StatusNeo

Mysteries In Deep Learning

"Modern age" neural network training—at least for image classification tasks and many other tasks is typically extremely stable under now-standard procedures like over-parameterization, batch-normalization, and adding residual links. No matter what random initialization or random data order is used during the training, the learnt models consistently outperform the unlearned ones using standard neural network designs and training methods (usually SGD with momentum). For instance, the mean test accuracy is 81.51 percent while the standard deviation is only 0.16 percent when the same WideResNet-28-10 architecture is trained on the CIFAR-100 dataset ten times with different random seeds.

Making neural Network more stable:

Gradient descent is used to train neural networks, and a subset of the training dataset is used to estimate the error that will be used to update the weights.
The batch size is a crucial hyperparameter that affects the dynamics of the learning algorithm. It refers to the quantity of examples from the training dataset used in the estimate of the error gradient.

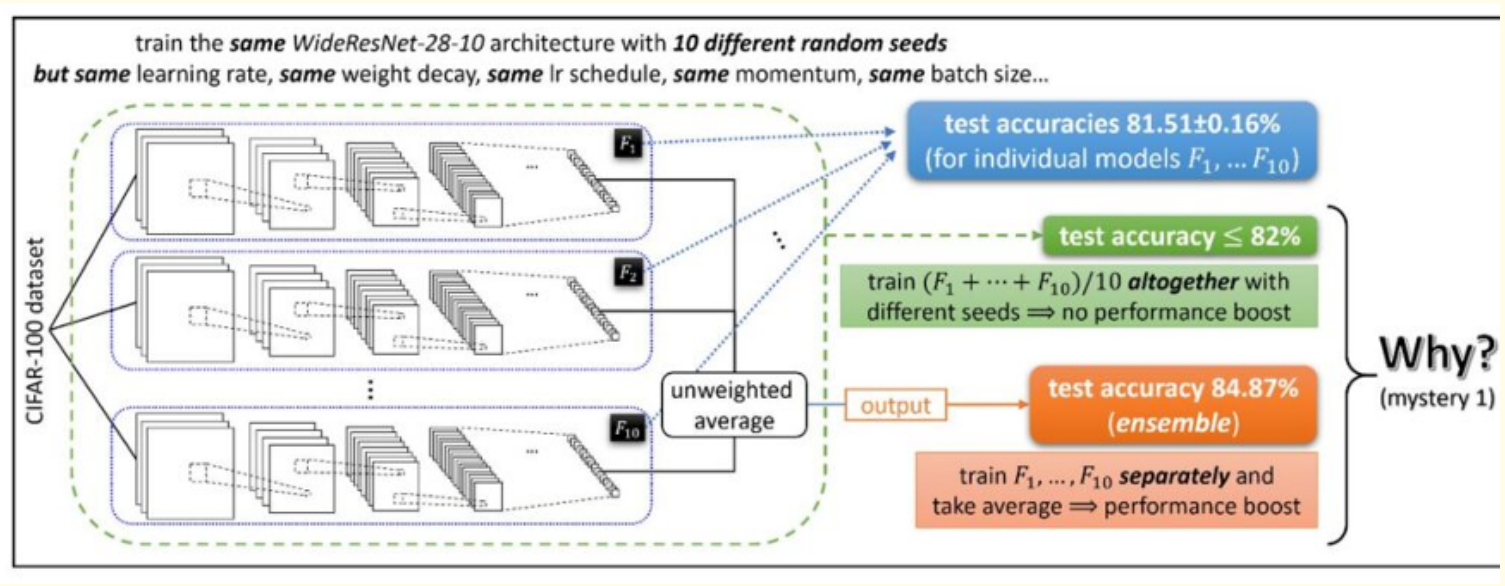
Mysteries In Deep Learning:

