國立臺灣大學電機資訊學院資訊工程學系
碩士論文

Department of Computer Science and Information Engineering
College of Electrical Engineering and Computer Science
National Taiwan University
Master Thesis

FedUA：一種適用於圖像分類的可感知不確定性的蒸餾式聯邦學習方案

FedUA : A Uncertainty-Aware Distillation Based Federated Learning Scheme for Image Classification

李紹銘

SHAO-MING Lee

指導教授：吳家麟 博士

Advisor: Ja-Ling Wu, Ph.D.

中華民國 111 年 8 月

August, 2022

doi:10.6342/NTU202202304

口試委員審定書

國立臺灣大學碩士學位論文
口試委員會審定書

FedUA：一種適用於圖像分類的可感知不確定性的蒸餾式聯邦學習方案

FedUA: A Uncertainty-Aware Distillation Based Federated Learning Scheme for Image Classification

本論文係李紹銘君（學號 R08922142）在國立臺灣大學資訊工程學系完成之碩士學位論文，於民國 111 年 8 月 10 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

（指導教授）

系主任

doi:10.6342/NTU202202304

致謝

這論文的完成，對我來說別具意義。曾經以為，這是人生中一段乏味的旅途，但卻意外讓我收穫滿滿。在研究所的就讀期間，歷經許多人生轉折，也讓研究的產出更加艱辛。從毫無頭緒到做出無愧的碩士畢業論文，所有的焦慮、困頓都是我未曾體會過的。不僅僅是在學術專業上的成長，也鍛鍊了我的韌性。

首先，要感謝家麟老師的指導和鼓勵。此外，也感謝學長姊們無私的幫助，在我所提的方法或實驗遇到瓶頸時，得以迎刃而解。還有實驗室的大家打理許多事務和分享各自研究的內容。

doi:10.6342/NTU202202304

摘要

近年來，聯邦式學習（Federated Learning）逐漸變成資訊領域重要的研究課題。而聯邦式學習強調學習任務是由鬆散的設備（或稱為用戶端）合力共同解決，此情境下存在的重要挑戰包括：用戶端間的資料是不平衡且非獨立同分布的 (Non-IID)，並且設備間的溝通受限於有限的傳輸頻寬而不可靠。上述議題對於聯邦式學習是棘手的挑戰。本文從深度神經網路的不確定性切入聯邦式學習的效能評估並提出新的模型聚合架構（Model Aggregation）。此架構是以知識蒸餾（Knowledge Distillation）為基礎佐以量化深度神經網路的不確定性（Uncertainty in DNN）相關評估方法，進而提升學習效果。實驗在圖像分類（Image Classification）的任務上，證實我們提出的模型聚合架構可有效的解決非獨立同分布的問題，尤其在限制傳輸成本的狀態下，有不錯的表現。

關鍵字：聯邦式學習，模型聚合，知識蒸餾，深度神經網路的不確定性。

doi:10.6342/NTU202202304

Abstract

In recent years, Federated Learning has gradually become an important research topic in the field of information theory. Federated learning emphasizes that learning tasks are jointly solved by loose devices (or clients). Important challenges in this situation include: the data among clients is unbalanced and non-IID, and the communication between devices is unreliable due to limited transmission bandwidth. The above issues are intractable to federated learning. This paper starts from the uncertainty of deep neural network to evaluate the effectiveness of federated learning and proposes a new model aggregation architecture. This scheme is based on knowledge distillation and quantifies the uncertainty in DNN related evaluation methods of deep neural networks to improve the learning performance. The experiments on the task of image classification confirm that our proposed model aggregation scheme can effectively solve the problem of non-IID data distribution, especially when the transmission cost is limited, and it has a good performance.

Keywords: Federated Learning, Model Aggregation, Knowledge Distillation, Uncertainty in Deep Neural Networks.

doi:10.6342/NTU202202304

目　錄

doi:10.6342/NTU202202304

圖目錄

doi:10.6342/NTU202202304

表目錄

doi:10.6342/NTU202202304

Chapter 1 Introduction

1.1 Non-IID issues in Federated Machine Learning

聯邦式學習（Federated Learning）的概念於 2016 年由 McMahan 等人提出[1]。其目標為當資料分布在不同的裝置且保有資料隱私性的條件下，可完成全域模型(global model) 的訓練。此外，該論文亦提出 Federated Averaging 演算法來完成全域之聚合任務，使用戶端可在各自保有本地的資料的前提下，順利完成模型的訓練。FedAvg 避免用戶上傳用戶端敏感的資料到伺服器端，改以上傳用戶端模型的梯度（gradients），再以此進行聚合成全域模型，達到保護隱私（privacy protection）和資料安全性（data security）的目的。

儘管 FedAvg 宣稱能夠處理資料非獨立同分布(Non-IID)問題，但許多研究指出在面臨非獨立同分布的情況下，其準確度會嚴重退化[2,3]，而原因主要來自非獨立同分布的資料會導致本地模型們的權重產生發散。更精確地說，在聯邦式訓練過程中，由於神經網路的損失函數為非凸性(non-convex)，FedAvg 取均值而得到的全域模型會與訓練在獨立同分布而得的理想模型（ideal model）持續增加差距，進而使得整體無法收斂並且讓學習表現惡化[4]。此外，FedAvg 也未能夠完全善用用戶端提供的所有資訊，例如梯度變異。

目前主要處理資料非獨立同分布(Non-IID)問題的方法可大略分為:基於資料（Data-Based）、系統(System-Based)及演算法(Algorithm-Based)三種類別[5]。基於資料類別的方法是透過資料共享[3,6]（data sharing）或者資料擴充[7]（data augmentation）直接而有效的解決問題，然而，這類的方法往往違反聯邦式學習的精神，因無法實踐資料去中心化而產生資料隱私洩露的風險。基於系統的方法則通常使採以分群（clustering）將用戶端進行類聚以建構多中心框架[8,9]，同群的用戶會有較相似的訓練資料。而其資料相似度（data similarity）的估算方法又進一步分為：估算損失值的相似程度及估算用戶端模型權重的相似程度兩種。基於演算法的方法其具體實現方式則非常多元，包含用戶端訓練時引入正規化[10,11]（regularization）、微調[12]（fine-tuning）、個性化層[13]

doi:10.6342/NTU202202304

（personalization layer），及一些機器學習常見的技術。例如，多任務學習[14]（multi-task learning）、終身學習[15]（lifelong learning）和知識蒸餾[16-20]（knowledge distillation）等手段。

