Qianru Sun 孙倩茹

Singapore Management University

Learning to learn
from small data

Source: Movie Scene from Pirates of the Caribbean

What to learn? e.g. image classification

L. Fei-Fei and O. Russakovsky, Analysis of Large-Scale Visual Recognition, Bay Area Vision Meeting, October, 2013
L. Fei-Fei, ImageNet: crowdsourcing, benchmarking & other cool things, CMU VASC Seminar, March, 2010

What to learn? e.g. image classification

Sea lion

L. Fei-Fei and O. Russakovsky, Analysis of Large-Scale Visual Recognition, Bay Area Vision Meeting, October, 2013
L. Fei-Fei, ImageNet: crowdsourcing, benchmarking & other cool things, CMU VASC Seminar, March, 2010

What to learn? e.g. image classification

Multiple classes

Strawberry

Backpack

Traffic light

Sea lion

Bathing cap

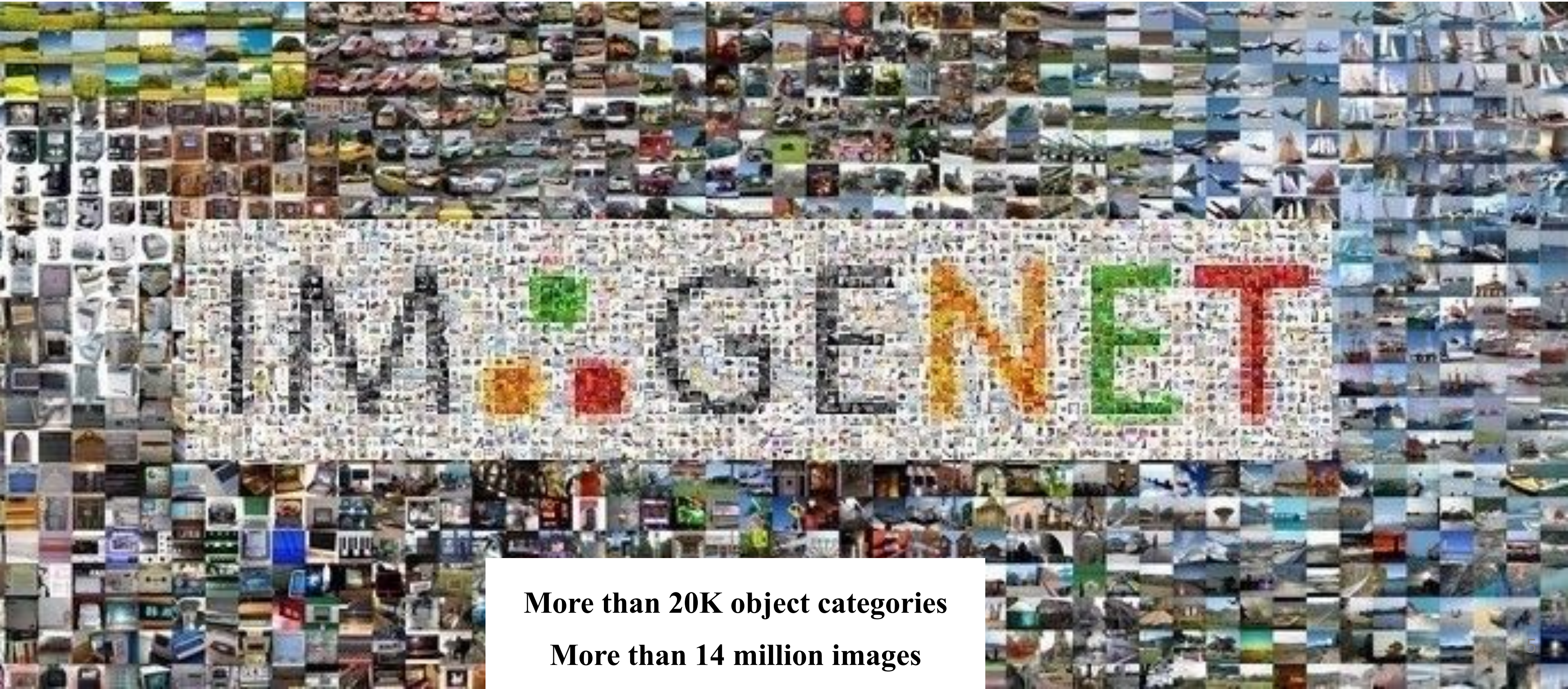
Matchstick

Racket

Flute

L. Fei-Fei and O. Russakovsky, Analysis of Large-Scale Visual Recognition, Bay Area Vision Meeting, October, 2013
L. Fei-Fei, ImageNet: crowdsourcing, benchmarking & other cool things, CMU VASC Seminar, March, 2010

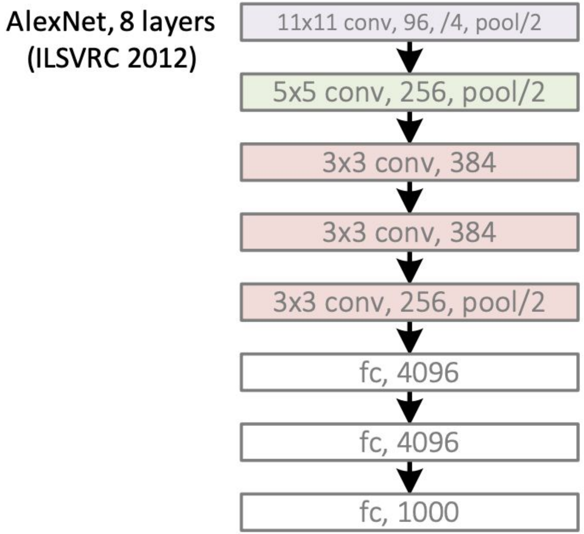
What to learn? benchmark: ImageNet



More than 20K object categories

More than 14 million images

How to learn? Deep Neural Networks



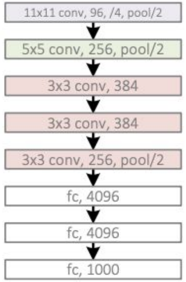
ILSVRC 2012:

ImageNet Large Scale Visual Recognition Competition in 2012 (Ended)

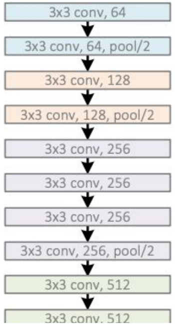
End-to-end training.
Easy to chopped up, modified, retrained.

A. Krizhevsky, I. Sutskever, G.E. Hinton. Imagenet Classification with Deep Convolutional Neural Networks. NIPS 2012.

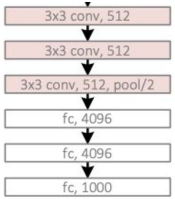
How to learn? Deeper Neural Networks



AlexNet, 8 layers
(ILSVRC 2012)



VGG, 19 layers
(ILSVRC 2014)



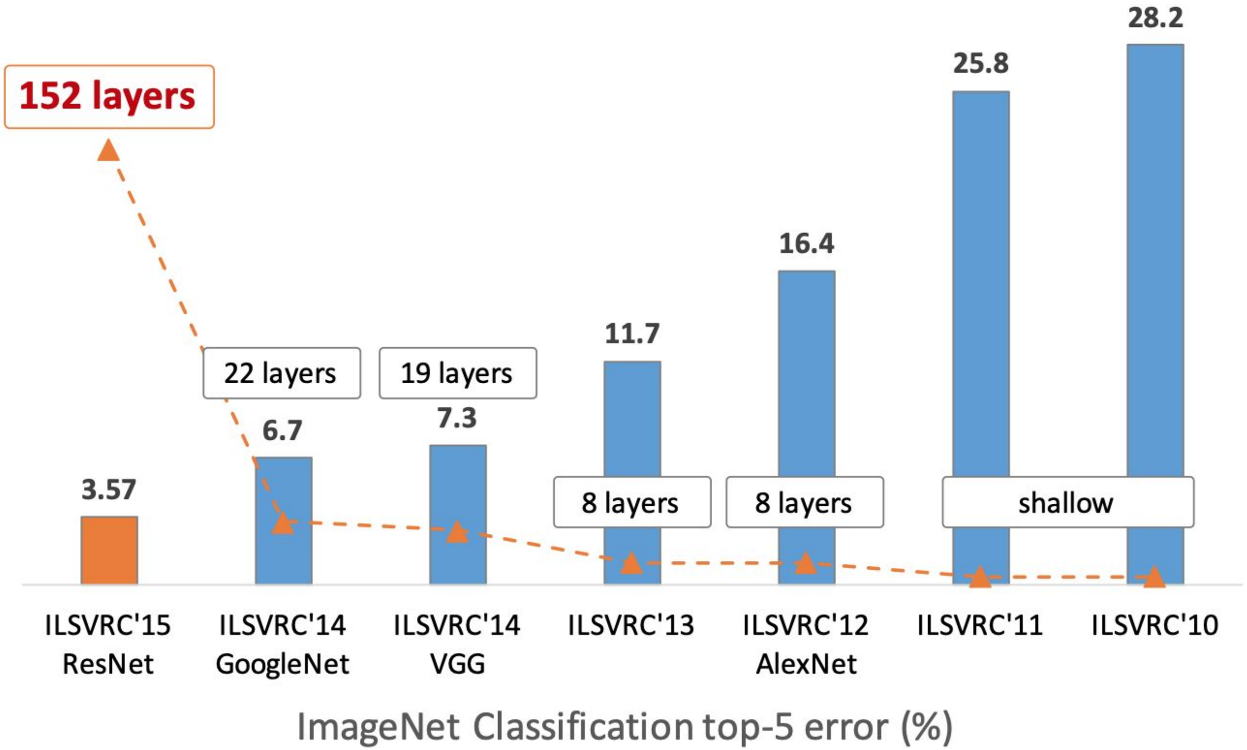
GoogleNet, 22 layers
(ILSVRC 2014)

