis and showed that the difference between $H(\mathcal{D}^{(k)})$ and $\hat{H}(\mathcal{D}^{(k)})$ is bounded. Utilizing the approximation method defined above, the server is capable of identifying clients with class-balanced datasets and gives priority to selecting those clients in order to achieve faster convergence and higher test accuracy of the global model.

2.3 Mixed-Precision Quantization for Federated Learning on Resource-Constrained Heterogeneous Devices

In distributed machine learning applications, devices such as mobile phones, wearables and/or smart homes often operate with heterogeneous resources. The clients with lower computation and memory budget may not afford learning local models with the same architecture as the clients with powerful resources. Approaches to customizing model architectures to clients with varied capabilities by pruning and quantization have been subjects of many studies in the FL community. However, existing techniques leave much to be desired. For example, quantization techniques, either Post-Training Quantization (PTQ) or Quantization-Aware Training (QAT), require a full precision model as a teacher of quantized models which can not be applied in FL directly – the clients with low resources cannot run full precision model due to the memory peak constraint. Our latest work aims to address the challenges of quantization in federated learning under heterogeneous source constraints. Specifically, we aim to develop a mixed-precision quantization (MPQ) method that takes into account limitations on each client's resources, without running a full precision model on devices.

To this end we introduce FedMPQ [7], a novel **Fed**erated learning algorithm with **M**ixed-**P**recision **Q**uantization, which enables training of quantized local models within the allocated bit-width budget. FedMPQ first initializes local models as fixed-precision quantized networks that satisfy clients' average bit-width budget, and then converts these quantized networks into a representation that allows bit-level sparsity-promoting training. In particular, while learning local models whose parameters admit binary representation, the clients deploy a group Lasso regularization term which imposes a

trade-off between the task loss and bit-sparsity. The precision of layers that end up having parameters which exhibit higher degree of sparsity is reduced to allow increasing precision of other layers, as illustrated in Figure.5. During the aggregation step, FedMPQ employs the *pruning-growing* strategy where the server aggregates clients' models (locally trained at potentially different bit-widths) to produce a full precision global model. Before transmitting the global model to a client, the bit-width of the model is adjusted to match the client's bit-width budget.

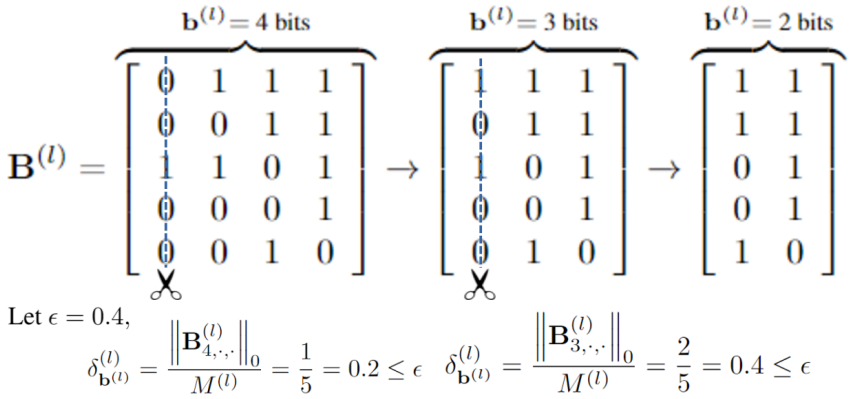


Figure 5: Procedure of bit-level pruning.

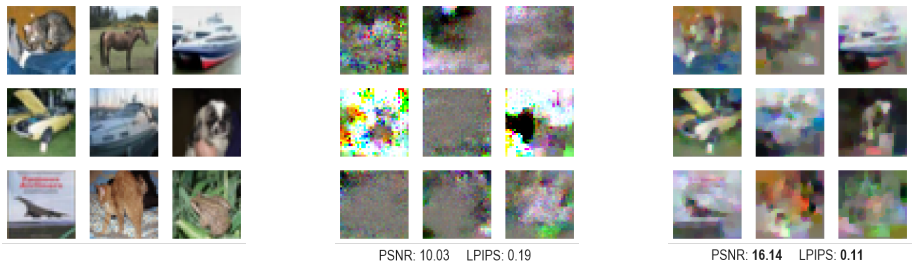
2.4 Recovering Labels from Local Updates in Federated Learning

While clients' privacy is a main motivation for federated learning, recent work [8] shows the possibility of extracting private data from the model updates via gradient inversion attack. This kind of attack can be described as an inverse problem

$$\mathbf{s} = \arg\min_{\mathbf{s},\mathbf{y}} \left\|\nabla_{\mathbf{x}} F_k(\mathbf{s},\mathbf{y}) - \nabla_{\mathbf{x}} F_k(\mathbf{s}^*,\mathbf{y})\right\|^2, \tag{7}$$

where $\mathbf{s}^*$ is the true private data and $\nabla_{\mathbf{x}} F_k(\mathbf{s}^*,\mathbf{y})$ denotes the measurements. Although the original attack in [8] performs well only when the federated learning system involves small batch size (for instance, 2 or 4), the subsequent studies improved the gradient inversion attack for larger batch sizes under some strong assumptions on an untrained model. To explore potential vulnerabilities of privacy in federated learning, our latest work is developing a novel label recovery scheme, *Recovering Labels from Local Updates* (RLU) [9], which provides near-perfect accuracy when attacking untrained (most vulnerable) models and achieves comparable performance even given a well-trained model.