1.2 Distillation-Based FL

Guha 等人提出 DOSFL[16]為“一次完成”的 FL 架構。不同於模型蒸餾手法，該架構使用資料集蒸餾(Dataset Distillation)方法：用戶端對本地資料進行蒸餾，並將合成資料和學習率上傳給伺服器。伺服器將來自用戶們的合成資料做合併後，訓練出全域模型。Jeong 等人提出 FD 架構[17]：用戶上傳的是各標籤的對數機率均值(per-label mean logit vectors)以加速通訊。此外，針對非獨立且同分佈問題，再使用 FAug 做處理。FAug 會讓所有用戶告知伺服器自己所缺乏的樣本，讓伺服器訓練一個 GAN，而後用戶下載該 GAN 以擴充自己的本地資料成獨立且同分佈型態。然而 FD 和 DOSFL 相較於 FedAvg，雖然都能在通訊成本大幅降低，但準確率表現卻些許不佳。Li 等人提出的架構 FedMD[18]：需要一個公開資料集(public dataset)，用戶先用公開資料集進行訓練，再以本地的私人資料做客製化訓練。在通訊階段，用戶上傳的是對公開資料集上做推算而得的對數機率，伺服器再將所有用戶上傳的對數機率取平均後進行學習。相較於 FedMD，Lin 等人提出的 FedDF[19]則是採用未標籤資料做蒸餾，並將蒸餾任務從用戶端轉移到伺服器端。結果顯示 FedDF 對蒸餾資料集的選擇有更好的穩固性，進一步適用於聯邦式學習的情境。Chen 等人提出 FedBE[20]，其基於 FedDF 的架構引入貝式推導 (Bayesian Inference)觀點採樣出更多模型並藉貝式集成(Bayesian Ensemble) 得到更好的全域模型。FedBE 證實對非獨立且同分佈問題有解決之效力，且可相容於對用戶端模型做正規化的其他架構。

doi:10.6342/NTU202202304

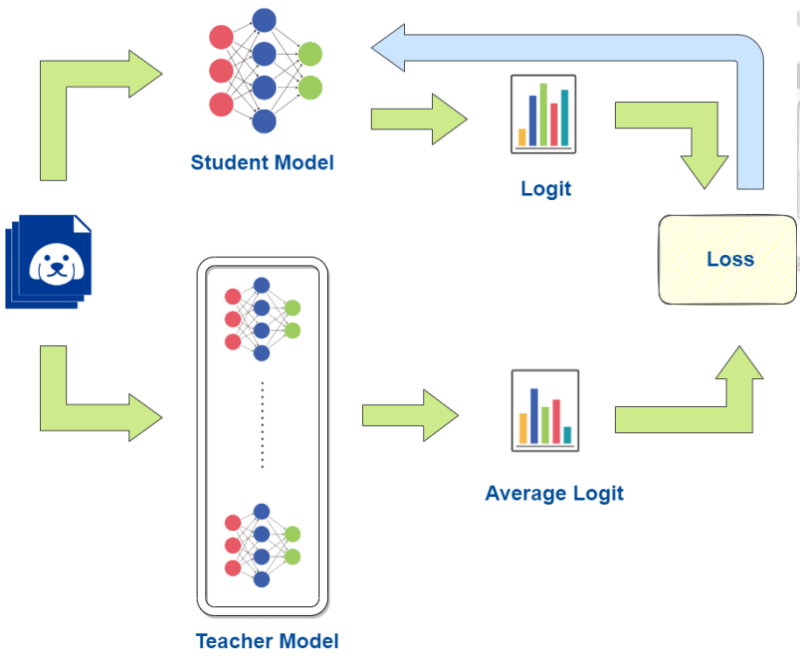

圖 1. FedDF 的示意圖

1.3 Possible contributions of this thesis

本文所提出的架構是基於知識蒸餾的聯邦學習。值得注意的可能貢獻包括：

1. 伺服器端量化上傳用戶端之網路不確定性，作為聚合模型的依據。
2. 伺服器端引入樣本評估，提高學習的效能。
3. 針對知識蒸餾的聚合架構，有效分離出不確定性和類間關係的資訊。
4. 有效的解決資料非獨立同分布的問題，且在限制傳輸成本的狀態下有不錯的學習表現。

doi:10.6342/NTU202202304

Chapter 2 Background

2.1 Knowledge Distillation

起初，知識蒸餾的提出是被應用於模型壓縮[21]（model compression），目標將一個或多個大模型（teacher model）壓縮到小模型（student model），使小模型有效學習預先訓練好的大模型中重要的知識，進而使其能夠保有一定的效能。通常知識蒸餾可使的小模型具備更好的泛化（generalization）能力。在分類問題上，可以藉由調控蒸餾溫度（temperature）的機制有效學習到類間關係（inter-class relations），前者又稱之為暗知識（dark knowledge），請參見圖2。

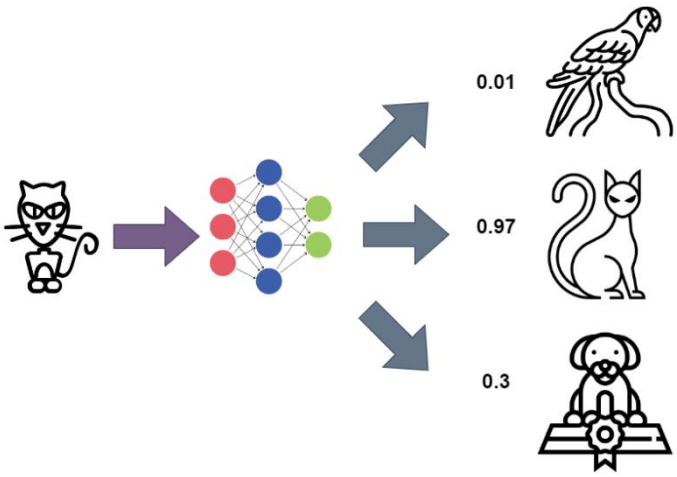

圖 2. 知識蒸餾藉由教師模型學習到暗知識

知識蒸餾對於聯邦學習而言是有前景的想法，兩個適合的原因：一、可緩解用戶端的過度擬合：在聯邦式學習的情境下，用戶若不能夠頻繁的執行模型聚合，用戶間模型的差異化會不斷累積，並且使用戶模型學習過多無用的本地資料特徵，這些相對於獨立同分佈而言屬於多餘的噪音，都會使得最終聚合得到近似於理想模型的困難度增加。知識蒸餾提供良好泛化的能力，應用其於聚合階段有助於全域模型獲取到該學習的資訊，緩解聯邦式學習下容易在用戶端發生的不完整資料偏差（incomplete-data bias）和過訓練（overtraining）的問題。二、有助全域模型學習到類間關係：在資料非獨立同分佈的情況越明顯

doi:10.6342/NTU202202304

時，本地模型學習到的類間關係可能不完整也不一致。此時不恰當的聚合方式可能無法使真實的類間關係有效傳遞給全域模型。如何辨認有效的類間關係，將影響學生模型能否汲取需要的知識。