ResNet, 152 layers
(ILSVRC 2015)

The best paper at CVPR'16

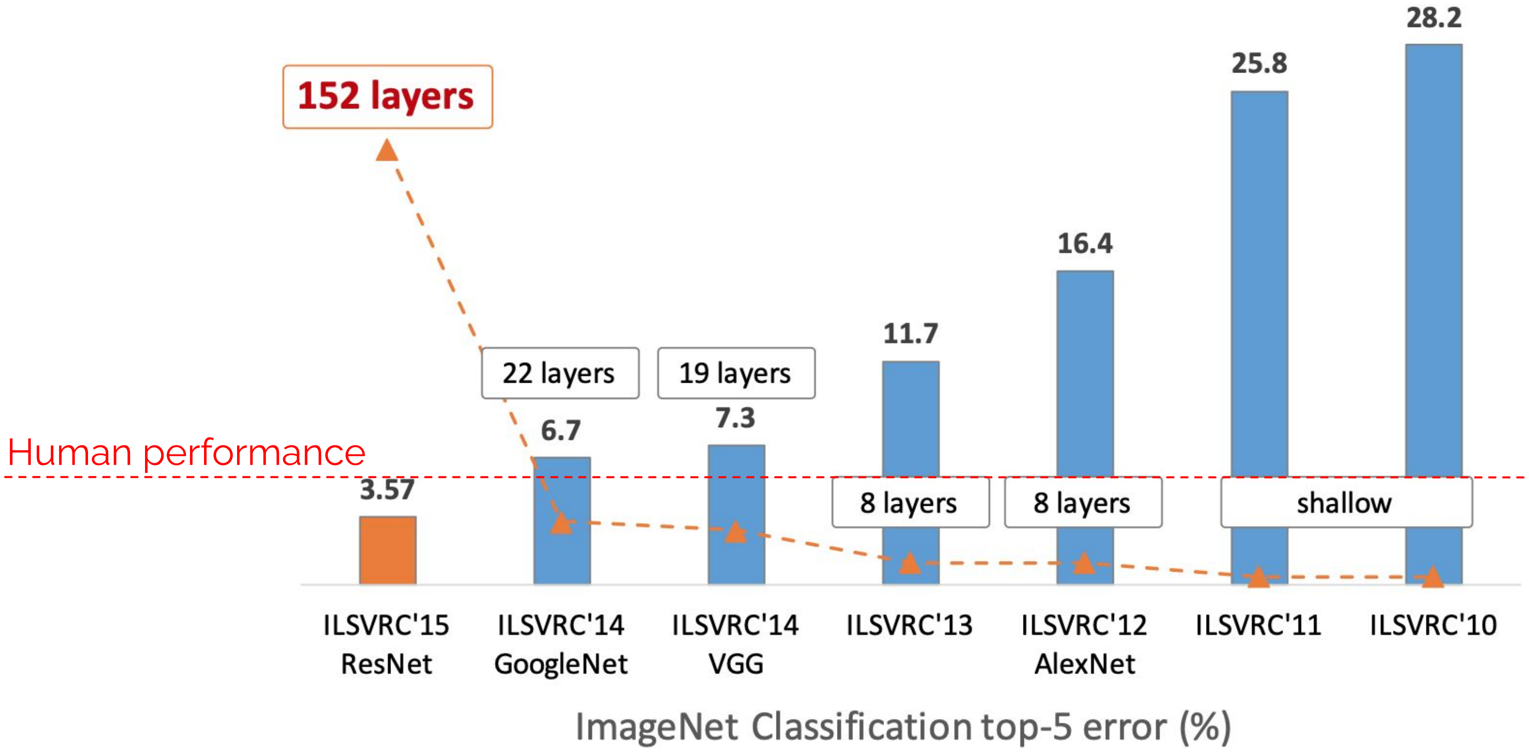
K. He, X. Zhang, S. Ren, and J. Sun. Identity Mappings in Deep Residual Networks. ECCV 2016. (survey)

How they perform? e.g. on ImageNet



K. He, X. Zhang, S. Ren, and J. Sun. Identity Mappings in Deep Residual Networks. ECCV 2016. (survey)

Even better than human!



K. He, X. Zhang, S. Ren, and J. Sun. Identity Mappings in Deep Residual Networks. ECCV 2016. (survey)

Even better than human!

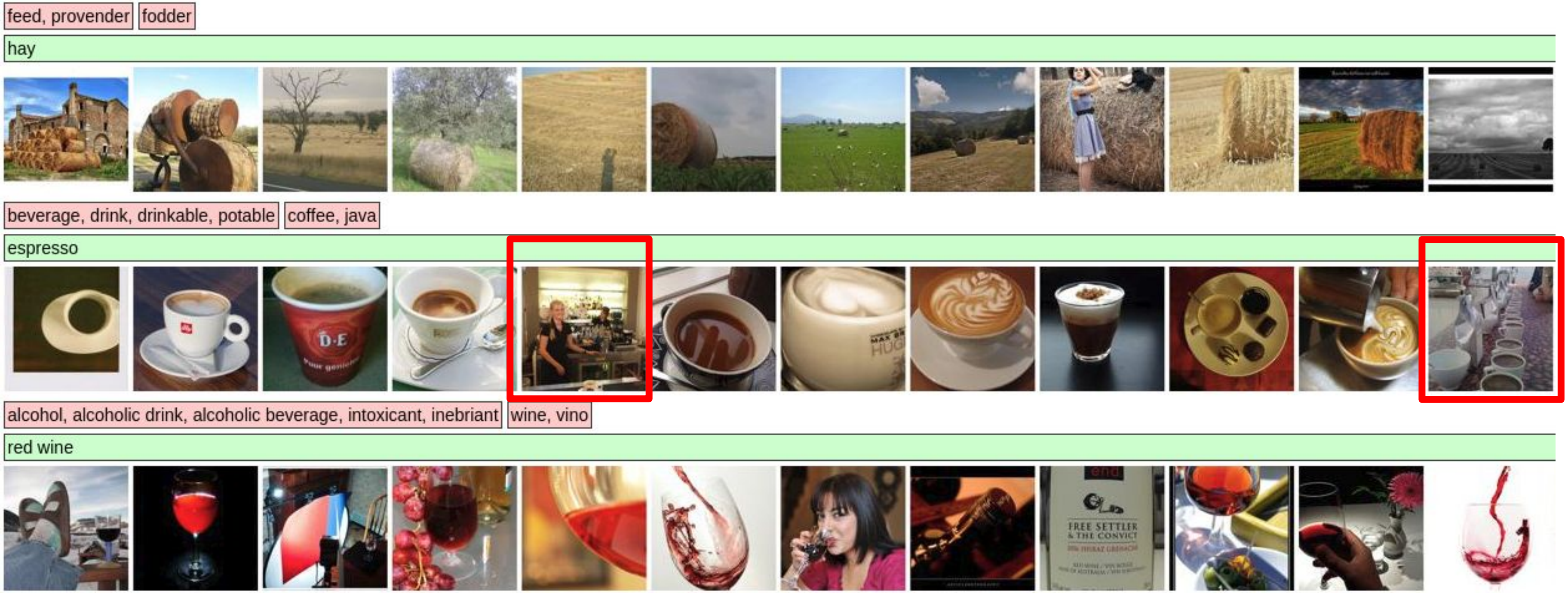
Learn from scratch!
Learn by self-playing!
Learn with ResNet!