- **What is the Mystery?**

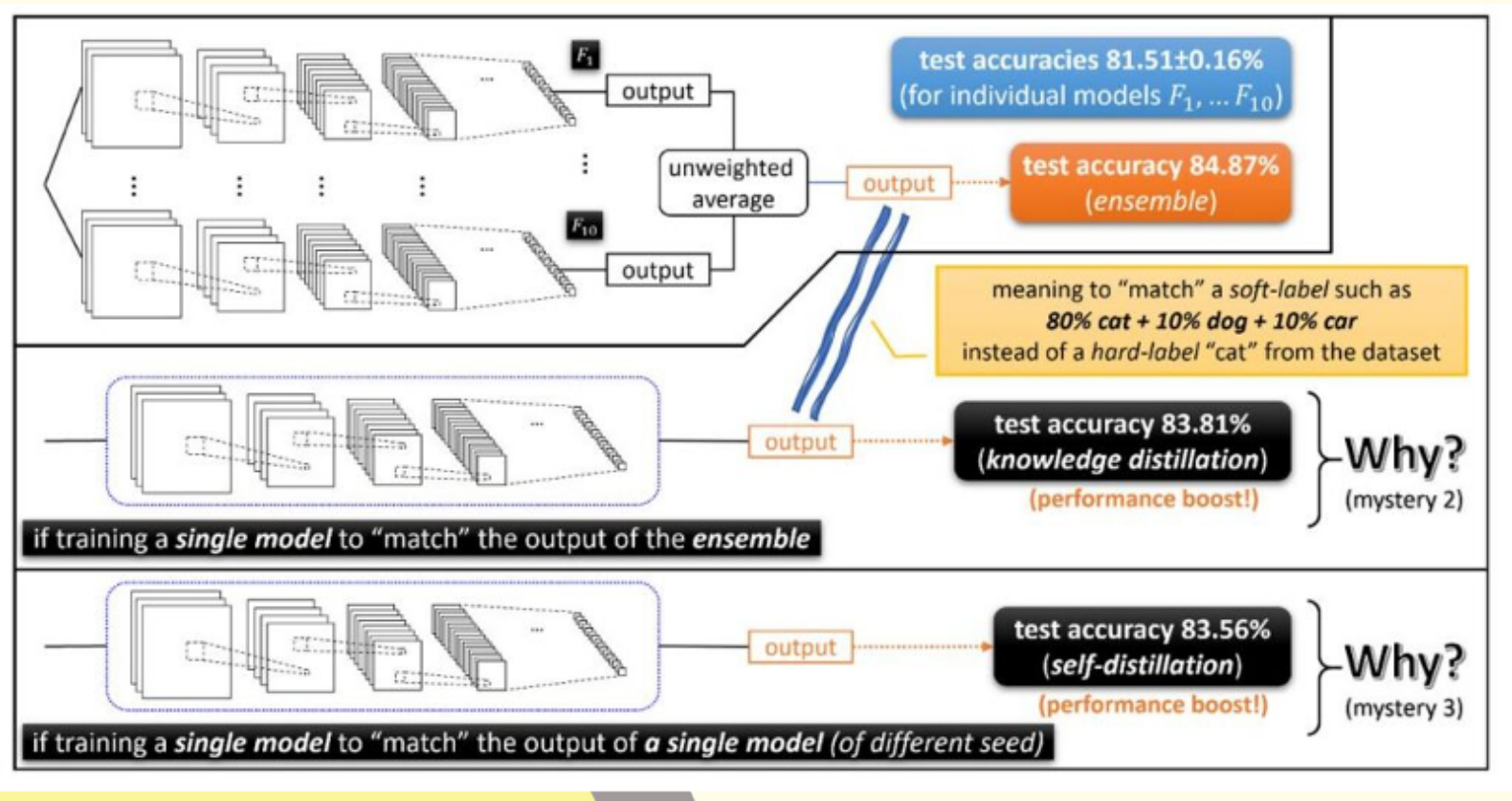
The easiest example from real life we can take is to consider the state of the economy. There are seven billion people on the planet, but none of us are particularly difficult in terms of economics, right? We have some wants and preferences as well as some financial resources. Then we make a purchase. Ultimately, the math that explains a person's economic conduct is not that difficult.
But when we combine seven billion of these, it becomes very challenging to predict how the global economy is doing and how it will act in a year. Even yet, there are still certain inherent risks, such the abrupt appearance of the coronavirus. Similar to that is the mystery surrounding neural networks: in modern models, there are hundreds of millions or even billions of units talking with one another, yet the aggregate behaviour is not well-understood mathematically.

There are three mysteries in Deep Learning.