然而，目前以蒸餾為基礎的聯邦學習架構並無法有效善用上述知識蒸餾方法的優勢。本篇論文提出的 FedUA 旨在讓知識蒸餾更容易融入聯邦學習，在非獨立且同分佈的和限制傳輸問題發生時可進一步提升學習效果。

2.2 Uncertainty in DNNs

一般的深度神經網路並沒有能力表達信心程度，然而，卻對特定應用領域信心程度的顯示越漸重要，例如安全攸關（safety-critical）任務和醫學應用等。因此，關於神經網路不確定性的研究也相繼被提出，包含定義了深度神經網路不確定的來源、量化不確定性、校正網路等不同的範疇。

一般來說，會將神經網路不確定分為以下三者：一、資料不確定性（data uncertainty）：資料天生就存在的不確定性，即便校正良好的模型此種不確定性依舊存在。二、模型不確定性： 模型因知識匱乏所致，一般而言此種不確定性可以改善透過訓練過程或校正模型來降低。三、分佈不確定性（distributional uncertainty）：預測分佈本身的不確定性。另一角度而言，此種不確定性可作異常檢測（out-of-distribution detection）的重要依據[22-25]。圖 3 顯示了類神經網路不確定性的分類。

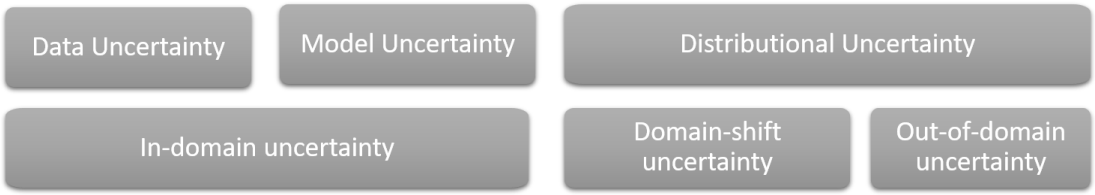

圖 3. 類神經網路不確定性的分類

doi:10.6342/NTU202202304

一般而言，資料不確定性會表現在最終預測上。例如，分類任務的歸一化指數函數輸出（softmax output）或迴歸任務的預測之標準差。然而，研究發現，神經網路常常發生過度自信且歸一化指數函數輸出常常校準不佳[26-28]，導致不精確的不確定估計。

若針對聯邦式學習，當用戶端資料引發分佈不確定（distributional uncertainty）時其負面效果更加明顯且處理起來更加棘手。因此，量化不確定性無疑對於模型的聚合是重要資訊依據。我們進一步考慮，以知識蒸餾為基礎的聯邦學習架構(請參見圖 4)。教師模型提取而來的對數機率作為學生模型訓練過程的主要依據，若能有效表現出目前樣本的信心，在多教師(multi-teacher)架構下，便能做出更彈性且有效率的知識傳授。

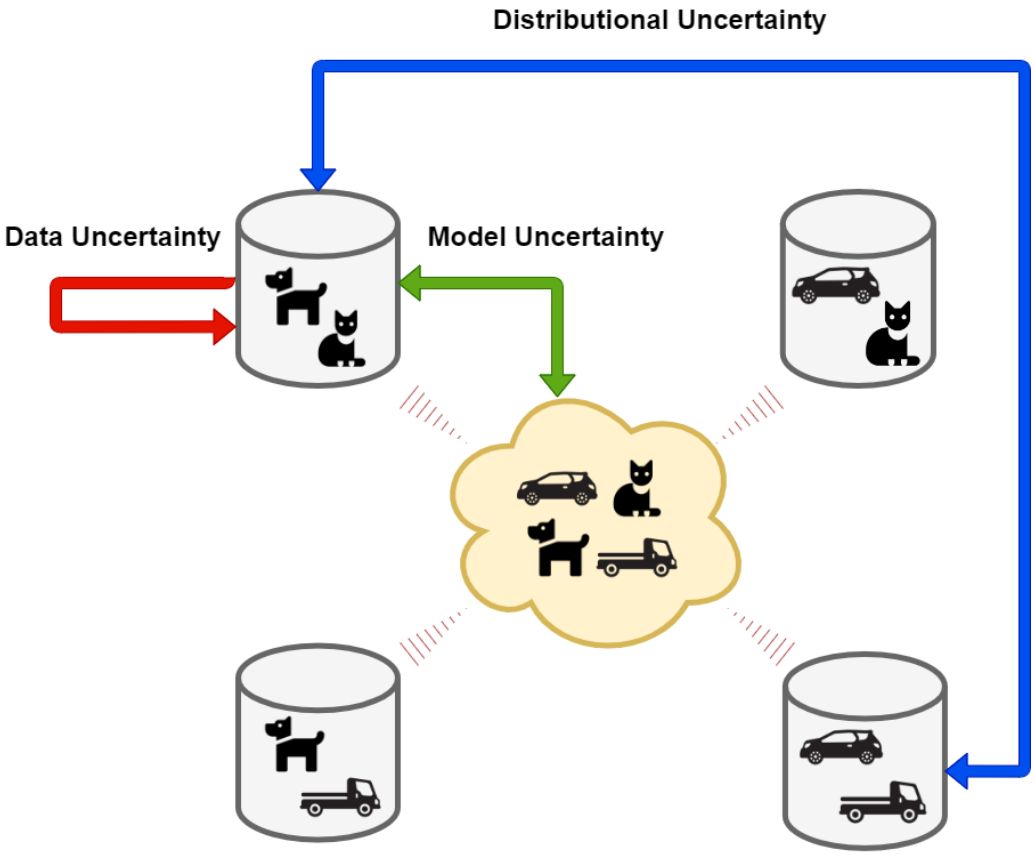

圖 4. 聯邦學習下的網路不確定性

doi:10.6342/NTU202202304

我們提出的 FedUA 考慮不確定性的影響，並加入不確定性量化步驟以有利聚合階段全域模型訓練目標的明確化。

doi:10.6342/NTU202202304

Chapter 3 Proposed Method

我們提出的聚合架構是以目前盛行的知識蒸餾為基礎，並加入**不確定性測量**與**樣本評估**兩個核心方法。以下說明此兩個方法的細節及整體架構。

3.1 Uncertainty Measurement

由於聯邦式學習的設定，用戶端各自以本地資料作為神經網路的訓練資料，導致本地模型即便在相同的訓練參數下，相同的資料輸入於測試階段仍可能會產生高度不一致的損失與預測。伺服器端在資料去中心化而導致無法窺探用戶模型擅長的預測資料情形發生。我們引入深度神經網路的不確定性測量以作為模型聚合過程中的重要依據資訊，以使伺服器端獲知參與用戶各自對於特定的資料輸入所產生的預測之信心程度，進而使後續整合所得的全域模型更加可靠。