https://deepmind.com/blog/article/alphago-zero-starting-scratch

Learning to learn
from small data

Why from small data?
in the era of big data

Big data is expensive!



https://cs.stanford.edu/people/karpathy/ilsvrc/

Big data is expensive!
e.g. to label the ImageNet

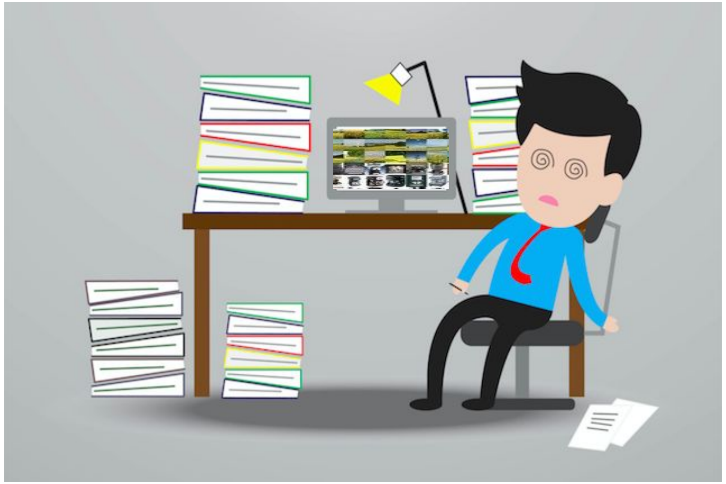
of classes: **40,000**

# of images per class: **10,000**

# of people needed to verify: **3**

Speed of verifying: **2 images/second**

$$40{,}000 \cdot 10{,}000 \cdot 3 / 2 = 600{,}000{,}000 \text{sec} \quad 19 \text{years}$$



No graduate student would like to do this 19 yrs project !

L. Fei-Fei, ImageNet: crowdsourcing, benchmarking & other cool things, CMU VASC Seminar, March, 2010.

Big data is not "personalized"!

e.g. person image generation

Input pose | Input image | **Large-scale training results** | **Learning to learn results** | Ground truth

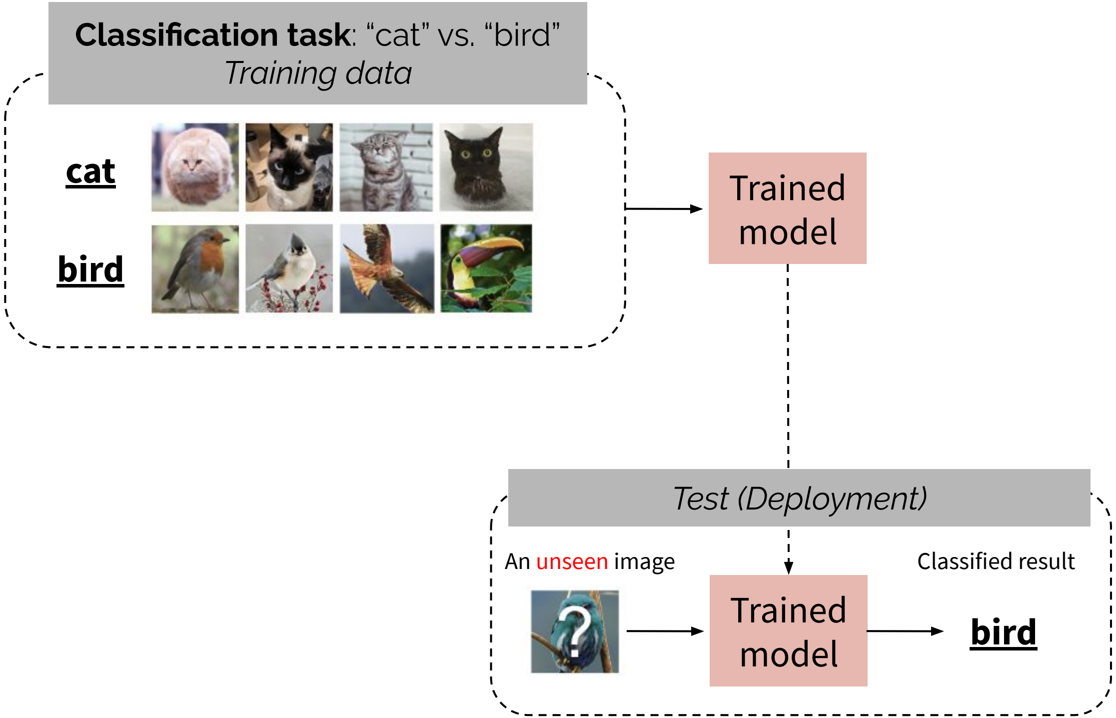
bad quality :(
missing the gender and age

good quality :)
closer to the ground truth

L. Ma, Q. Sun, B. Schiele, L. Van Gool, A Novel BiLevel Paradigm for Image-to-Image Translation. arXiv 2019.

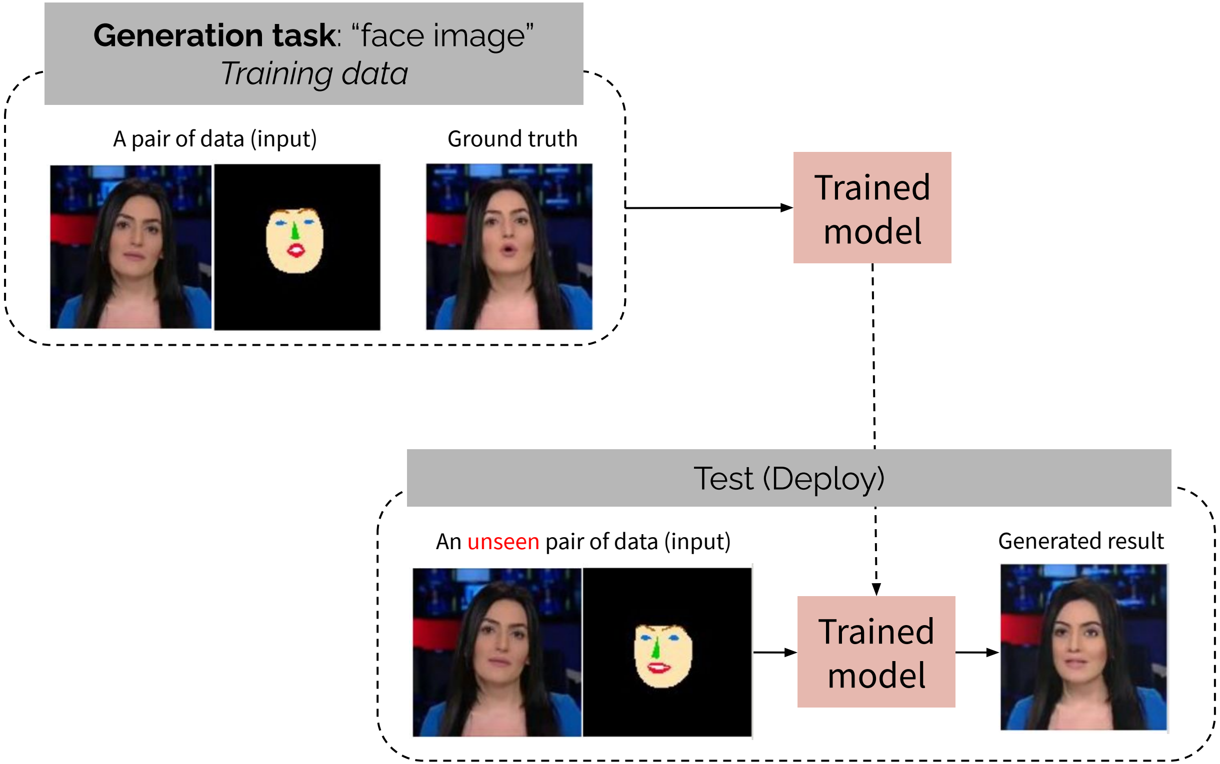
Small data is cheap! e.g. classification

Only 4 labeled examples per class



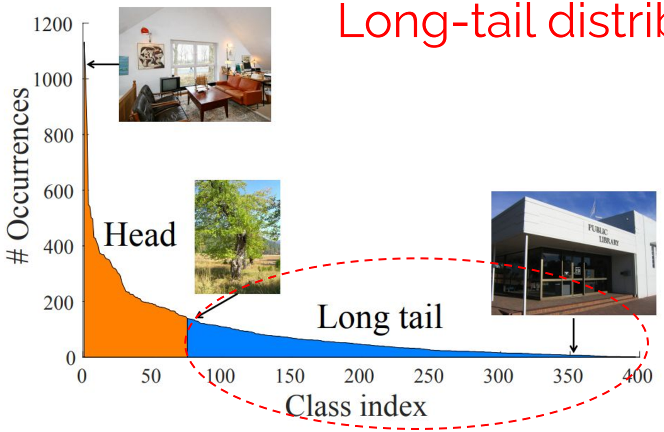
Small data is cheap! e.g. generation

Only 1 pair of data per identity



Small data is around us!