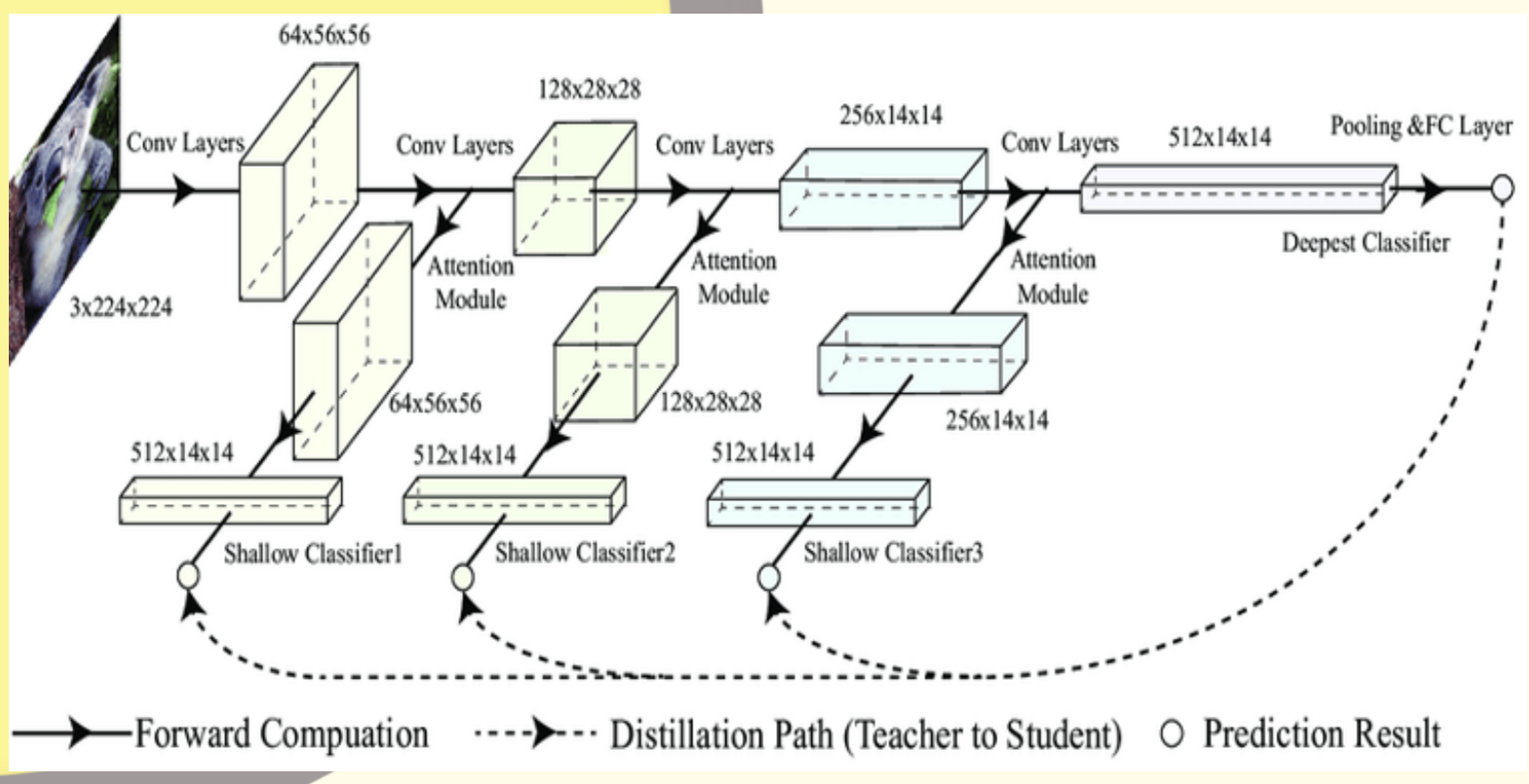
1. **Ensemble**: Multiple separate models are combined using ensemble learning to improve generalisation performance. Deep learning models with multilayer processing architecture are now outperforming shallow or conventional classification models in terms of performance.
   - High performance ensemble models are incredibly expensive and time consuming to run. A single model can be trained using the knowledge distillation technique to match the performance of the ensemble. The second enigma of ensembles is furthered by the effectiveness of knowledge distillation techniques.
   - **Why does it works with a sudden performance boost?**
     It is the model of Deep learning model performance, and it is enhanced by ensemble or model averaging techniques. Simply averaging the results of a few neural networks that were independently trained using the same training data set that will perform. It is well known that using this approach will result in significantly higher prediction accuracy over the test set than using only one model individually. The performance improvements will also hold true if all architectures are the same and if the same training data set and training algorithm are used.



2. **Knowledge Distillation:** Knowledge distillation is the idea of compressing a model by employing a larger, fully trained network to gradually teach a smaller network exactly what to perform. The output feature maps produced by the larger network after each convolution layer are referred to as "soft labels." The smaller network is then instructed to mimic the larger network's outputs at every level to learn its identical behaviour.



