DISTILLATION LEARNING APPLICATIONS

RISQ/MRM

C'EST VOUS L'AVENIR

SOCIETE GENERALE

OUTLINES

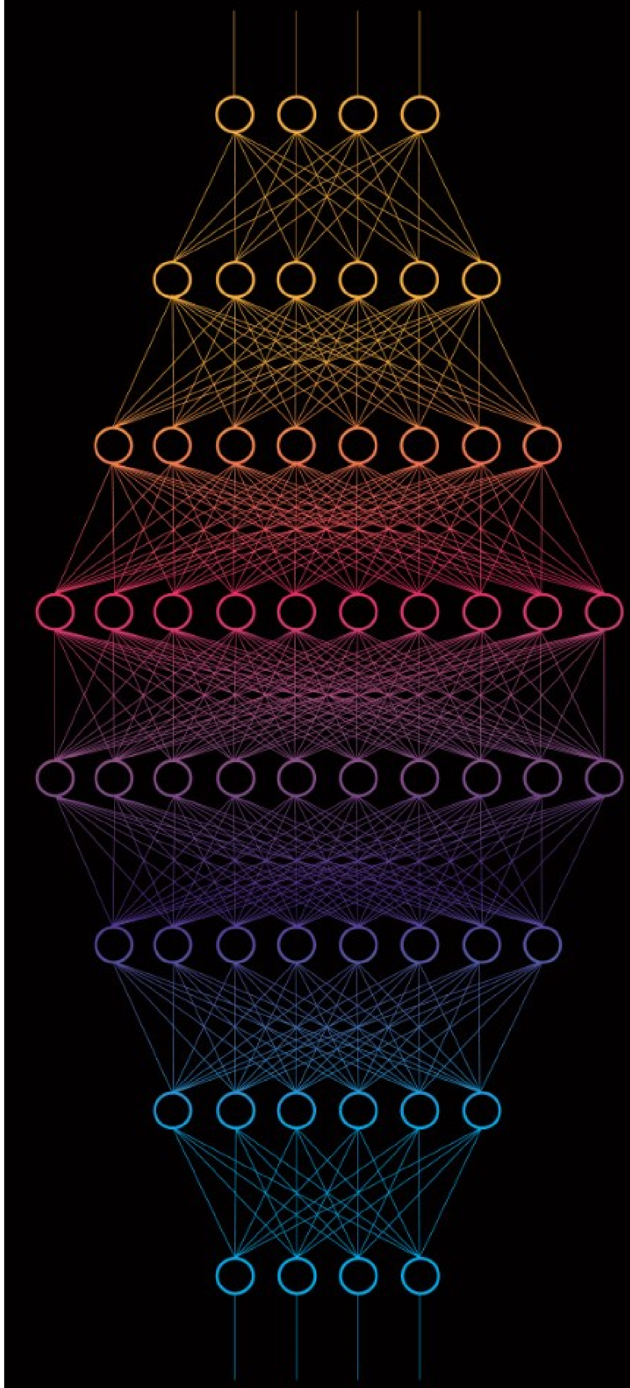
1. SUMMARY OVERVIEW

A. What is Knowledge Distillation ?

B. Context and Motivation

C. Main Use of Knowledge Distillation

SOCIETE GENERALE



A. WHAT IS KNOWLEDGE DISTILLATION ?

Definition of Caruana and al.,2006 : it is a model compression technique that uses a fast and compact model to approximate the function learned by a slower, larger, but better performing model. The name of knowledge distillation is introduced in 2015 by Hinton and al, 2015. However, distillation learning has also been largely used to ameliorate models' performance without compression and in transfer learning tasks.

Teacher model: A complex model that we would like to distill in a simplified model.

Student model: A compact model, often a simplified version of the teacher model, in which we would like to distill the knowledge of the teacher model.

Knowledge: weights, feature maps, activation function, relationships and distributions, etc. learned by the teacher model and that we would like to distill in the student model.

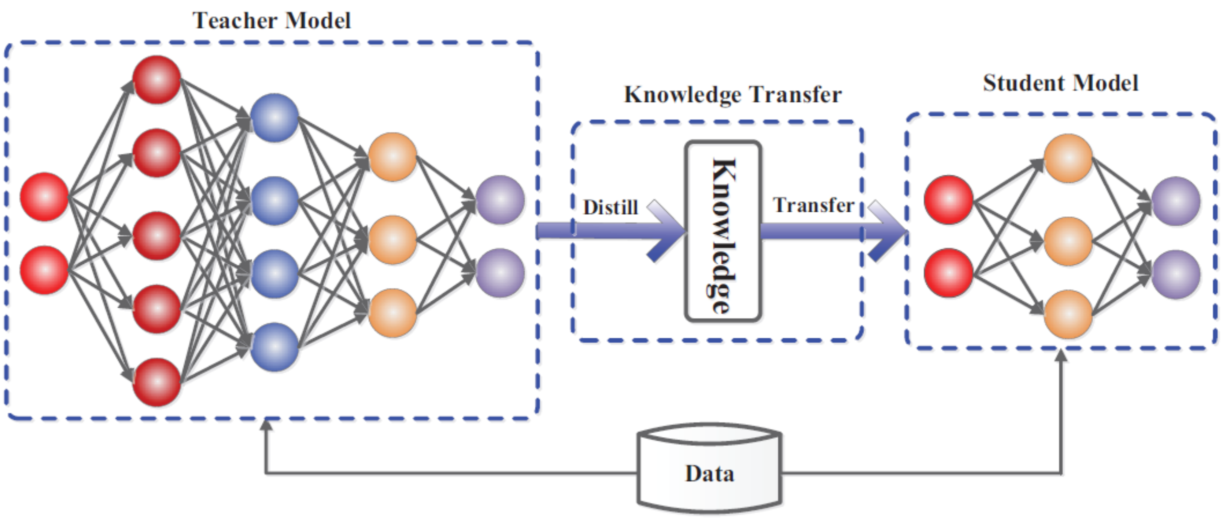


Fig. 1 The generic teacher-student framework for knowledge distillation, source arxiv

SOCIETE GENERALE

B. CONTEXT AND MOTIVATION

Context	Motivation
• Increasing use of **complex models** such as deep learning models and ensemble models. • Model **Transparency Requirement** • Costs of cumbersome models during inference: Latency, pollution, etc.,	• Necessity to deploy at scale machine learning models in production with minimum latency. • The willingness of reducing/simplifying /explaining or enhancing models.