Long-tail distribution, expensive images, expensive annotation …



Long-tail distribution

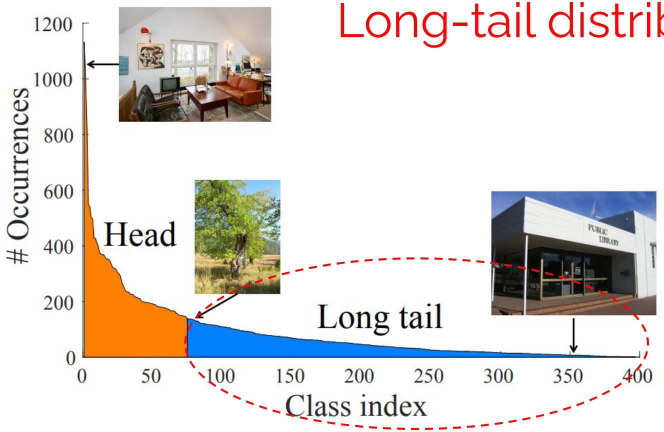
Long-tail distribution on the SUN-397 dataset.

Y. Wang, D. Ramanan and M. Hebert. Learning to Model the Tail. NIPS 2017.

Small data is around us!

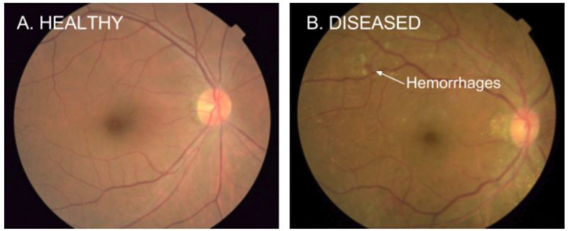
Long-tail distribution, expensive images, expensive annotation …

Long-tail distribution



Long-tail distribution on the SUN-397 dataset.

Detect Diabetic Retinopathy Rare images



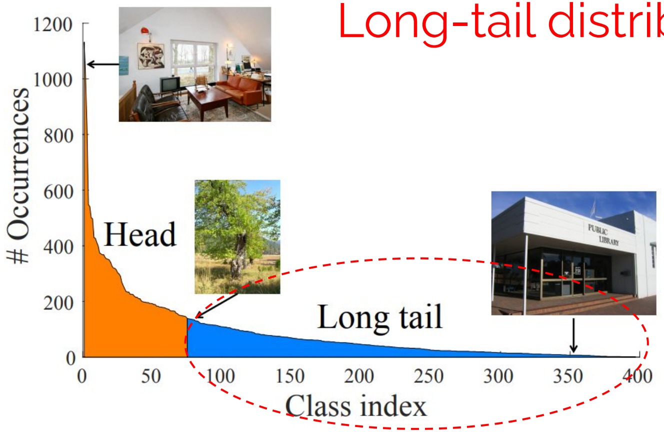
Diabetic Retinopathy, a diabetes complication that affects eyes. Source: https://ai.googleblog.com/2016/11/deep-learning-for-detection-of-diabetic.html

Y. Wang, D. Ramanan and M. Hebert. Learning to Model the Tail. NIPS 2017.

Small data is around us!

Long-tail distribution, expensive images, expensive annotation ...

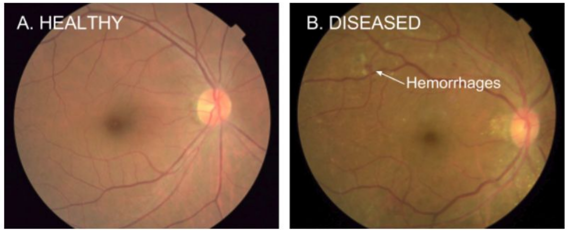
Long-tail distribution



Long-tail distribution on the SUN-397 dataset.

Y. Wang, D. Ramanan and M. Hebert. Learning to Model the Tail. NIPS 2017.

Detect Diabetic Retinopathy Rare images



Diabetic Retinopathy, a diabetes complication that affects eyes. Source: https://ai.googleblog.com/2016/11/deep-learning-for-detection-of-diabetic.html

Expensive annotation

Input Image | Semantic Segmentation | Instance Segmentation

Small data is around us!

Long-tail distribution, expensive images, expensive annotation ...

The realistic example is often as follows ...

Instance Segmentation

Small data is easy for human!

[1] B. M. Lake, R. Salakhutdinov, and J.B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 2015.

[2] E. G. Miller, N.E. Matsakis, and P.A. Viola. Learning from one example through shared densities on transformations. *CVPR*, 2000.

[3] B. M. Lake, R. Salakhutdinov, and J.B. Tenenbaum. Concept learning as motor program induction: A large-scale empirical study. *Annual Conference of the Cognitive Science Society*, 2012.

Easy for you to remember me by one glance…

Why? A conservative estimate of images a person has seen.

Let's assume a person sees a distinct image every 30 seconds.

By the time a person enters his/her 25 years old,

he/she has seen:

= 25 years * 365.24 days/year *16 hours/day * 60 minutes/hour * 2 images / minute

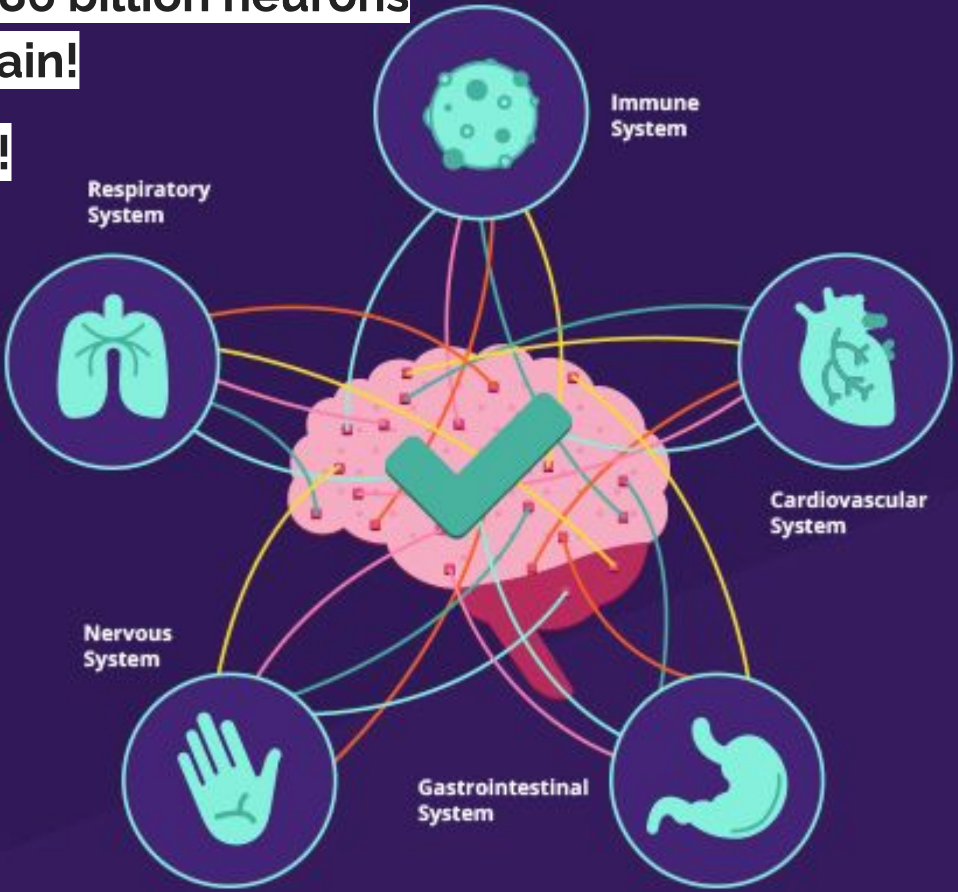
= 17, 531, 520 images

More than an ImageNet (14 million images)!

https://towardsdatascience.com/why-small-data-is-the-future-of-ai

AND, more that 86 billion neurons in the human brain!