考量蒸餾式的聯邦學習架構，我們希望在聚合階段可以借助資訊量測，去趨近理想的教師模型，讓知識的傳承上可以更完善。根據模型不確定性，在模型學習階段，若某特定類別擁有越大量的輸入資料時，則已訓練模型在推論階段，在該類別的輸出應有更高的信心表現。關於分布不確定性，透過有效的測量能讓所有的教師模型更能『揚長避短』地參與學生模型的學習，進一步使伺服器端的學生模型能夠有更全面性的分類能力，。

針對每個用戶端，我們利用高斯判別分析 (Gaussian Discriminant Analysis) 建立其特徵空間密度之高斯混合模型(Gaussian Mixture Model)，給定一組 (X,Y) 建立方式如下:

doi:10.6342/NTU202202304

for each class c with samples $X_c \subset X$ **do**

$$w_c \leftarrow \frac{|X_c|}{|X|}$$

$$\mu_c \leftarrow \frac{1}{|X_c|}\sum_{X_c} f_w(X_c)$$

$$\sigma_c \leftarrow \frac{1}{|X_c|}\ (f_w(X_c)-\mu_c)\ (f_w(X_c)-\mu_c)^T$$

在模型聚合前，利用高斯混合模型，量化目前樣本對特定用戶端模型的認知不確定性（epistemic uncertainty），過程如下:

$$z \leftarrow f_w(x) \qquad f: feature\ \ extractor$$

$$p(z) \leftarrow \sum_c w_c\ N(z;\mu_c;\sigma_c) \qquad N: Gaussian\ \ model$$

先給定用戶端模型，將樣本輸入 x 做特徵提取(feature extraction)，得到特徵向量Z。透過此密度模型，推算目前伺服器端樣本在各用戶端模型的特徵空間密度機率。

值得注意的是，此時，我們採用的不確定性測量方法為單一決定性模型(Single Deterministic Models)，旨在減少在模型在訓練和測試時的計算負擔。

另外，我們以特徵空間(Feature Space)作為量化對象，而非歸一化指數函數(Softmax) 。原因是：在蒸餾式聯邦學習架構下，資料的類間關係（class correlation）是有利於聚合模型的，此關係反映在偶然不確定性（aleatoric uncertainty）上。因此，我們希望聚合過程可排除該因素的影響以避免造成目標不協調（objective mismatch）的情形發生。針對此項考量，在我們的實驗中也針對此項做了比較分析。

3.2 Sample Assessment

對於典型的知識蒸餾，學生模型與教師模型的訓練資料為獨立同分布，讓兩者能達到有效率而穩定的知識傳承。但考慮聯邦式學習下所對應之多教師的情形，用戶上傳的教師模型容易過度擬合於本地資料，我們希望伺服器聚合階

doi:10.6342/NTU202202304

段能有效將全域模型帶往更泛化的方向。

為了達成上述目的，應當對於學生的訓練資料審慎選擇，因此，我們在 FedUA 加入樣本評估階段。在此階段，我們效法積極學習（Active Learning）的精神，以篩選出高認知不確定性（epistemic uncertainty）的樣本作為教師模型的訓練資料。我們採用 Bayesian Active Learning by Disagreement (BALD) [29]，該方法是基於貝式觀點對樣本的不確定性做量化，參見下方數學式(1)。

$$I(y; w|x, D) = H(y|x, D) - E_{p(w|D)}[H(y|X, w, D)] \quad (1)$$

因此，我們的目標為找出最大化模型輸出與模型參數互資訊（mutual information）的樣本。由資訊理論觀點來說，我們尋求的樣本為: 一.在平均輸出方面為低度信心。二.對單一採樣模型輸出則為高度信心。綜合以上，代表互資訊越大的樣本，其模型間越無法取得輸出上的共識。

應用上，我們藉由蒙地卡羅近似法（Monte Carlo Approximation），可以化為數學式(2):

$$I(y; w|x, D) \approx -\sum_c (\frac{1}{T}\sum_t p_c^t)\log\left(\frac{1}{T}\sum_t p_c^t\right) + \frac{1}{T}\sum_{c,t} p_c^t \log\left(p_c^t\right) \quad (2)$$

將其應用於聯邦式學習時，我們可以理解為獲取本地模型間無法達成一致共識的樣本。具備該特性的樣本，在訓練階段對於模型收斂方向會產生較大分歧，因此對於伺服器在全域模型的優化相對重要。

3.3 Overall Architecture

我們所提出的架構係基於 FedDF 之上，在伺服器加入不確定性測量及樣本評估兩個階段。這兩者有助於聚合階段可更好且有效率的訓練出全域模型，並且不會造成過多的額外運算成本，也不會剝奪知識蒸餾架構的完整性。

doi:10.6342/NTU202202304

在每回合的聚合操作中，伺服器會將該回合上傳的所有用戶端的模型視為教師模型，在執行教師模型評估時，我們會擷取前向傳遞時特定神經網路層的輸出，提供給上述兩個階段做進一步分析。不確定性測量藉由用戶端模型的特徵表示輸出權重，對數機率合併階段（Logits Combination），則利用此權重取代 FedDF 取均值的方式，得到組合對數機率（Ensemble Logits）。同時，利用模型預測做樣本評估，通過的樣本將作為教師模型的訓練資料。執行這些預處理後，會先將學生模型參數初始化為教師模型們的平均參數（即 FedAvg），而後我們便可以進行知識蒸餾，在結束後將訓練完的學生模型作為全域模型回傳給用戶端做本地訓練。圖 5 顯示了我們所提出的 FedUA 之整體架構示意圖。