More significantly, RLU achieves high performance even in realistic real-world settings where the clients in an FL system run multiple local epochs, train on heterogeneous data, and deploy various optimizers to minimize different objective functions. Specifically, RLU estimates labels by solving a least-square problem that emerges from the analysis of the correlation between labels of the data



(a) Ground Truth (b) IG (c) Improved by RLU

Figure 6: Batch image reconstruction (batch size set to 9) on CIFAR10 compared to IG [8]. We select the best reconstructed batch for visualization and display the average metrics of the selected batches.

points used in a training round and the resulting update of the output layer. The experimental results on several datasets, architectures, and data heterogeneity scenarios demonstrate that the proposed method consistently outperforms existing baselines, and helps improve quality of the reconstructed images in GI attacks in terms of both PSNR and LPIPS, as illustrated in Fig. 6.

3 Low-Rank Adaptation for Continual Learning with Pre-Trained Models

In the era of foundation models, we revisit continual learning (CL), which aims to enable pre-trained vision transformers (ViTs) to learn new tasks over time. However, as the scale of these models increases, catastrophic forgetting remains a more serious challenge, particularly in the presence of significant domain shifts across tasks. Recent studies highlight a crossover between CL techniques and parameter-efficient fine-tuning (PEFT), which focuses on fine-tuning only a small set of trainable parameters to adapt to downstream tasks, such as low-rank adaptation (LoRA). While LoRA achieves faster convergence and requires fewer trainable parameters, it has seldom been explored in the context of continual learning.

To address this gap, we propose a novel PEFT-CL method called Dual Low-Rank Adaptation (DualLoRA) [10], which introduces both an orthogonal LoRA adapter and a residual LoRA adapter parallel to pre-trained weights in each layer. These components are orchestrated by a dynamic memory mechanism to strike a balance between stability and plasticity. The orthogonal LoRA adapter's parameters are updated in an orthogonal subspace of previous tasks to mitigate catastrophic forgetting, while the residual LoRA adapter's parameters are updated in the residual subspace spanned by task-specific bases without interaction across tasks, offering complementary capabilities for fine-tuning new tasks. On ViT-based models, we demonstrate that DualLoRA offers significant advantages in accuracy, inference speed, and computation efficiency over existing CL methods across multiple benchmarks.

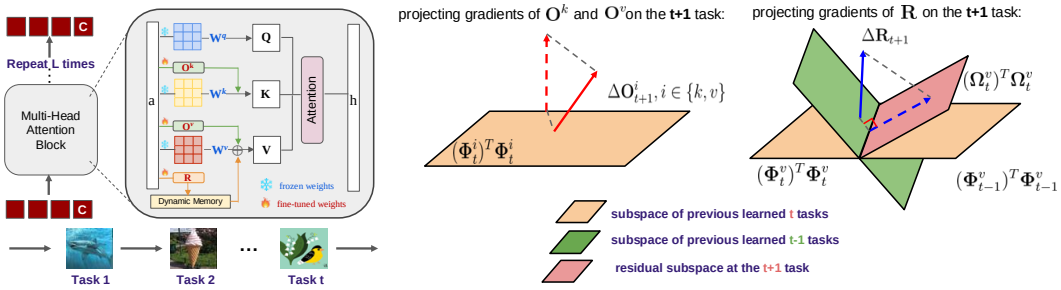


Figure 7: Illustration of our proposed DualLoRA paradigm (left) and design insights of orthogonal adapter and residual adapter (right), where the solid arrow denotes the original update and the dashed arrow denotes the projected update.

4 Boundary Attention Constrained Zero-Shot Layout-To-Image Generation

Recent text-to-image diffusion models excel at generating high-resolution images from text but struggle with precise control over spatial composition and object counting. To address these challenges, several studies developed layout-to-image (L2I) approaches that incorporate layout instructions into text-to-image models. However, existing L2I methods typically require either fine-tuning pretrained parameters or training additional control modules for the diffusion models.

In this work [11], we propose a novel zero-shot L2I approach, BACON (Boundary Attention Constrained generation), which eliminates the need for additional modules or fine-tuning. Specifically, we use text-visual cross-attention feature maps to quantify inconsistencies between the layout of the generated images and the provided instructions, and then compute loss functions to optimize latent features during the diffusion reverse process. To enhance spatial controllability and mitigate semantic failures in complex layout instructions, we leverage pixel-to-pixel correlations in the self-attention feature maps to align cross-attention maps and combine three loss functions constrained by boundary attention to update latent features. Comprehensive experimental results on both L2I and non-L2I

pretrained diffusion models demonstrate that our method outperforms existing zero-shot L2I techniuqes both quantitatively and qualitatively in terms of image composition on the DrawBench and HRS benchmarks.