Table 1. Context and Motivation Behind the Use of Knowledge Distillation

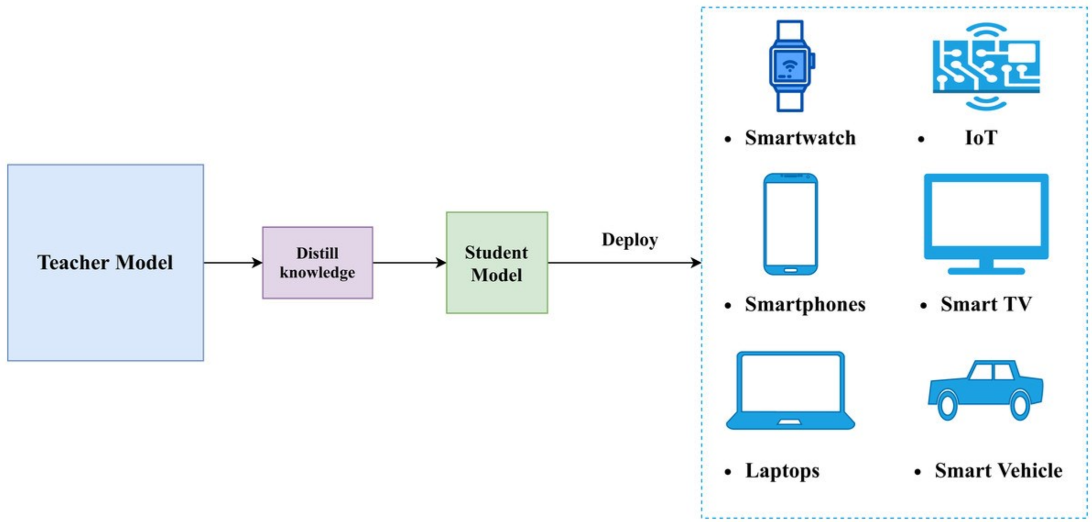


Fig 2. Knowledge Distillation for Device Model Embedding

C. MAIN USE OF KNOWLEDGE DISTILLATION

3 Main Potential Uses of Knowledge Distillation in Model Risk Management

1 — XAI

- If we have an inexplainable teacher such as a deep neural network or a random forest, we can use distillation of the teacher to train an explainable and transparent model such as a decision tree along with being close to the teacher performance.
- In this case, the trade off performance/interpretability must be balanced depending on the situation.
- Usually, we use the teacher for inference alongside with student's explainability insights.

2 — Enhancing Simple Models (ESM)

- A simple model is such as logistic regression, random forest, decision tree, linear regression or a simple neural network.
- For instance, training a logistic regression directly will perform less in the test phase than training a deep neural network and distilling it into the same logistic regression.
- Training a simple model through distillation of a more complex model usually outperforms training directly the same simple model.
- In MRM context , performance of PD estimation models can be enhanced by training a complex model such as deep neural network and then distilling it into a student model.

3 — Enhancing Teacher Models (ETM)

- In the context of compression and due to capacity gap, the student cannot outperform the teacher in general.
- Experiments have shown that we can enhance neural networks performance by distillation into ensemble trees or in some self-distillation specific frameworks.
- Distillation of neural networks in gradient boosted trees has enhanced the performance according to Che and al., 2015.

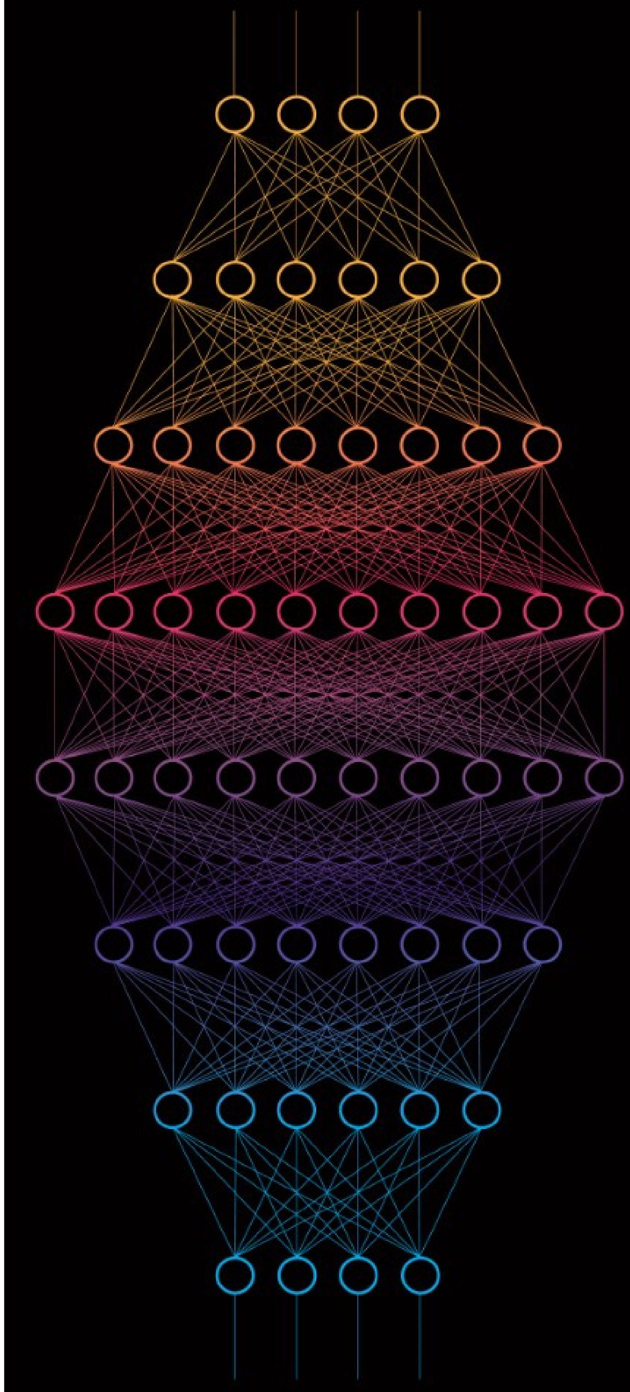
SOCIETE GENERALE

2. KNOWLEDGE DISTILLATION TECHNIQUES

A. Knowledge

B. Training Modes

C. Some Distillation Frameworks



A. KNOWLEDGE

Knowledge is the information that we would like to distill

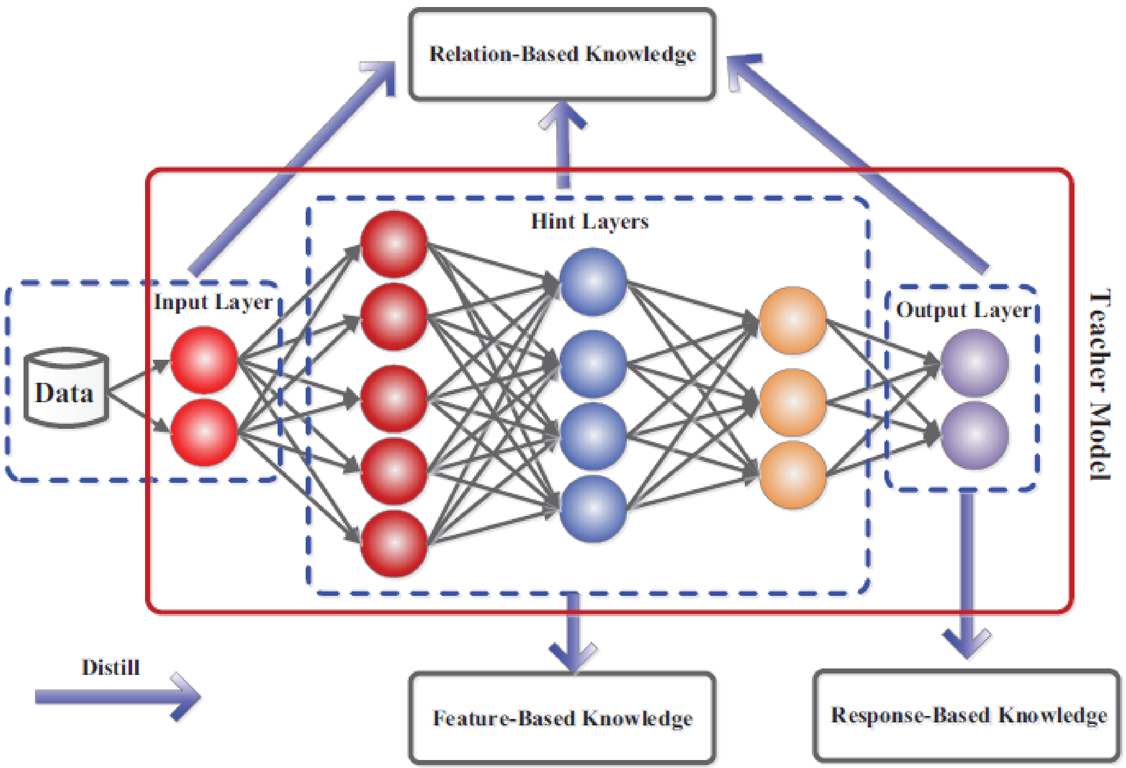


Fig 3. The schematic illustrations of sources of response-based knowledge, feature-based knowledge and relation-based knowledge in a deep teacher network.