3. **Self-Distillation:** The third mystery of ensemble is very related to the second one but even more puzzling. Knowledge distillation shows that a smaller model can match the performance of a bigger ensemble. But a parallel phenomenon is referred to as a self-distillation and is even more puzzling. Self-distillation relies on the performing knowledge distillation against the individual models which can also increases the performance.
Basically, the self-distillation relies on the training the identical model using itself as an educator.



Q1) Deep learning Ensembles vs. Feature Mapping

The most popular type of ensemble learning is random feature mappings, which train models using a random distribution of features. This kind of method is well understood and performs well in linear models, making it a good starting point for evaluating the effectiveness of deep learning ensembles. Deep learning ensembles exhibited behaviour that was very comparable to that of feature mappings, according to preliminary Microsoft Research experiment findings. The process of knowledge distillation isn't quite the same, though.

Random Feature Mappings *(e.g. NTK or other neural kernel methods)*	accuracy of directly training the average of 10 models	$\gg$ ensemble accuracy (over 10)	$\gg$ individual model accuracies	$\geq$ knowledge distillation / self-distillation accuracies
WRN-10-4-***NTK*** on CIFAR10	72.86%	70.54%	66.68%	66.01% / 61.92%
WRN-10-4-***NTK*** on CIFAR100	41.47%	38.32%	31.90%	31.38% / 27.64%
	(more examples in our paper)			
Deep Learning	ensemble accuracy (over 10)	$\approx$ knowledge distillation / self-distillation accuracies	$\gg$ accuracy of directly training the average of 10 models	$\approx$ individual model accuracies
WRN-28-10 on CIFAR10	97.20%	97.22% / 97.13%	96.46%	96.70±0.21%
WRN-28-10 on CIFAR100	84.69%	83.81% / 83.56%	81.83%	81.51±0.16%
	(more examples in our paper)			

- **Ensemble versus reducing variance of Individual models**

Based on the makeup of the data, one of the study's most illuminating findings comes from Microsoft Research. Each class of the data in a multi-view dataset is based on a structure with several view features. For instance, the headlights, wheels, or windows of an automobile image can help identify it as a car. According to Microsoft Research, datasets having multi-view structures are more likely to improve ensemble model performance than datasets without such features.

- **Can Ensemble reduce the variance?**

Yes, it measures the precision or specificity of the match and had a high variance means a weak match.

- **Multi-view data: A fresh method to support the Deep Learning Ensembles**

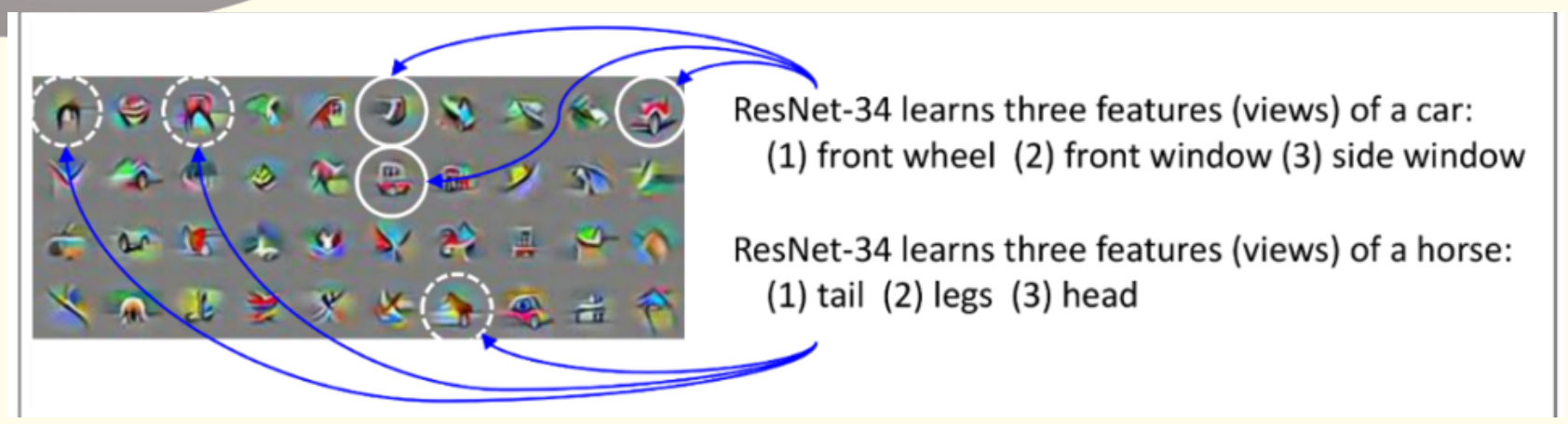
To create more accurate models, multi-view learning incorporates information from various heterogeneous sources and makes use of their complementing data.