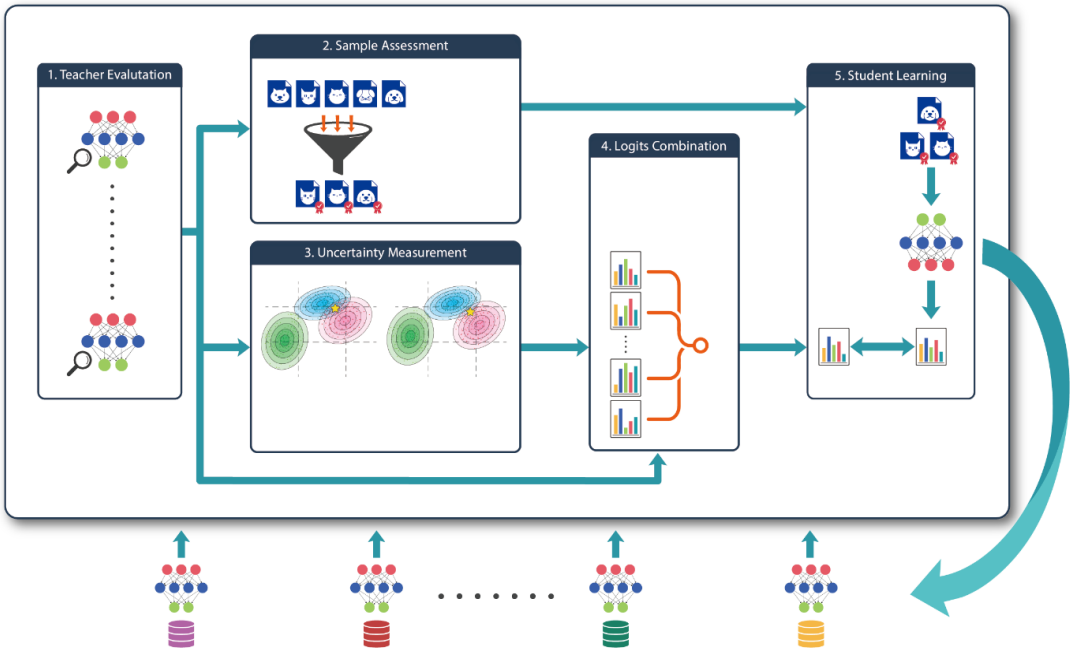

圖 5. 我們提出的 FedUA 架構示意圖

doi:10.6342/NTU202202304

Chapter 4 Experiments

為了驗證核心方法的有效性，我們個別對於**不確定性測量**與**樣本評估**做消融測試。在整體架構部分，FedAvg 和 FedDF 可以做為比較對象。所有實驗重複執行五次以上，並統計平均和變異做為實驗結果。

4.1 Setup

- **Datasets and Network Models**

我們將提出的 FedUA 架構實驗於圖像分類應用中，並選用 ResNet-32 為神經網路模型，CIFAR -10 做為訓練資料集。我們從訓練資料中選擇 40000 張圖作為用戶端本地訓練用的標籤資料。其餘 10000 張圖則作為伺服器端蒸餾式聚合所需的未標籤資料。

對於用戶端訓練用的標籤資料，我們採用 Step 法[20]當基線實現非獨立同分布，搭配 Dirichlet 做出不同類型非獨立同分布的型態。在 Step 法下，每個用戶端擁有大量特定 2 個類別的圖，其餘 8 個類別的圖則是少量。Dirichlet 法利用集中參數， α (concentration parameter)，調控 Dirichlet 分布以達到不同分散程度的資料。

doi:10.6342/NTU202202304

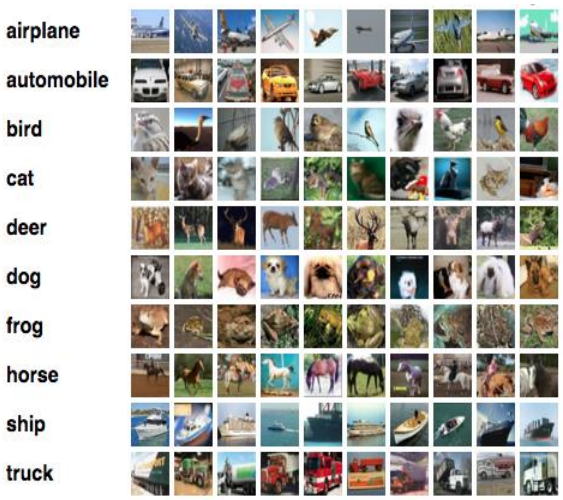

圖 6: CIFAR10 的圖像類別

CIFAR-10 資料及之圖像(請參見圖 6)係由交通工具和動物組成的 10 種類別，存在類間關係有利我們探究知識蒸餾，不確定性測量，聯邦學習架構三者相互的關係，有助確認聯邦學習的聚合階段，兼顧類間關係的學習和判定教師模型的有效性。

- **Details**

在聯邦學習的設定部分，為了方便比較並考慮其他論文的參數設定，我們將 40 回和數 E (rounds) 設為上限。客戶端數量假設共 10 個，回報比率 C 初始化為 1.0 (Reporting Fraction)，該參數決定每一輪隨機選擇客戶的數量，做為模型已上傳比例以便進行後續聚合。

每輪本地端或伺服器端的訓練皆為 20 期(epoch)並採用隨機梯度下降法(stochastic gradient descent)進行訓練。本地端的批次大小(batch size)設為 32，伺服器端則為 128。此外，不同於本地端使用交叉熵(cross entropy)作為損失函數，為了適應知識蒸餾，在伺服器端使用的是相對熵 (KL divergence)。

4.2 Results and Analysis

doi:10.6342/NTU202202304

4.2.1 Ablation Analysis

- Sample Assessment

我們在樣本評估的消融分析上，除了 Bayesian Active Learning by Disagreement (BALD)的驗證外，進一步實驗未標籤資料在不同的取樣比率下 (Sample Ratio)，對於伺服器端的影響。

除了在相同的未標籤資料的取樣比率下，BALD 有更佳的表現外，證實利用 BALD 篩選出樣本組成的批次，相較於傳統隨機組成的批次，對於全域模型的模型優化是更有學習意義的。我們也發現，無論何種篩選方式，最好的表現皆非落在擁有完整資料集的迭代訓練上。相關結果呈現於表 1 中。

	SR = 0.2	SR = 0.4	SR = 0.6	SR = 0.8	SR = 1.0
Random	71.5 ± 0.61	71.3 ± 0.78	71.8 ± 0.66	**72.3 ± 0.66**	72.1 ± 0.55
BALD	72.0 ± 0.24	72.5 ± 0.36	**73.9 ± 0.46**	73.7 ± 0.33	73.2 ± 0.48

表 1. 比較不同的樣本評估設定之表現

藉由加入我們提出的樣本評估，不僅是對蒸餾式聯邦學習的表現上有利，同時在減輕伺服器端的運算負擔上是有幫助的，儘管後者並非聯邦式學習的首要考量。

- Uncertainty Measurement

在此章節，我們著重在不確定性測量的表現上，我們觀察用戶端在原始的學習設置並以非獨立同分布的資料下學習，用戶端各自上傳後，其在伺服器端作為教師模型的推論模式下，特徵空間密度輸出和歸一化指數函數輸出的表現。

doi:10.6342/NTU202202304

如果以神經網路的不確定性進行探討，CIFAR-10 存在高度的 (1) 資料不確定性(Data Uncertainty)，因為動物間存在一定的相關性，交通工具間也是如此。由於非獨立且同步的假設下，(2) 模型不確定性(Model Uncertainty) 和 (3) 分布不確定性(Distributional Uncertainty)亦是我們該關注的部分。前者是因為本地模型受類別層級的資料不平衡(Data Imbalance)影響，導致訓練樣本缺少的類別，其不確定性應越高。而後者，是針對上傳的本地模型相較於理想模型(Ideal Model)而言，訓練樣本集和真實樣本集落差而引發的分布不確定性。