SOCIETE GENERALE

1. RESPONSE-BASED KNOWLEDGE

The classical framework for knowledge distillation. The student tries to mimic as good as possible the output predictions of the teacher model in a response-based manner. Practically, we use logits (Neurons outputs before SoftMax) because they contain dark knowledge which is the deep knowledge learnt by the teacher.

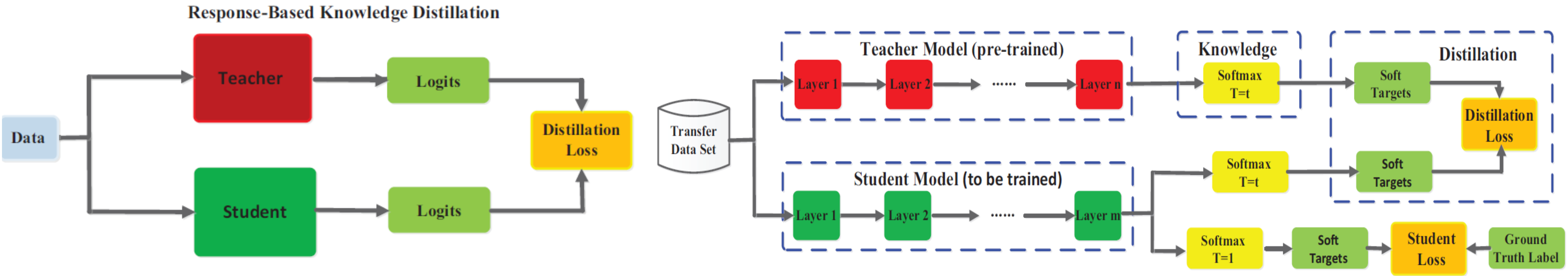


Fig 4. The specific architecture of the benchmark knowledge distillation. The student model can learn to mimic teacher's predictions and also ground truth labels.

Pros	Knowledge	Limits
Easy-to-use, straight-forward Fast, efficient	Predictions of the teacher model Dark knowledge embedded in soft targets or in logits (Hinton and al, 2015, Caruana and al, 2014).	Limited to supervised learning Relies on the final output fails to address intermediate-level supervision

Table 2. Response-based distillation investigation

$$\mathcal{L}_{KD} = \sum_{(x_t, y_t) \in (X_t, Y_t)} [\alpha \mathcal{L}_{CE}(f_S, x_t, y_t) + \beta \mathcal{L}_{KL}(f_S, f_T, x_t)]$$

Formula 1. Hinton Loss for Response-Based KD, Source, Hinton and al, 2015

$$p(z_i, T) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

Formula 2. Hinton Soft-Targets for Response-Based KD, Source, Hinton and al, 2015; very high T values correspond approximately to matching logits.

2. FEATURE-BASED KNOWLEDGE DISTILLATION

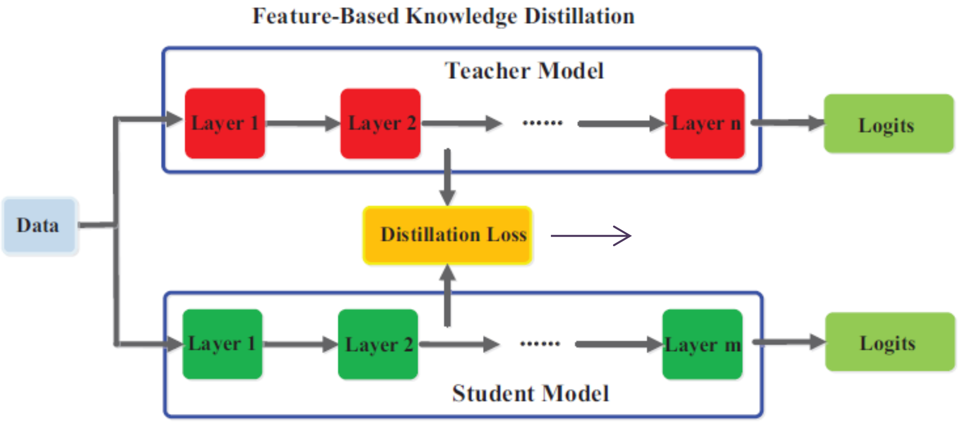


Fig 5. The generic feature-based know

$$L_{FeaD}(f_t(x), f_s(x)) = \mathcal{L}_F(\Phi_t(f_t(x)), \Phi_s(f_s(x)))$$

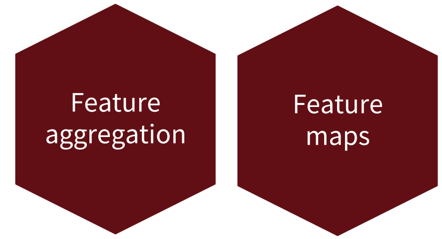


Fig 6. Some types of feature-based knowledge

Pros	Knowledge	Limits
Learn multiple levels of feature representation.	1) Feature representation, hint layers (Romero et al., 2015) 2)Parameter distribution, multi-layer group (Liu et al., 2019c) 3)Feature Maps, hint layers (Chen et al., 2021)	Effectively choose the hint layers from the teacher model and the guided layers from the student model with optimum training complexity is questionable.

Table 3. Feature-based distillation investigation

3. RELATION-BASED KNOWLEDGE DISTILLATION

Based on a relational construction such as correlation, Probability distribution , etc.

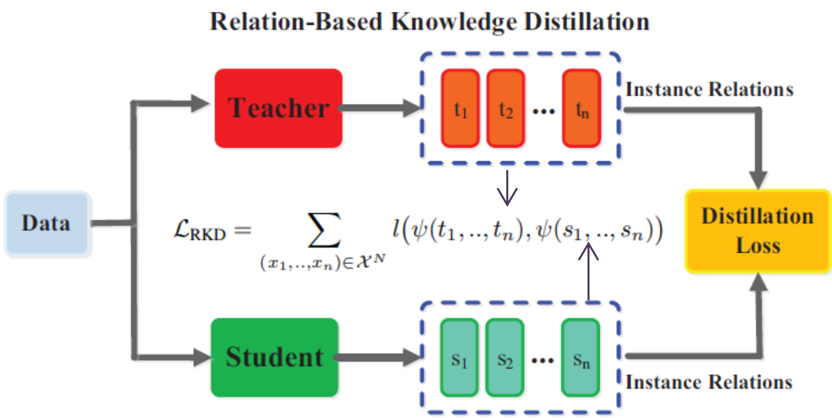


Fig 7. The generic relation-based knowledge distillation.