A huge network!

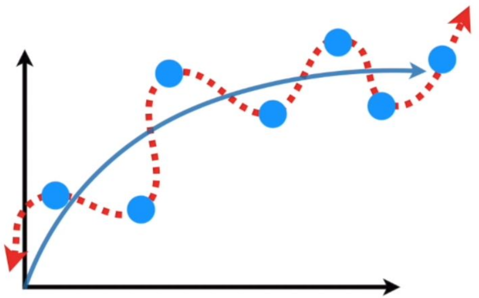


https://www.brainfacts.org/core-concepts/your-complex-brain

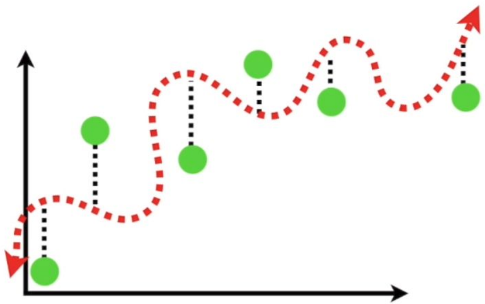
Small data is hard for machine!

Two main problems:

- Over-fitting becomes much harder to avoid.
- Outliers become much more dangerous.



Overfitting in training

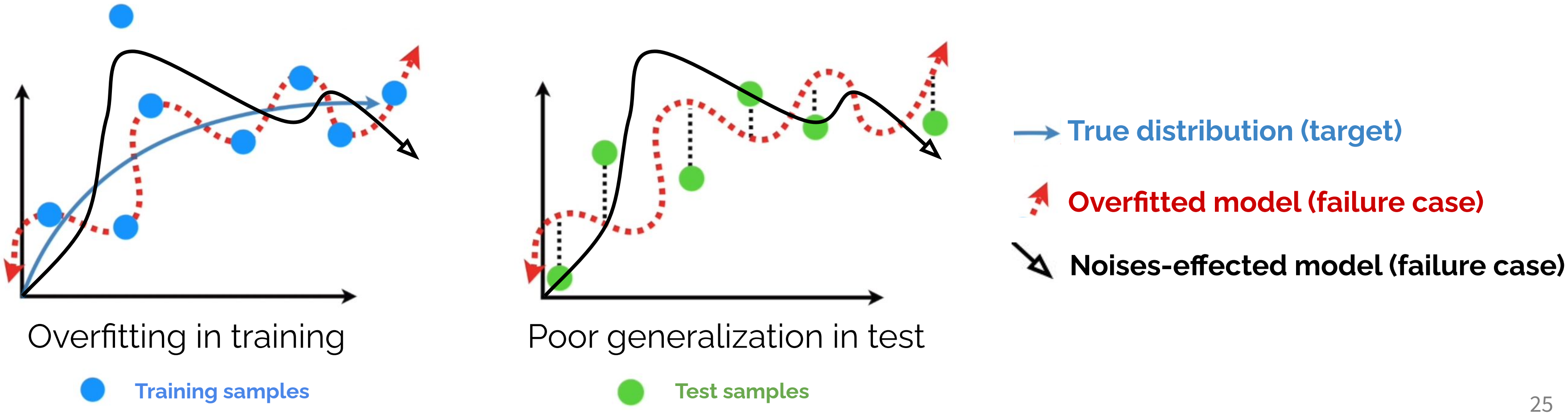


Poor generalization in test

Small data is hard for machine!

Two main problems:

- Over-fitting becomes much harder to avoid.
- Outliers become much more dangerous.



Can machine learn from humans?

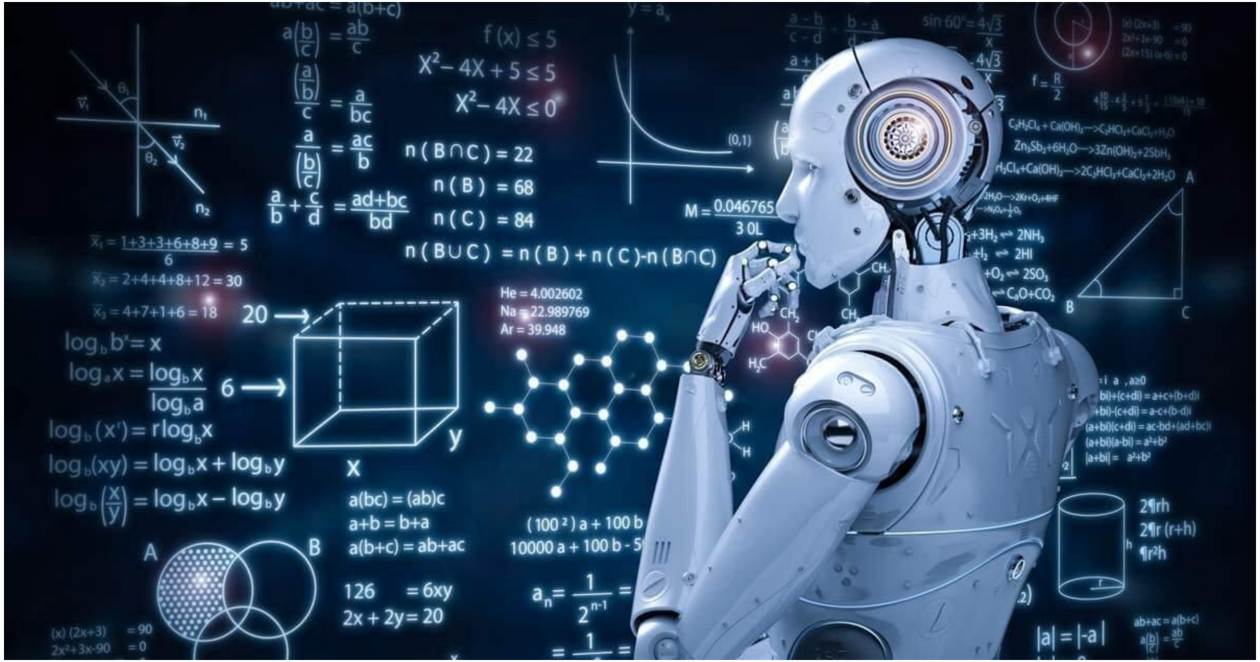
Humans learn experiences from related tasks

⇩

“meta-learning” or “learning to learn”

(Machine Learning)

Meta-Learning (Learning to learn)



zdnet.com

My recent works - Learning to learn
for different capabilities in computer vision tasks

Learning to **transfer** “memory” --- “experiences from other large-scale tasks”

Learning to **extract** “data” --- “images potentially useful for future training”

Learning to **combine** “models” --- “trained network parameters”

My recent works - Learning to learn
for different capabilities in computer vision tasks

Learning to **transfer** “memory” --- “experiences from other large-scale tasks”

Learning to **extract** “data” --- “images potentially useful for future training”

Learning to **combine** “models” --- “trained network parameters”

Learning to transfer **“memory”**

Given an existing model pre-trained on a large-scale task

Given the conditions

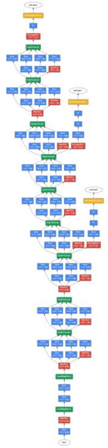
Any big dataset

Any network
(scratch)

training

<u>knowledge</u>

Network parameters
(learned)

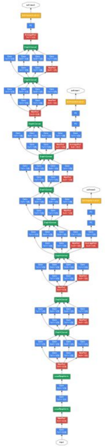


Learning to transfer **“memory”**

The general memory of image patterns is already in the model

knowledge

Network parameters
(learned)



A small dataset
e.g. 4-shot

cat

bird

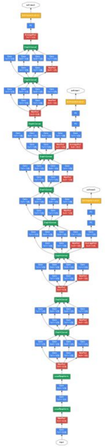
Few-shot
model

Learning to transfer **"memory"**

A traditional method is "Fine-Tuning"

knowledge

Network parameters
(learned)



A small dataset
e.g. 4-shot

cat

bird

Fine-Tuning(FT) [Pan et al. 2011]

Few-shot
model

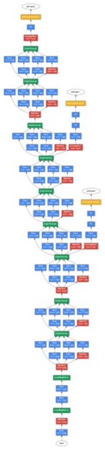
S. J. Pan, et al. Domain Adaptation via Transfer Component Analysis. IEEE Trans. Neural Networks, 2011.