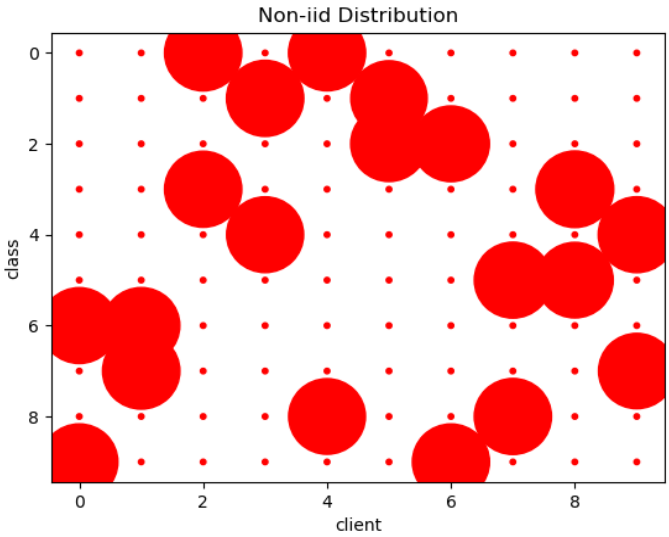

圖 7. 用戶端非獨立同分佈的訓練資料分布圖

以圖 7 的非獨立同分布做說明，在此例中，編號 0、2、3、5、6、7 號客戶端擁有的大量圖像類別，包含一種交通工具和一種動物，因此這些客戶端在粗粒度分類(coarse-grained)的能力上較為出色；編號 1、8、9 號客戶端則有兩種大量的動物類別圖像及編號 4 號則是唯一擁有兩種交通工具的類別圖像，這些客戶則傾向具備在相關性高的類別中做出細粒度(fine-grained)分類的能力。

舉例，當輸入樣本屬於動物類別時，此時 4 號教師模型在動物類別無論在粗粒度和細粒度的分類能力上都較差，我們應該抑制學生模型參考 4 號教師模型的程度。此外，促成 1、8、9 教師模型能在動物類別細粒度能力的發揮，是學生模