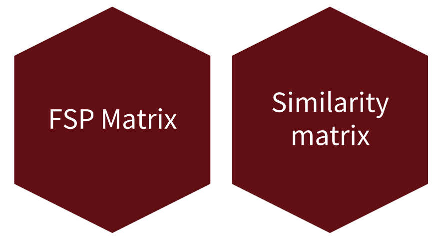


Fig 8. Some types of relation-based knowledge

Technic	Knowledge	Limits
Explores the relationship between layers or data samples.	1) FSP matrix, End of multi-layer group (Yim et al., 2017) 2) Logits graph, hint layers (Zhang and Peng, 2018) 3) Similarity Matrix, hint layers (Tung and Mori, 2019)	Relation modeling difficulties

Table 4. Relation-based distillation investigation

A. TRAINING MODES (1/2)

How can we perform distillation between the teacher and the student ?

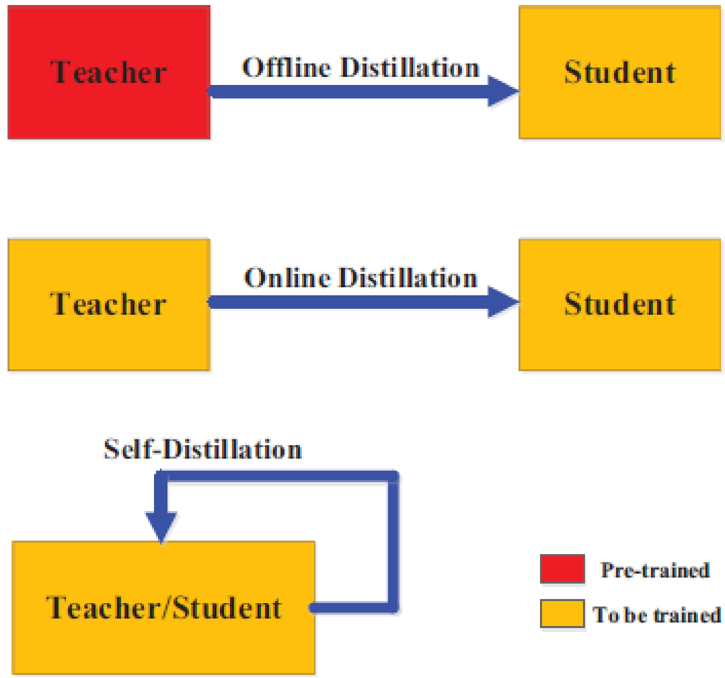


Fig 9. Different Distillation Training Modes . The red color for "pre-trained" means networks are learned before distillation and the yellow color for "to be trained" means networks are learned during distillation.

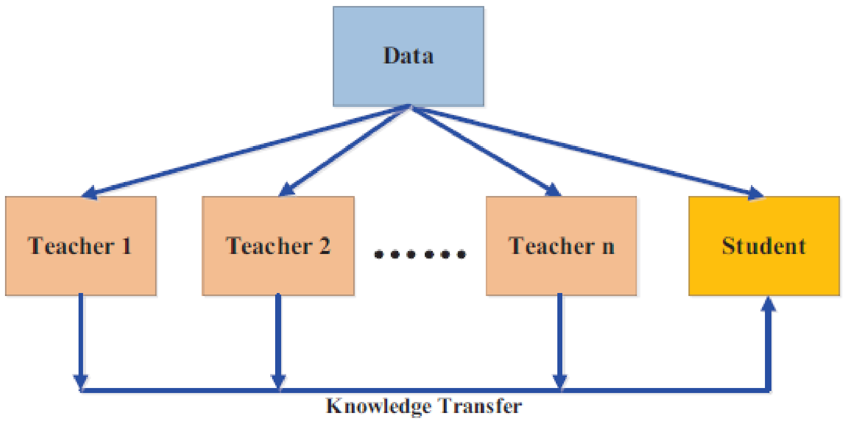
A. TRAINING MODES (2/2)

Training scheme	Usages	Advantages	Limits and Cons
Offline distillation	A good teacher model already exists.	Simple, easy to implement Large usages Most of previous work in KD is offline	1) One-way transfer 2) Two-phase training 3) Student dependency 4) Unavailability of a pre-trained teacher
Online distillation	A good teacher model is not available. Its training is part of distillation.	One-phase, end-to-end training scheme	1) Capacity gap 2) High Training Complexity
Self-distillation	Teacher and student are the same with similar architectures. Used in specific frameworks to enhance baseline models performance.		Capacity gap

Table 5. Distillation Learning Training Modes Assessment

C. SOME DISTILLATION FRAMEWORKS

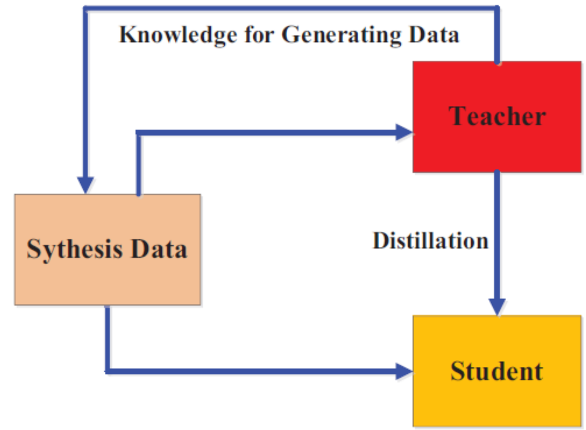
1. Multi-teacher Distillation



Problem	Usage example	Pros
1) Bias coming from one the teacher 2) Lack of knowledge using one teacher	2 teachers, one transfers response-based knowledge and the other transfers feature-based knowledge (Chen et al. 2019b).	Provide richer knowledge to the student Straightforward

Table 6. Multi-teacher Distillation Framework's detailed explanation.

2. Data-Free Distillation



Problem	Usage example	Pros
1) Unavailable data arising from privacy, legality, security and confidentiality	Data is generated from the feature representations from the pre-trained teacher model and used to train the student model in addition to traditional distillation.	Powerful, Uses GAN Straightforward

Table 7. Data-Free Distillation Framework's detailed explanation.

3. ADVERSARIAL KNOWLEDGE DISTILLATION

An effective framework to enhance the power of student learning via the teacher knowledge distillation using GAN. This framework tackles two main problems;1) Difficulty for the teacher to learn the true data distribution (lack of data, unrepresentative data, small model, etc.); 2) Small capacity of the student and difficulties to mimic accurately the teacher (Capacity gap, Unreliable teachers)

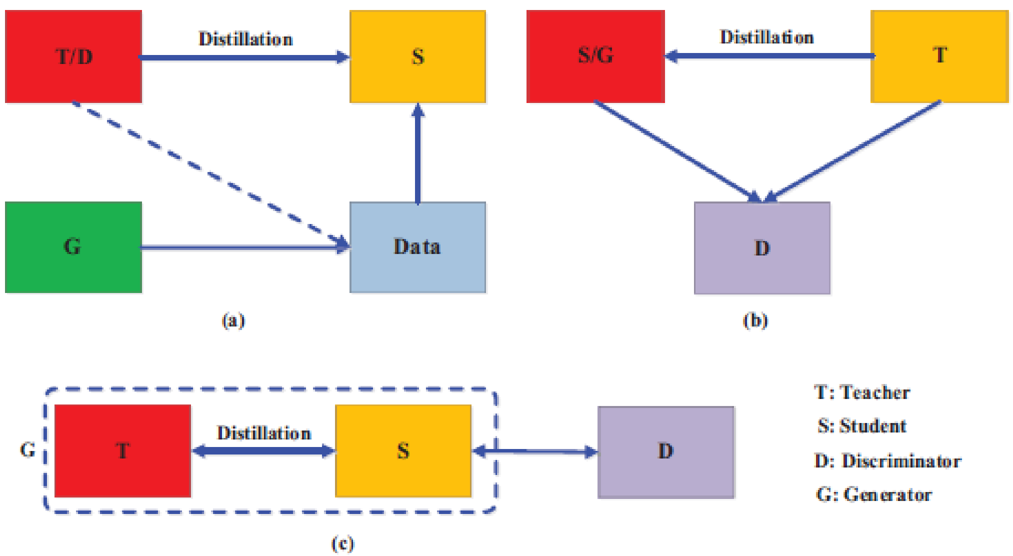


Fig 10. The different categories of the main adversarial distillation methods.