Learning to transfer **“memory”**

A traditional method is “Fine-Tuning”, while it is problematic!

knowledge

Network parameters
(learned)



A small dataset
e.g. 4-shot

cat

bird

Few-shot model

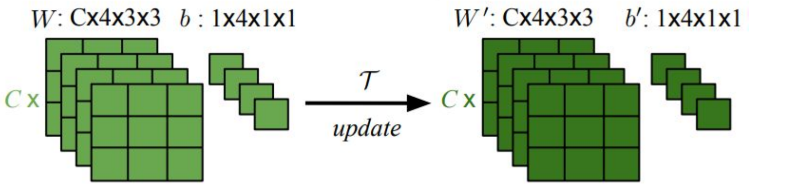
Fine-Tuning(FT) [Pan et al. 2011]
Problem: “catastrophic forgetting”
[Lopez-Paz et al NIPS’17]

S. J. Pan, et al. Domain Adaptation via Transfer Component Analysis. IEEE Trans. Neural Networks, 2011.
D. Lopez-Paz and M. Ranzato. Gradient episodic memory for continual learning. In NIPS, 2017.

Learning to transfer "memory"

<u>Fine-Tuning</u> [Pan et al. 2011]

Problem: "catastrophic forgetting"

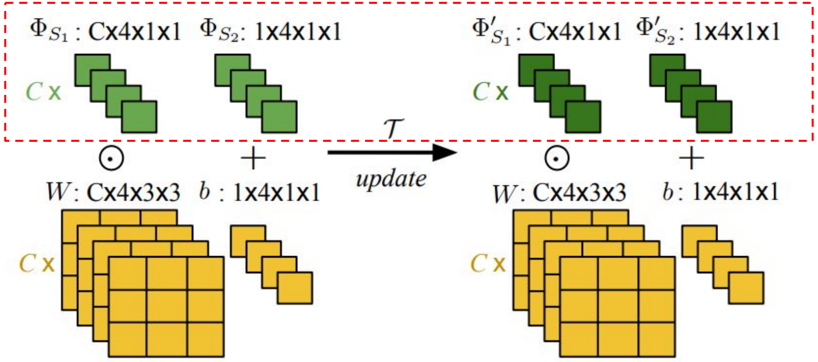


(a) Parameter-level *Fine-Tuning* (*FT*)

<u>Meta-Transfer Learning</u> [Sun et al. CVPR'19]

Scaling and shifting (*SS*) of frozen neurons

Only SS parameters get updated.

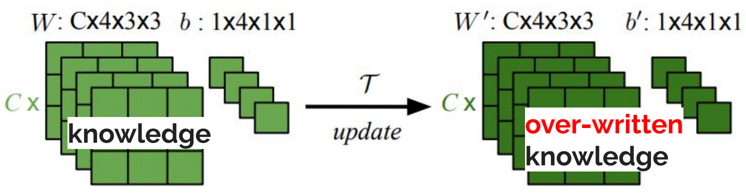


(b) Our *Scaling* S_1 and *Shifting* S_2

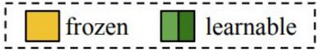
S. J. Pan, et al. Domain Adaptation via Transfer Component Analysis. IEEE Trans. Neural Networks, 2011.
Q. Sun, et al. Meta Transfer Learning for Few-Shot Learning. CVPR 2019.

Learning to transfer **"memory"**

Fine-Tuning [Pan et al. 2011]
Problem: "catastrophic forgetting"

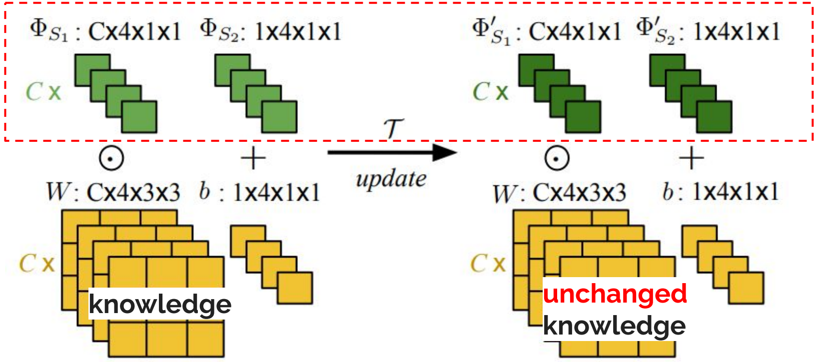


(a) Parameter-level *Fine-Tuning* (*FT*)



Meta-Transfer Learning [Sun et al. CVPR'19]
Scaling and shifting (*SS*) of frozen neurons

Only SS parameters get updated.



(b) Our *Scaling* S_1 and *Shifting* S_2

S. J. Pan, et al. Domain Adaptation via Transfer Component Analysis. IEEE Trans. Neural Networks, 2011.
Q. Sun, et al. Meta Transfer Learning for Few-Shot Learning. CVPR 2019.

Learning to transfer **"memory"**

Meta-Transfer Learning [Sun et al. CVPR'19]

Results on two benchmarks: **miniImageNet** [Vinyals et al. NIPS'16]; **FC100** [Oreshkin et al. NeurIPS'18]

Table. Classification accuracy (%). Higher is better.

	miniImageNet		FC100		
	1 (shot)	5	1	5	10
update $[\Theta; \theta]$	45.3	64.6	38.4	52.6	58.6
update θ	50.0	66.7	39.3	51.8	61.0
FT θ	55.9	71.4	41.6	54.9	61.1
FT $[\Theta4; \theta]$	57.2	71.6	40.9	54.3	61.3
FT $[\Theta; \theta]$	58.3	71.6	41.6	54.4	61.2
SS $[\Theta4; \theta]$	59.2	73.1	42.4	55.1	61.6
SS $[\Theta; \theta]$**(Ours)**	60.2	74.3	43.6	55.4	62.4

Update[...]: Without "learning to learn"

Our improvements(minimum) over *Update[...]*:
miniImageNet: **10.5%(1-shot)** and 7.6%(5-shot);
FC100: **4.3%(1-shot)**, 3.6%(5-shot) and 1.4%(10-shot).
Ours performs better on lower-shot settings.

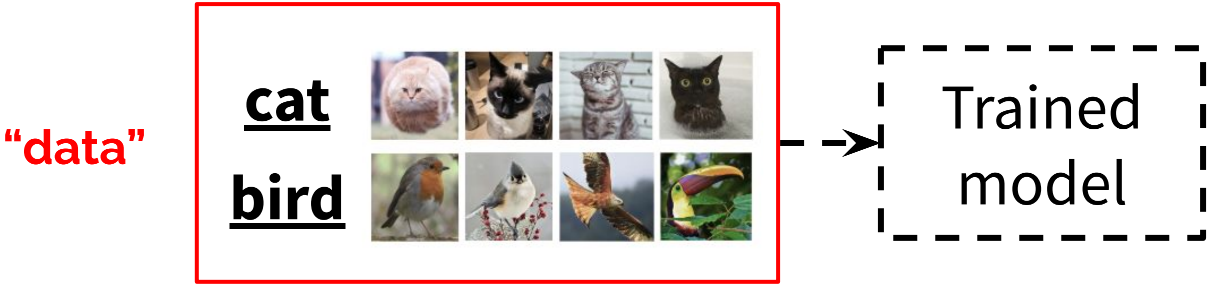
Q. Sun, et al. Meta Transfer Learning for Few-Shot Learning. CVPR 2019.
O. Vinyals, et al. Matching Networks for One Shot Learning. NIPS 2016.
B. N. Oreshkin, et al. TADAM: Task Dependent Adaptive Metric for Improved Few-Shot Learning. NeurIPS 2018.