In multi-instance learning, examples are represented by labelled bags that contain collections of instances. Due to the varying cardinality and feature space of the various multi-instance views, data fusion of these views cannot be simply concatenated into a single set of features.

The ensemble strategy that is suggested in this research combines view learners and seeks consensus among the weighted class predictions to benefit from the complimentary information from many viewpoints.

Importantly, the ensemble must cope with the many feature spaces that each of the views brings in, even though the views may only fully represent the data for the bags.

Let us take an example: A car image can be classified as a car by looking at the headlights, the wheels, or the windows. We all can see all those characteristics for a normal placement of a car in photos, and we only need to use one of the characteristics to identify it as a car. However, there are some car images taken from a particular angle, where one or more of these features are missing Like, an image of a car facing forward might be missing the wheel feature.



This Structure, where each class of the data contains numerous view features, it is known as a multi-view. All the view features will appear in most of the data, but some view features may not be be present.

The multi-view structure can be seen in both the intermediate layers and the input pixel space.

- **Knowledge distillation: Requiring a single model to understand various viewpoints**

As we continue to demonstrate the process of the Knowledge distillation in this new study. In actual situation, certain car photos could appear “more like a cat” than others; as an illustration, some of car images might have headlights that resemble cat eyes.

In these circumstances, the ensemble model can offer useful dark knowledge, such as the vehicle picture X is 10% like a cat.

The Observation: When the training an individual level the remaining views can still correctly identify image (assuming) X as a car during the training of a single neural network model, they cannot be used to match the hidden knowledge that “image X is 10% like a cat”.

In other words, the individual model is compelled to learn every view characteristic feasible in order to match the performance of the ensemble during the knowledge distillation.

The key to deep learning’s knowledge distillation is that each model, which is a neural network, performs feature learning and is thus capable of learning every aspect of the ensemble.

	CIFAR10 test accuracy				CIFAR100 test accuracy			
	single model (over 10)	ensemble (over 10)	10 runs of knowledge distill	ensemble over knowledge distill	single model (over 10)	ensemble (over 10)	10 runs of knowledge distill	ensemble over knowledge distill
ResNet-28-2	95.22±0.14%	96.33%	95.89±0.07%	96.21%	76.38±0.23%	81.13%	78.94±0.21%	80.35%
ResNet-34	93.65±0.19%	94.97%	94.37±0.13%	94.88%	71.66±0.43%	76.85%	73.57±0.34%	75.60%
ResNet-34-2	95.45±0.14%	96.55%	96.00±0.12%	96.42%	77.01±0.35%	81.48%	79.43±0.23%	81.56%
ResNet-16-10	96.08±0.16%	96.80%	96.73±0.07%	96.76%	80.03±0.17%	83.18%	82.51±0.14%	83.36%
ResNet-22-10	96.44±0.09%	97.12%	97.01±0.09%	97.09%	81.17±0.23%	84.33%	83.54±0.19%	84.27%
ResNet-28-10	96.70±0.21%	97.20%	97.06±0.08%	97.24%	81.51±0.16%	84.69%	83.75±0.16%	84.87%

- **Self-distillation: Implicitly integrating ensemble and knowledge distillation is self-distillation**

Self-distillation also provides the theoretical justification for knowledge in this new work. There seems to be some performance benefit from training one individual model to match the output of another identical individual model but using a different random seed.
At a higher level, we see self-distillation as a more condensed version of the ensemble and knowledge distillation. One can anticipate that an individual model F2 will only learn a portion of the features based on its own random initialization when learning the model from a random initialization to match the output of an independently trained individual model F1.

Additionally, the F2 is motivated to learn the portion of characteristics that F1 has already learnt. This procedure can be thought of as “Ensemble learning two individual models Like F1, F2 and distilling It to F2”.

- **Conclusion and aim to go forward**

In this work, we show best of out knowledge, this paper is the first theoretical demonstration of how deep learning ensembles function. To back up our theory and our “multi-view” data premise, we also offer the empirical data. What we think that the different contexts can use our framework.
For instance, random cropping used to enhance the data may be seen as an additional strategy for forcing the network to acquire “multi-views”. But the fact is that our improved theoretical understanding of how neural networks learn features during the training, along with the new principled methodologies, it should let us in practise create neural networks with the test accuracy that is as good as or better than ensembles.