Scheme	Explanation
(a)	A generator is trained on true data distribution. Generated Data go then through teacher discrimination based on its proper data distribution. Student learns then teacher knowledge from 2 sources; 1) classical distillation process, 2) through generated data embedding teacher's internal feature representation.
(b)	A discriminator is trained on teacher's feature distribution. In addition to traditional distillation process, the student will generate new data based on its internal feature distribution corrected each time by the discriminator. The generated data is not used for training.
(c)	A discriminator is trained on true data distribution and corrects feature distribution of generators which are the student and teacher in an online setting.

Table 8. Adversarial Knowledge Distillation Framework's detailed explanation.

4. EXPLAINABILITY DISTILLATION (1/2)

Teacher explanation are important features driving a specific prediction. However, traditional distillation doesn't distill explanation and thus, student predictions are not driven by the same features due to explanation inconsistency between the teacher and the student.

Alharbi and al., 2021 have proposed a novel framework to distill explanation in addition to dark knowledge called XDistillation (XD). The framework has outperformed all traditional distillation methods.

BEFORE

AFTER

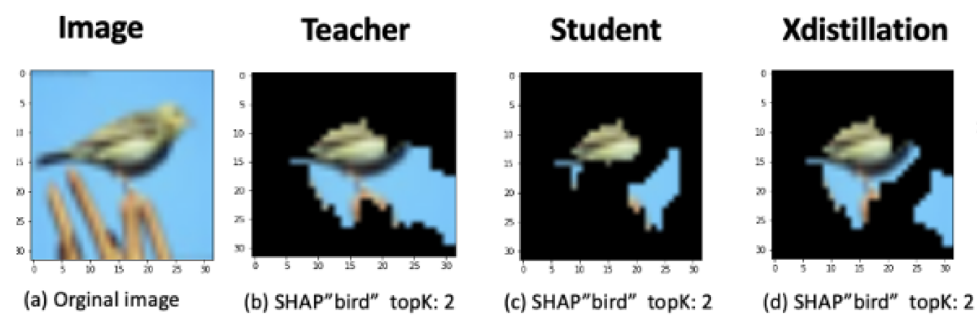


Fig 11. Inconsistency between teacher and student explanation

Fig 12. The overlapping explanation area of teacher, KD and XD.

4. EXPLAINABILITY DISTILLATION (2/2)

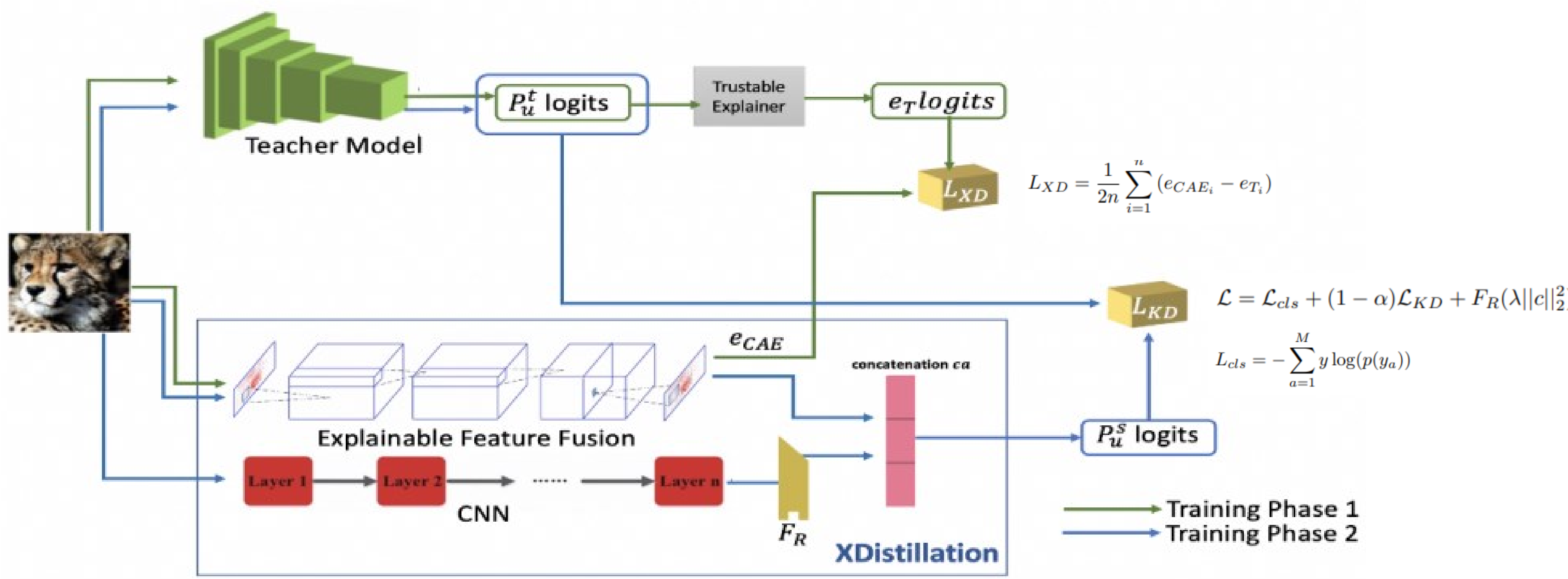


Fig 13. The overall architecture of Xdistillation; The novel idea is the feature fusion component which approximate teacher explanation.