My recent works - Learning to learn
for different capabilities in computer vision tasks

Learning to **transfer** “memory” --- “experiences from other large-scale tasks”

Learning to **extract** “data” --- “images potentially useful for future training”

Learning to **combine** “models” --- “trained network parameters”

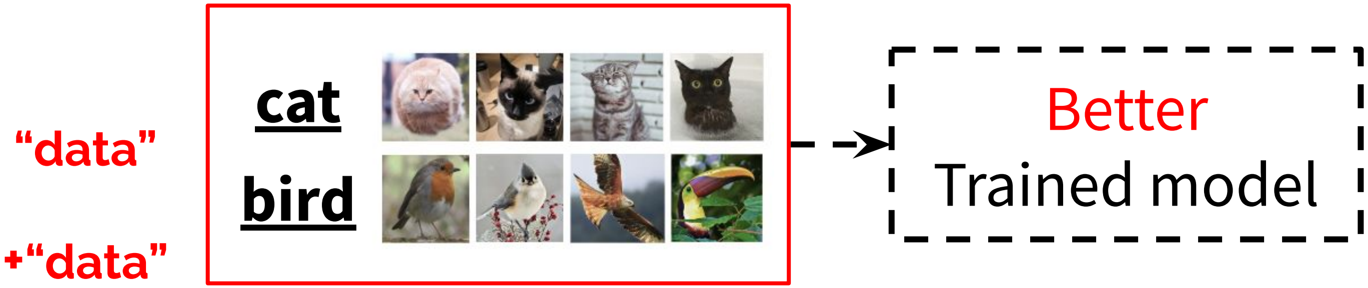


My recent works - Learning to learn
for different capabilities in computer vision tasks

Learning to **transfer** “memory” --- “experiences from other large-scale tasks”

Learning to **extract** “data” --- “images potentially useful for future training”

Learning to **combine** “models” --- “trained network parameters”



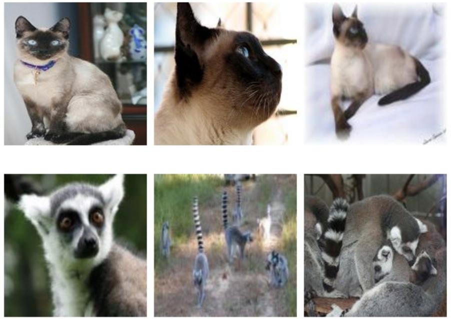
Learning to extract "data"

"data" + "unlabeled data"

cat

bird

+

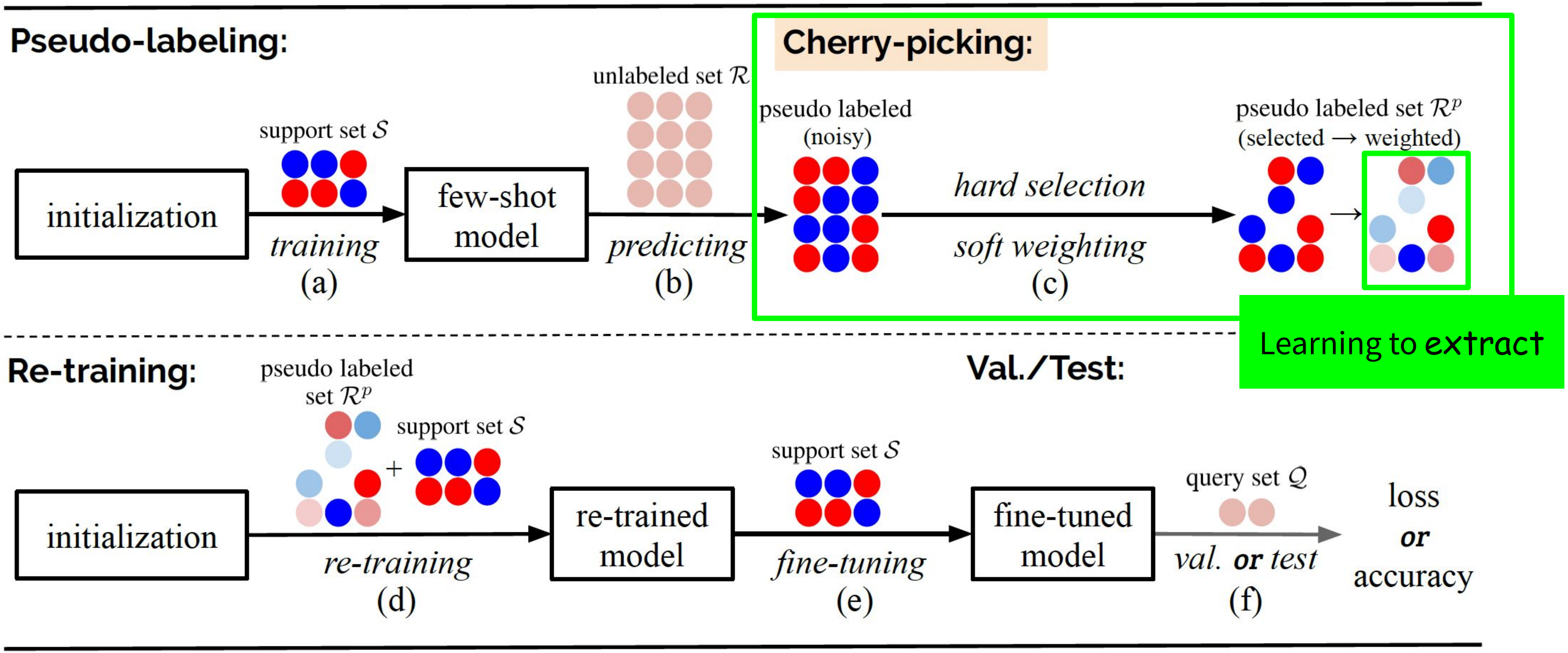


Better
Trained model

Semi-supervised learning
(data is cheap but label is expensive)

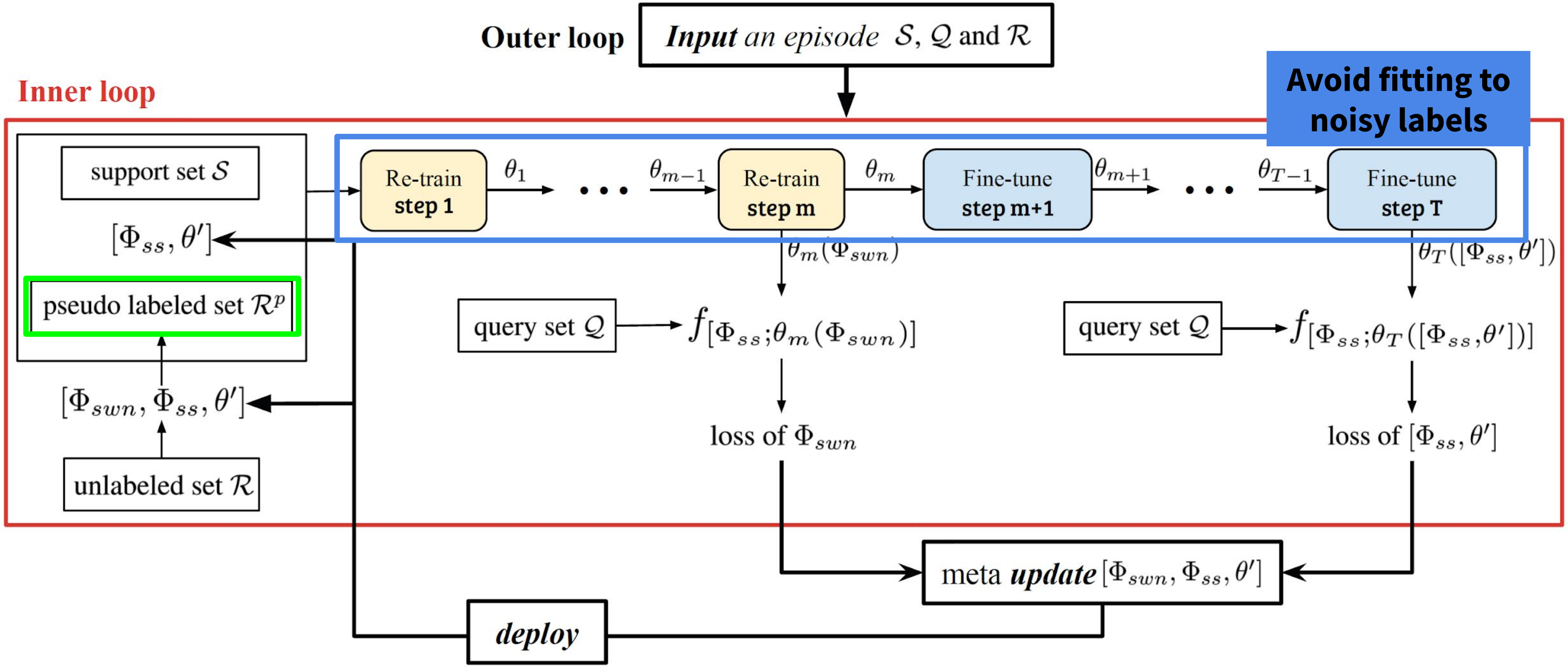
Xinzhe Li, et al. Learning to self-train for semi-supervised few-shot classification. NeurIPS 2019.