doi:10.6342/NTU202202304

型訓練在後期持續提升的關鍵。反之，若輸入樣本屬於交通工具時，我們應調降 1、8、9 號並調高 4 號教師模型在聚合階段的影響力。

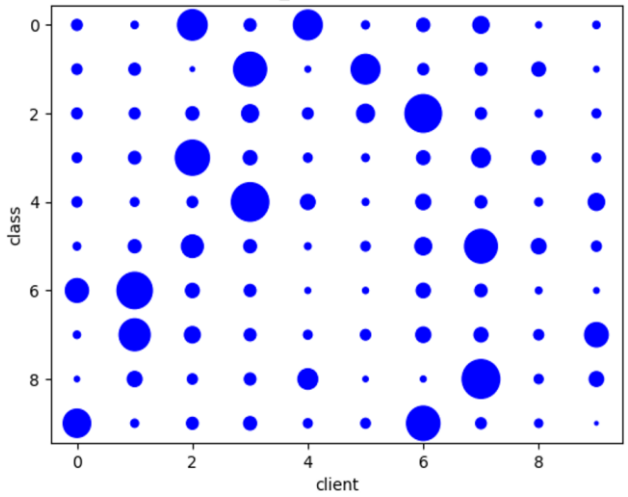

圖 8. Top1 歸一化函數輸出之熵(Softmax Entropy)的統計分佈

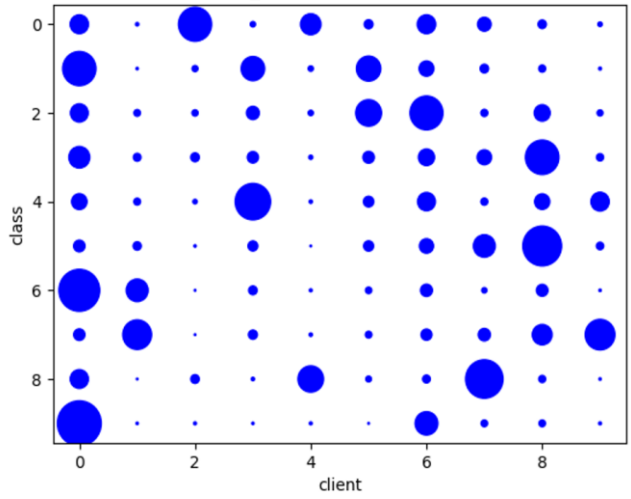

圖 9. Top1 特徵空間密度 (Feature Space Density) 的統計分佈

doi:10.6342/NTU202202304

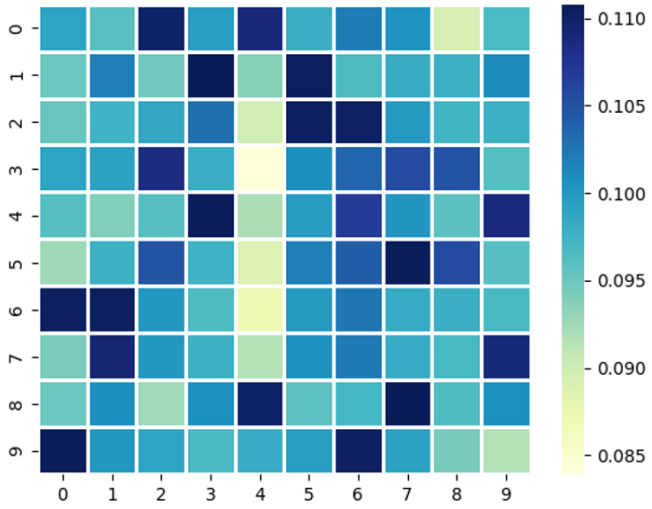

圖 10. 歸一化函數輸出之熵(Softmax Entropy)的平均值

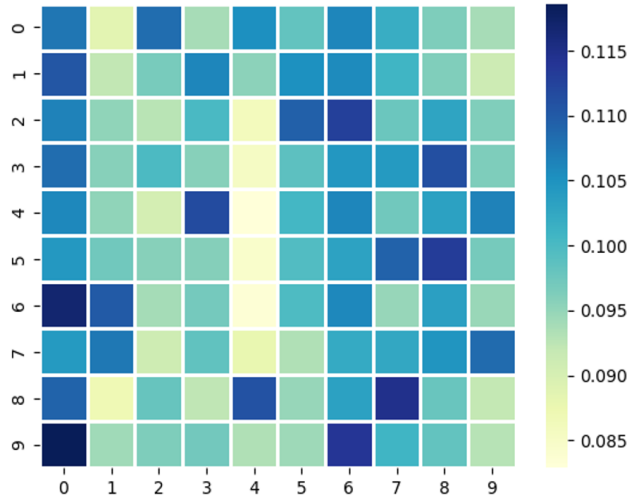

圖 11. 特徵空間密度 (Feature Space Density) 的平均值

我們從圖 8 至圖 11 的觀察結果可以得知，無論是歸一化函數輸出之熵或特徵空間密度方法，其在資料不確定性及模型不確定性的表現上都不錯，但倘若考量聯邦式學習所需的分布式不確定性，後者得到的結果呈現了更好的資訊，因此我們提出的 FedUA 最終採用的是特徵空間密度法。

4.2.2 Overall Performance

doi:10.6342/NTU202202304

- Non-IID

我們將 FedUA 依據前述設置，並以 Step 法做為獨立同分佈資料集實現，並將結果與 FedDF、FedAvg 做比較。

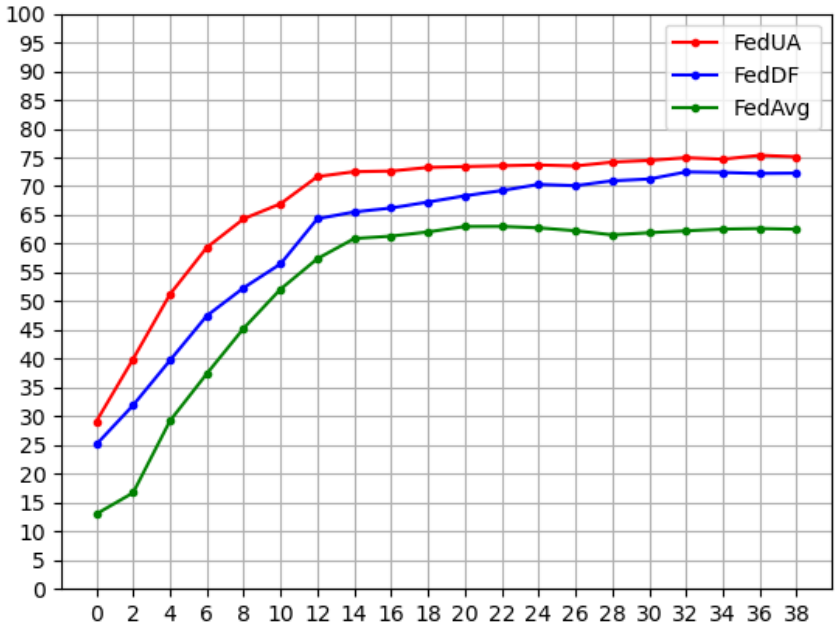

圖 12 .FedUA、FedDF、FedAvg 的學習曲線(Learning Curve)

在初期，受益於知識蒸餾，FedUA、FedDF 的全域模型的準確率明顯優於 FedAvg。三者收斂的速度接近，約在 14 輪左右便開始趨緩。最終，FedUA 的準確率相較於 FedDF 約有 2-3%的進步。(請參見圖 12)

- Different Non-IID Data Partition

我們比較在不同的非獨立且同步的資料型態下，FedUA、FedDF、FedAvg 的表現是否不同。圖 13，由左而右分別是 Step 法、Dirichlet (α=0.1)、Dirichlet (α=0.5)的結果。

doi:10.6342/NTU202202304

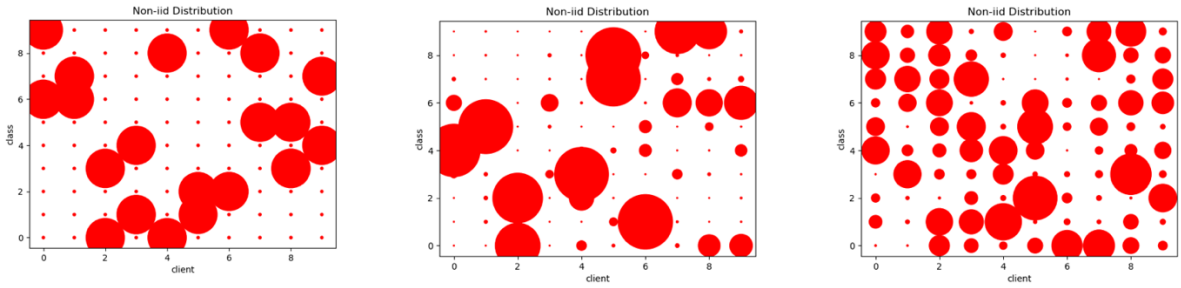

圖 13. 不同的非獨立同分布設定下的實際分布，由左至右依序為 Step、Dirichlet(0.1)、Dirchlet(0.5)

	Step	Dirichlet(0.1)	Dirichlet(0.5)
FedAvg	62.6 ± 0.23	59.6 ± 1.03	80.1 ± 0.45
FedDF	72.3 ± 0.49	64.4 ± 0.93	82.8 ± 0.47
FedUA	74.8 ± 0.45	65.3 ± 0.78	83.4 ± 0.24

表 2. 比較 FedAvg、FedDF、FedUA 在不同的非獨立同分布設定下之表現

由圖 13 及表 2 可以觀察到，在 Dirichlet(α=0.1)設定下，客戶端間存在更嚴重的資料不平衡和分布不確定性，FedUA 核心方法帶來的優勢不如 Step 法突出，約莫相較 FedDF 只有 1%的成長，但有賴於知識蒸餾比起 FedAvg 仍有明顯的進步。若在 Dirichlet(α=0.5)下，客戶端的資料更趨近獨立且同分佈，此時無論是 FedUA 或 FedDF 都表現一般。

推測原因在以 Dirichlet 做資料分割下，衍生出更嚴重的資料不平衡和更複雜的特徵空間密度型式，導致模型不確定性和分布式不確定性情況更難以被估量。然而，若考慮聯邦學習在目前真實的使用情境，裝置間的本地資料應當傾向 Step 類。