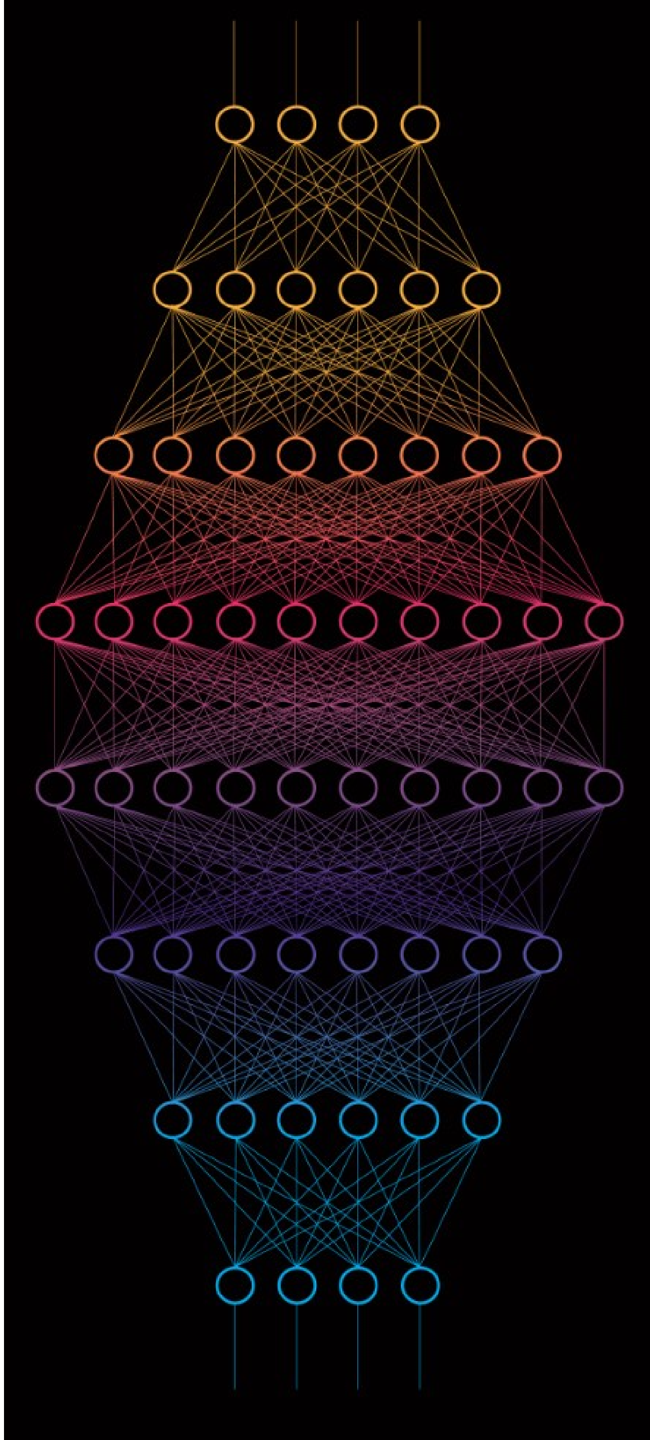
Model	Accuracy %	#parameters
Teacher	**93,78**	14,728, 266
Baseline student	89.2	2,781,386
Knowledge distillation (KD)	90.2	2,781,386
Attention transfer (AT)	90	2,781,386
Neural selective transfer (NST)	89.47	2,781,386
Activation boundary (AB)	89.36	2,781,386
Xdistillation	**90.9**	3,521,276

Table 9. Performance comparison

3. TESTING EXAMPLES

A. Distillation for Neural Network Explanation

B. Logistic Regression's Performance Enhancing



DISTILLATION FOR NEURAL NETWORK EXPLANATION (1/2)

Frosst & Hinton, 2017

- *MNIST is a computer vision task where we use a large database of handwritten digits from 0 to 9 to build a model that recognizes those digits on an image. Usually, we use a convolutional neural networks (CNN) as a baseline. Our goal is to build a decision tree to perform the same task using distillation for explicability matters.*

- *We have 3 models: 1) Convnet (CNN) which is the teacher. The model is already trained and fine-tuned. It can be imported from Keras Python Library. 2) Soft Binary Decision Tree (SBDT) which is the student decision tree but trained traditionally without distillation independently from the teacher. 3) SBDT Distilled is the student trained using teacher's distillation.*

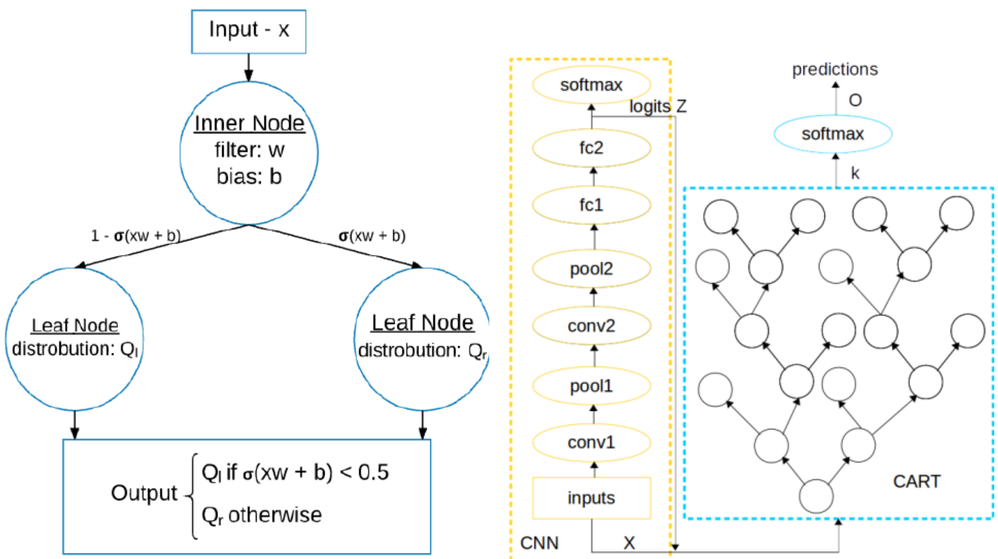


Fig 14. Soft Binary Decision Tree (SBDT) distillation architecture. Type of distillation is response-based trained in offline mode.

Model	Depth	Labels	Batch size	Epochs	Accuracy
Convnet (CNN)	-	Hard	16	12	99.29%
SBDT	4	Hard	4	40	80.88%
SBDT Distilled	4	Soft	4	40	90.71%

Table 10. Distillation Performance. Distillation training outperforms traditional training but performs worse than the teacher. However, we gain in explicability

DISTILLATION FOR NEURAL NETWORK EXPLANATION (2/2)

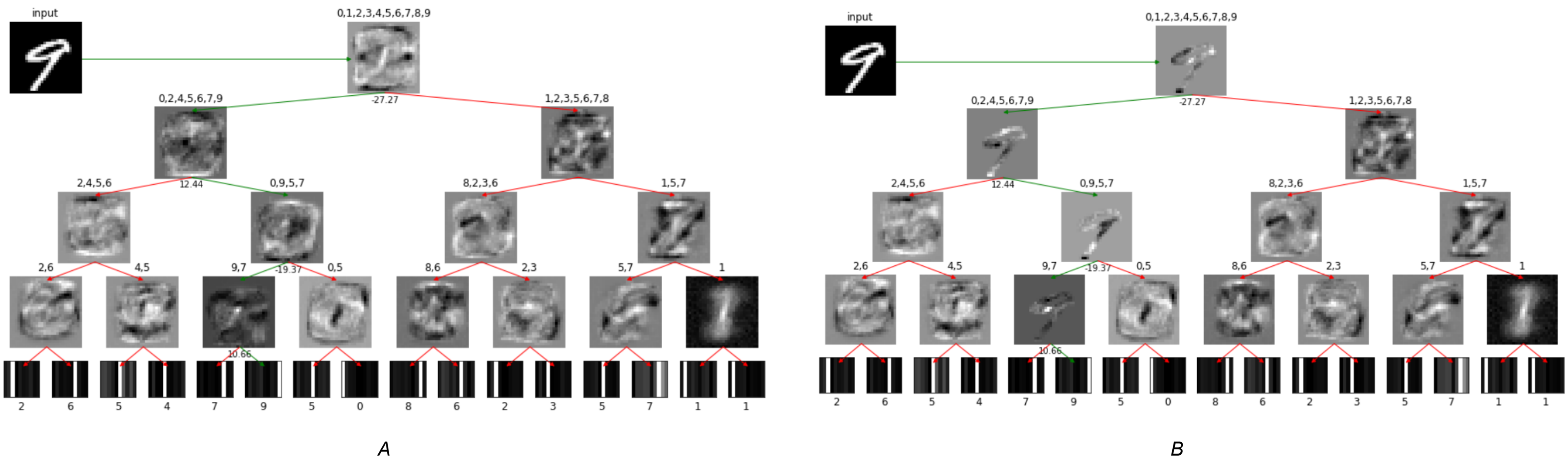
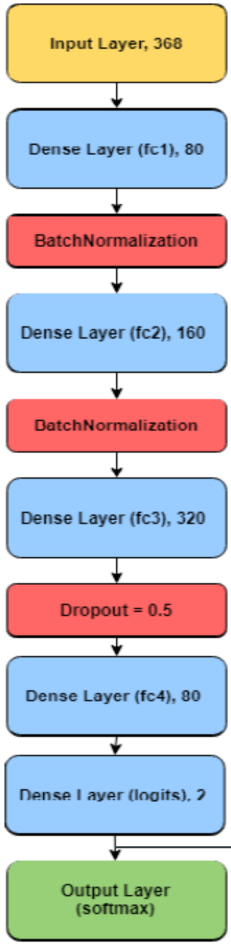