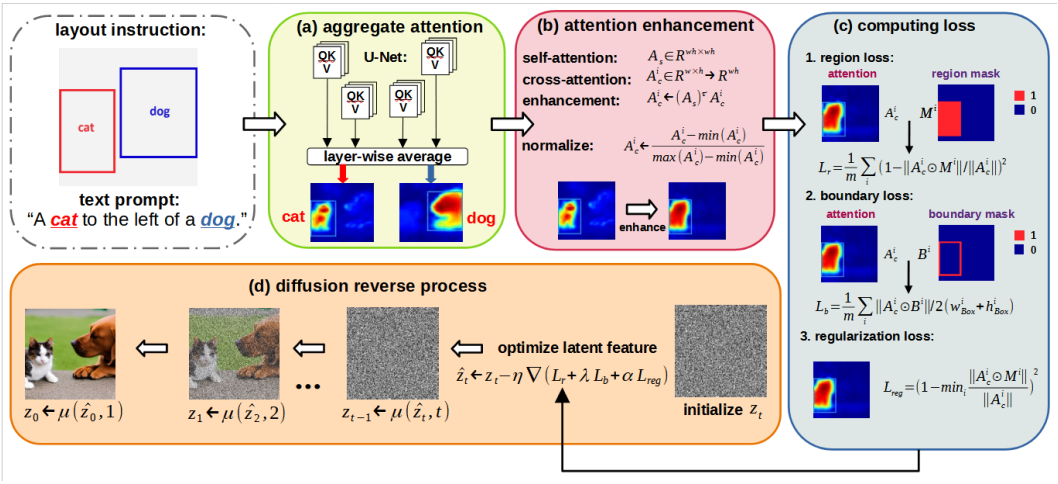


Figure 8: The overall framework of BACON for zero-shot L2I generation consists of four steps: (a) aggregating cross-attention and self-attention maps obtained from the language-vision fusion blocks by layer-wise averaging; (b) enhancing cross-attention maps with self-attention maps; (c) computing three losses and combining them with coefficients; (d) optimizing z_t by gradient descent in diffusion reverse process.

Figure 9: An illustration of zero-shot layout-to-image (L2I) generation using various diffusion models. Text prompts and layout information, where specific concepts are localized within corresponding bounding boxes, are provided as input to the pretrained diffusion models to generate images that align with the instructions.

5 Future Directions

In addition to the topics outlined in the previous section, there are other interesting direction that I would like to explore in my research. These include:

- **Robust Aggregation in Federated Learning.** The vanilla Federated Learning, FedAvg, aggregates local models into the global model by naively computing their weighted average. Such simple aggregation is vulnerable to Byzantine Clients who are providing detrimental local updates or conducting backdoor attack. It is critical to develop more robust aggregation methods to deploying FL systems in real work.

- **Federated Continual Learning.** Continual learning (CL) is an extensively researched subject focused on developing a comprehensive model capable of adapting to new data domains through fine-tuning while maintaining optimal performance on previously encountered data domains. However, privacy constraints in federated learning pose challenges for directly applying schemes from continual learning, such as data replay or knowledge distillation. Furthermore, data heterogeneity among clients in federated learning exacerbates the problem of "catastrophic forgetting" observed in continual learning. In this setting, I am interested in pursuing development of effective continual learning frameworks for privacy-preserving federated learning systems.

References

[1] 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. PMLR, 2017.

[2] **Huancheng Chen** and Haris Vikalo. Federated learning in non-iid settings aided by differentially private synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5026–5035, June 2023.

[3] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

[4] **Huancheng Chen**, Chaining Wang, and Haris Vikalo. The best of both worlds: Accurate global and personalized models through federated learning with data-free hyper-knowledge distillation. In *The Eleventh International Conference on Learning Representations*, 2023.

[5] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[6] **Huancheng Chen** and Haris Vikalo. Heterogeneity-guided client sampling: Towards fast and efficient non-iid federated learning. *Conference on Neural Information Processing Systems*, 2024.

[7] **Huancheng Chen** and Haris Vikalo. Mixed-precision quantization for federated learning on resource-constrained heterogeneous devices. *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2024.

[8] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33:16937–16947, 2020.

[9] **Huancheng Chen** and Haris Vikalo. Recovering labels from local updates in federated learning. In *Forty-first International Conference on Machine Learning*, 2024.

[10] **Huancheng Chen**, Jingtao Li, Nidham Gazagnadou, Weiming Zhuang, Chen Chen, and Lingjuan Lyu. Dual low-rank adaptation for continual learning with pre-trained models. *arXiv preprint arXiv:2411.00623*, 2024.

[11] **Huancheng Chen**, Jingtao Li, Weiming Zhuang, Haris Vikalo, and Lingjuan Lyu. Boundary attention constrained zero-shot layout-to-image generation. *arXiv preprint arXiv:2411.10495*, 2024.