- Limited Communication Cost

最後，我們考慮聯邦式學習實務上面臨到的通訊成本受限情境。普遍會藉由調整每輪上傳至伺服器的客戶端參與比率 (Participate Ratio : C)來進行模擬。

doi:10.6342/NTU202202304

我們設置在不同的參與比率下，進而觀察對應之結果(請參閱表 4.3)。

	C=1.0	C=0.7	C=0.4
FedAvg	62.6 ± 0.23	61.3 ± 0.35	58.1 ± 0.26
FedDF	72.3 ± 0.49	68.1 ± 0.61	63.8 ± 1.03
FedUA	74.8 ± 0.45	71.8 ± 0.56	68.3 ± 0.85

表 3. 比較 FedAvg、FedDF、FedUA 在不同參與比率的結果

從結果而言，若參與比率調降為 0.7 設置以下，FedAvg 並未明顯衰退，但 FedDF 約莫下降 4%。甚至在參與比率 0.4 時，已快衰退將近 10%。

原生的蒸餾式聯邦學習在參與比率下降時，教師模型不足時，產生知識匱乏的現象，可能會使伺服器聚合階段使用的未標籤資料沒有正確的學習目標，導致 FedDF 十分依賴客戶端穩定地參與聚合。

但 FedUA 因為引入樣本評估和不確定性測量，某種程度上，緩衝上述問題造成的影響，規避了無意義甚至是錯誤的學習模式情境發生時對學生模型造成的損害。

doi:10.6342/NTU202202304

Chapter 5 Conclusion & Future Work

近年來聯邦式學習蓬勃發展，但非獨立同分布一直是其必須面對重要的挑戰。聯邦式學習不僅受限於安全性和通訊成本的考量上，多元複雜的非獨立同分布型態難以得到通用的解決方式。

實務上，過去採取知識蒸餾式學習的聯邦式學習架構，通常由於本地訓練資料的不平衡，容易造成的本地模型學習到的類間關係不完整，進而使聚合階段的全域模型學習不夠完備。

因此，在面對非獨立同分布的問題上，我們考量真實世界的用戶資料的情形，希望我們提出的不確定性測量及樣本評估確實能夠帶來幫助，提供聚合階段更多的資訊，並做出更有效的學習。

但我們目前的討論僅適用於圖像分類的問題上，也未能嘗試實驗於更複雜的圖像資料集。我們認為在不確定性方法和蒸餾式聯邦學習上，或許可以整合兩者的優勢進一步做出不同的協作模式，這將是我們未來努力的主要方向。

doi:10.6342/NTU202202304

Reference

[1] McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." *Artificial intelligence and statistics*. PMLR, 2017.

[2] Kairouz, Peter, et al. "Advances and open problems in federated learning." *Foundations and Trends® in Machine Learning* 14.1–2 (2021): 1-210.

[3] Zhao, Yue, et al. "Federated learning with non-iid data." *arXiv preprint arXiv:1806.00582* (2018).

[4] Xiao, Peng, et al. "Averaging is probably not the optimum way of aggregating parameters in federated learning." *Entropy* 22.3 (2020): 314.

[5] Zhu, Hangyu, et al. "Federated learning on non-IID data: A survey." *Neurocomputing* 465 (2021): 371-390.

[6] Yoshida, Naoya, et al. "Hybrid-FL for wireless networks: Cooperative learning mechanism using non-IID data." *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE, 2020.

[7] Duan, Moming, et al. "Astraea: Self-balancing federated learning for improving classification accuracy of mobile deep learning applications." *2019 IEEE 37th international conference on computer design (ICCD)*. IEEE, 2019.

[8] Ghosh, Avishek, et al. "Robust federated learning in a heterogeneous environment." *arXiv preprint arXiv:1906.06629* (2019).

[9] Ghosh, Avishek, et al. "An efficient framework for clustered federated learning." *Advances in Neural Information Processing Systems* 33 (2020): 19586-19597.

[10] Li, Tian, et al. "Federated optimization in heterogeneous networks." *Proceedings of Machine Learning and Systems* 2 (2020): 429-450.

[11] Hsu, Tzu-Ming Harry, Hang Qi, and Matthew Brown. "Measuring the effects of non-identical data distribution for federated visual classification." *arXiv preprint arXiv:1909.06335* (2019).

doi:10.6342/NTU202202304

[12] Wang, Kangkang, et al. "Federated evaluation of on-device personalization." *arXiv preprint arXiv:1910.10252* (2019).

[13] Arivazhagan, Manoj Ghuhan, et al. "Federated learning with personalization layers." *arXiv preprint arXiv:1912.00818* (2019).

[14] Smith, Virginia, et al. "Federated multi-task learning." *Advances in neural information processing systems* 30 (2017).

[15] Liu, Boyi, Lujia Wang, and Ming Liu. "Lifelong federated reinforcement learning: a learning architecture for navigation in cloud robotic systems." *IEEE Robotics and Automation Letters* 4.4 (2019): 4555-4562.

[16] Zhou, Yanlin, et al. "Distilled one-shot federated learning." *arXiv preprint arXiv:2009.07999* (2020).

[17] Jeong, Eunjeong, et al. "Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data." *arXiv preprint arXiv:1811.11479* (2018).

[18] Li, Daliang, and Junpu Wang. "Fedmd: Heterogenous federated learning via model distillation." *arXiv preprint arXiv:1910.03581* (2019).

[19] Lin, Tao, et al. "Ensemble distillation for robust model fusion in federated learning." *Advances in Neural Information Processing Systems* 33 (2020): 2351-2363.

[20] Chen, Hong-You, and Wei-Lun Chao. "Fedbe: Making bayesian model ensemble applicable to federated learning." *arXiv preprint arXiv:2009.01974* (2020).

[21] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." *arXiv preprint arXiv:1503.02531* 2.7 (2015).

[22] Gawlikowski, Jakob, et al. "A survey of uncertainty in deep neural networks." *arXiv preprint arXiv:2107.03342* (2021).

[23] Hendrycks, Dan, and Kevin Gimpel. "A baseline for detecting misclassified and out-of-distribution examples in neural networks." *arXiv preprint arXiv:1610.02136* (2016).

[24] Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation:

doi:10.6342/NTU202202304

Representing model uncertainty in deep learning." *international conference on machine learning*. PMLR, 2016.

[25] Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles." *Advances in neural information processing systems* 30 (2017).

[26] Ovadia, Yaniv, et al. "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift." *Advances in neural information processing systems* 32 (2019).

[27] Guo, Chuan, et al. "On calibration of modern neural networks." *International conference on machine learning*. PMLR, 2017.

[28] Mukhoti, Jishnu, et al. "Deep Deterministic Uncertainty: A Simple Baseline." *arXiv e-prints* (2021): arXiv-2102.

[29] Gal, Yarin, Riashat Islam, and Zoubin Ghahramani. "Deep bayesian active learning with image data." *International Conference on Machine Learning*. PMLR, 2017.

doi:10.6342/NTU202202304