Fig 15. Tree's maximum probability path for Classification explanation; A) Explanation's filters provided by SBDT trained traditionally without distillation; B) Explanation's filters provided by SBDT with ConvNet distillation

LOGISTIC REGRESSION PERFORMANCE ENHANCING

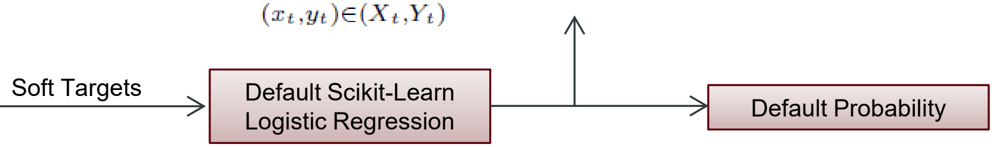


- *Lending Club dataset contains over 2 million loans issued between 2007 and 2018. The goal is to predict loan defaults based on the loan features and borrower characteristics such as loan amount, interest rate, loan grade, employment status, home ownership, credit score, and more.*
- *Here, we compare 3 models: 1) Feedforward neural network; 2) Logistic regression without distillation (baseline); 3) Logistic regression using distillation*

Dataset	Baseline Logistic Regression	Teacher Neural Network	Student Logistic Regression
Lending Club	0.5083	0.88	0.6507

Table 12. Distillation learning is an effective method to enhance simple models' performance.

$$\mathcal{L}_{KD} = \sum_{(x_t, y_t) \in (X_t, Y_t)} [\alpha \mathcal{L}_{CE}(f_S, x_t, y_t) + \beta \mathcal{L}_{KL}(f_S, f_T, x_t)]$$



	Apple	Pear	Banana	Car
Hard Targets	0	1	0	0
Soft Targets	0.1	0.9	10^{-5}	10^{-9}

Table 11. Soft targets contain richer information than hard targets

Fig 16. The teacher tuned based on the validation AUC Performance. The last Dense layer named 'logits' is given two hidden nodes, to match output and to obtain the logits for Knowledge Distillation. Multiple regularization methods were applied to the model, including L1 and L2 regularization, Dropout layers, and Batch Normalization layers. Additionally, multiple activation functions have been tested such as tanh and sigmoid, but ReLU provided the best overall performance. The optimizer that proved to give the best performance is the Adam optimizer, whilst also improving the speed of learning.

FUTURE WORK

TEST, MRM TOOL, MEMORY REDACTION

Task	Duration	Due Date
Test on PD Estimation models	1 month	7/ 10/23
Test on other MRM use cases: speech recognition, distilled GPT	1 month	7/30/23
Workshop 2: Tests' Results	1 day	to define
MRM Distillation Learning Tool: Packages development.	1 Month	9/25/23

Table. 13 Internship GANT Chart

REFERENCES

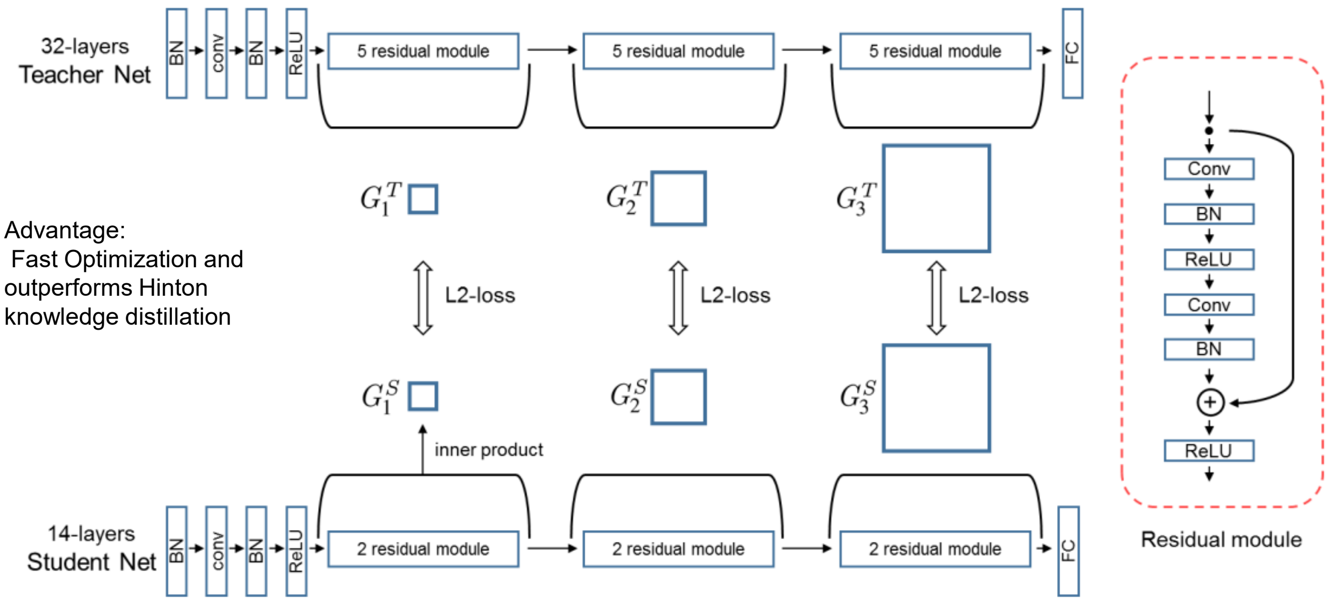
1. Craven and al., 1995: Extracting Tree-Structured Representations of Trained Networks
2. Caruana and al.,2006: model compression
3. Hinton and al.,2015: Distilling the Knowledge in a Neural Network
4. Han and al., 2015: Learning both Weights and Connections for Efficient
5. Hoffman and al., 2015: Cross Modal Distillation for Supervision Transfer
6. Zagoruyko and al., 2017: Attention Transfer
7. Huang and al., 2017: Knowledge Distill via Neuron Selectivity Transfer
8. Caruana and al., 2017: Interpretable & Explorable Approximations of Black Box Models
9. Yim and al, 2017: A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning
10. Burda, Edwards and al., 2018: Exploration by Random Network Distillation
11. Caruana and al., 2018: Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation
12. Liu and al., 2018: Improving the Interpretability of Deep Neural Networks with Knowledge Distillation
13. Asadulaev and al., 2019: Interpretable Few-Shot Learning via Linear Distillation
14. Bastani and al., 2019: Interpreting Blackbox Models via Model Extraction
15. Zhang and al, 2021: Adversarial co-distillation learning for image recognition

APPENDIX (1/3)

EXAMPLE OF RELATION-BASED DISTILLATION FRAMEWORK

A novel example is distilling the flow of solution process (FSP) matrix Yim and al, 2017. The FSP matrix is generated by crossing feature representation of two selected layers across data using inner product, gram representation, or other capturing-information product depending on the problem.

Advantage:
Fast Optimization and outperforms Hinton knowledge distillation



Algorithm 1 Transfer the distilled knowledge

Stage 1: Learning the FSP matrix
Weights of the student and teacher networks: W_s, W_t
1: $W_s = \arg\min_{W_s} L_{FSP}(W_t, W_s)$
Stage 2: Training for the original task
1: $W_s = \arg\min_{W_s} L_{ori}(W_s)$

$$G_{i,j}(x;W) = \sum_{s=1}^{h}\sum_{t=1}^{w} \frac{F^1_{s,t,i}(x;W) \times F^2_{s,t,j}(x;W)}{h \times w}$$

$$L_{FSP}(W_t, W_s) = \frac{1}{N}\sum_{x}\sum_{i=1}^{n} \lambda_i \times \|(G_i^T(x;W_t) - G_i^S(x;W_s)\|_2^2$$

Fig 17. Complete architecture of FSP KD. There are two stages. In stage 1, the student network is trained to minimize the distance between the FSP matrices of the student and teacher networks. Then, the pretrained weights of the student DNN are used for the initial weight in stage 2. Stage 2 represents the normal training procedure

Example of FSP construction as in Yim and al, 2017: 1) Take the feature map matrix of layer i and layer j. 2) Construct the Gramian matrix; 3) sum up over all terms; 4) repeat the process across selected layers. NB: 1) the number of selected layers must be the same in the teacher and the student; 2) this is only a construction example.

APPENDIX (2/3)

EXAMPLE OF SELF-DISTILLATION FRAMEWORK

Self-distillation is a variant of online distillation where the teacher and student share the same architecture. It is often used to enhance neural networks performance comparing to the traditional training mode. As a framework's example, adversarial co-distillation (ACN) by Zhang and al, 2021 is a novel technique to enhance the performance of a CNN in the image recognition task by generating divergent examples where models do not totally agree. The goal is to have them make the same prediction accurately based on a majority vote mindset.

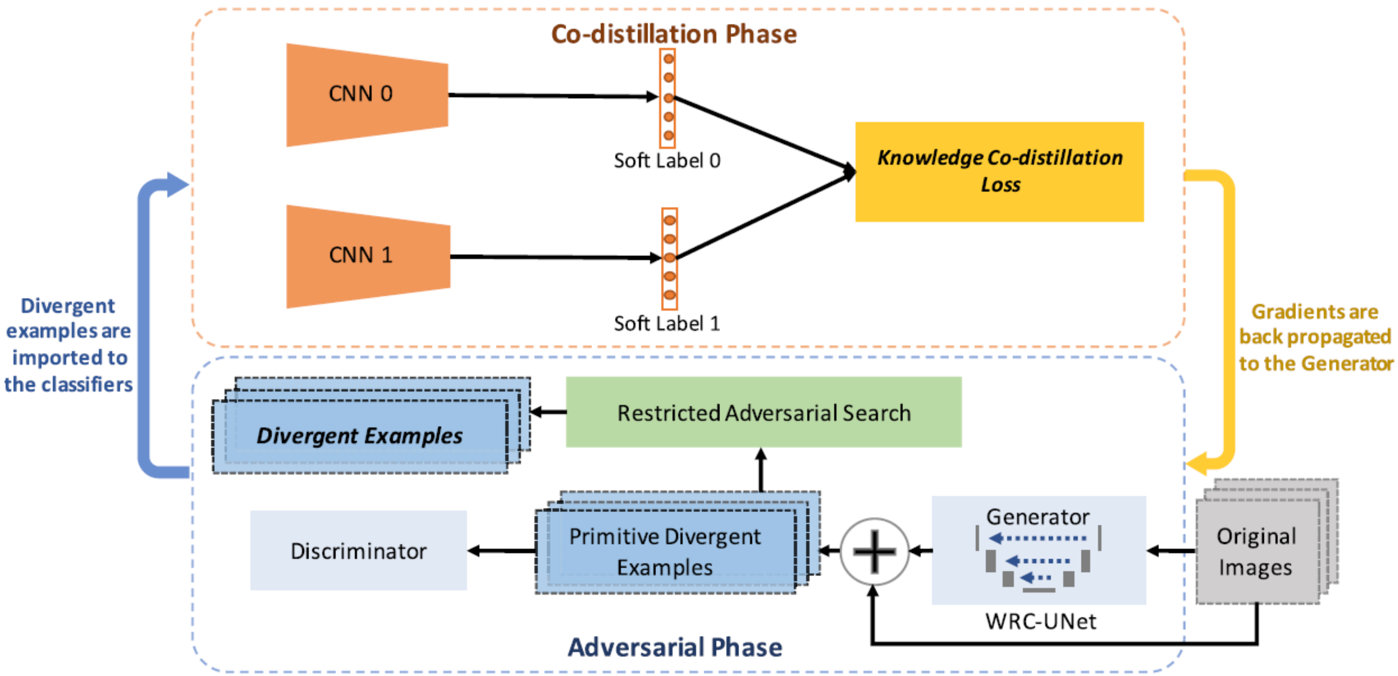


Fig. 18. The framework illustration of ACNs. ACNs consist of an Adversarial Phase and a Co-distillation Phase. The Adversarial Phase generates the divergent examples, and the Co-distillation Phase learn the divergent examples. The Adversarial Phase is designed according to the GANs framework.

Model	Original trained	ACN
Resnet-20	68.22%	70.67%
VGG11	67.38%	70.11%
AlexNet	39.45%	46.27%

Table 14. Distillation learning can be used to enhance complex models' performance without compression

APPENDIX (3/3)

OTHER APPLICATION EXAMPLES

Article	Use	Task Description	Baseline	Teacher Model	Student Model Performance	Limitations	Distillation Mode
Liu and al., 2018	XAI, ESM	MNIST	DT (acc: 84%)	CNN (acc: 99.25%)	DT (acc: 86.6 %)	Unfaithfulness Risk	Offline, ResK
Che and al., 2015	XAI, ESM, ETM	Medical setting, VENT dataset, Mortality Task, Binary Classification	GBT (AUC: 72%)	DNN (AUC : 73%) SDA (AUC : 74%) LSTM (AUC : 76.55%)	GBTmimic-DNN (AUC: 77%) GBTmimic-SDA (AUC :78%) GBTmimic-LSTM (AUC: 75.5%)	GBT lack of interpretably Complex Student	Offline, ResK
Cachola and al., 2022	XAI, ESM, NLP	Medical setting, MIMIC-III, Assigning clinical notes to ICD codes	Logistic Regression (Micro-AUC: 93%)	DR CAML (Micro-AUC: 97.2 %) HAN (Micro-AUC: 96.7%) Trans ICD (Micro-AUC: 92%)	Linear Regression: DR CAML (Micro-AUC: 96.7 %) HAN (Micro-AUC: 96.2%) Trans ICD (Micro-AUC: 90%)	Unfaithfulness Risk	Offline, ResK
Caruana and al., 2018	XAI, ESM, ETM	COMPAS, Predicting recidivism risk	Linear Model (AUC: 73%) RF (AUC: 73%) iGAM (AUC: 74%)	COMPAS, Unknown Model (Acc: Unknown, average 65%)	iGAM (acc: 75%)	Unfaithfulness Risk	Offline, ResK

Table 15. Relevant work in knowledge distillation especially in XAI, ESM, ETM applications.

C'EST VOUS
L'AVENIR

SOCIETE
GENERALE