Learning to extract "data"



Xinzhe Li, et al. Learning to self-train for semi-supervised few-shot classification. NeurIPS 2019.

Learning to extract **"data"**



Xinzhe Li, et al. Learning to self-train for semi-supervised few-shot classification. NeurIPS 2019.

Learning to extract "data"

$\mathcal{D}$ for Distracting classes

		mini		tiered		mini w/$\mathcal{D}$		tiered w/$\mathcal{D}$	
		1(shot)	5	1	5	1	5	1	5
fully supervised (upper bound)		80.4	83.3	86.5	88.7	-	-	-	-
no meta	*no selection*	59.7	75.2	67.4	81.1	54.4	73.3	66.1	79.4
	hard	63.0	76.3	69.8	81.5	61.6	75.3	68.8	81.1
	recursive,hard	64.6	77.2	72.1	82.4	61.2	75.7	68.3	81.1
meta	*hard* (Φ_{ss}, θ')	64.1	76.9	74.7	83.2	62.9	75.4	73.4	82.5
	soft	62.8	75.9	73.1	82.8	61.1	74.6	72.1	81.7
	hard,soft	65.0	77.8	75.4	83.4	63.7	76.2	**74.1**	82.9
	recursive,hard,soft	**70.1**	**78.7**	**77.7**	**85.2**	64.1	**77.4**	73.5	83.4
	mixing,hard,soft	66.2	77.9	75.6	84.6	**64.5**	76.5	73.6	**83.8**
Masked Soft k-Means *with* MTL		62.1	73.6	68.6	81.0	61.0	72.0	66.9	80.2
TPN *with* MTL		62.7	74.2	72.1	83.3	61.3	72.4	71.5	82.7
Masked Soft k-Means [24]		50.4	64.4	52.4	69.9	49.0	63.0	51.4	69.1
TPN [13]		52.8	66.4	55.7	71.0	50.4	64.9	53.5	69.9

+5.5%

Our Ablation Study

Comparable to Ours

Table 2: Classification accuracy (%) in ablative settings (middle blocks) and related SSFSC works

A significant improvement by using "learning to extract"!!!

Xinzhe Li, et al. Learning to self-train for semi-supervised few-shot classification. NeurIPS 2019.
MTL: Qianru Sun, et al. Meta Transfer Learning for Few-Shot Learning. CVPR 2019.

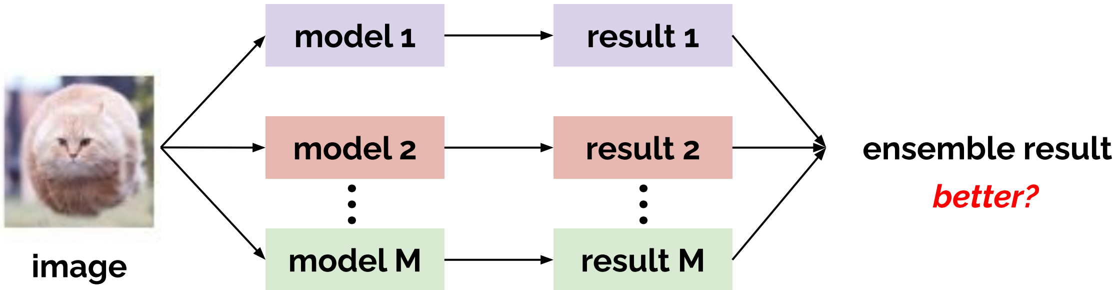
My recent works - Learning to learn
for different capabilities in computer vision tasks

Learning to **transfer** “memory” --- “experiences from other large-scale tasks”

Learning to **extract** “data” --- “images potentially useful for future training”

Learning to **combine** “models” --- “trained network parameters”

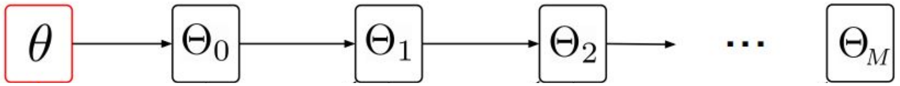
idea:



Learning to customize **models**

"Multiple models" == "models with different architectures, learning rates, data inputs, loss functions …"

initialization **model 1** **model 2** **model *M***

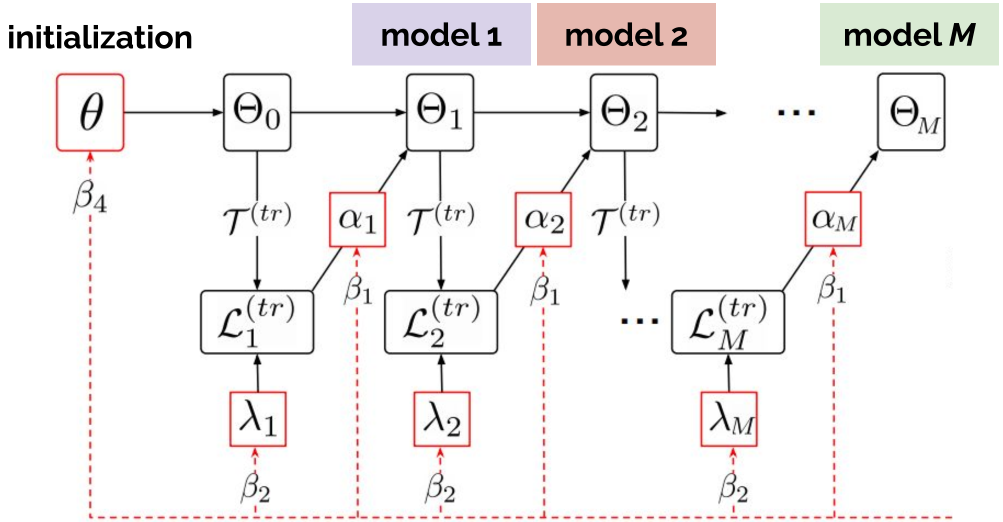


How to get?

Yaoyao Liu, et al. An Ensemble of Epoch-wise Empirical Bayes for Few-Shot Learning. ECCV 2020.

Learning to customize **models**

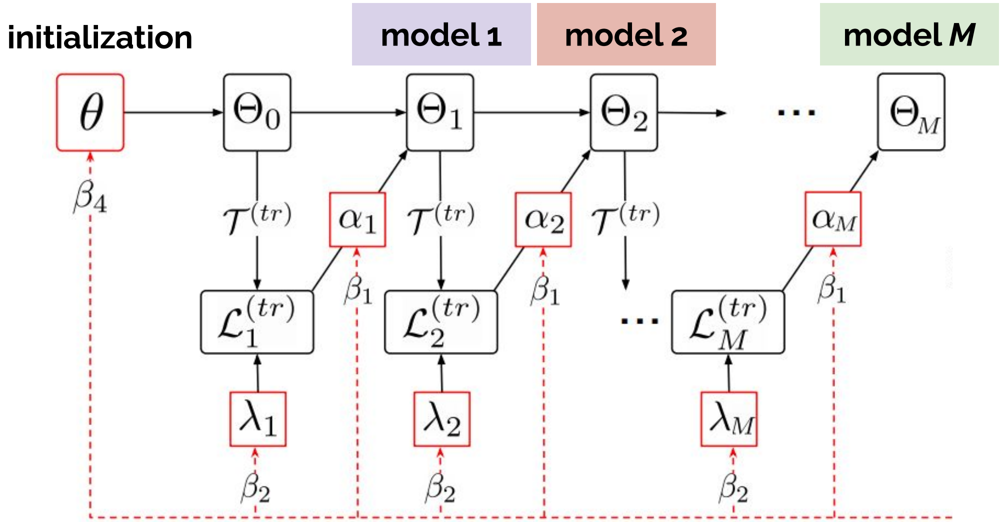
“Multiple models” == “models with different architectures, learning rates, data inputs, loss functions …”



Yaoyao Liu, et al. An Ensemble of Epoch-wise Empirical Bayes for Few-Shot Learning. ECCV 2020.

Learning to customize **models**

"Multiple models" == "models with different architectures, learning rates, data inputs, loss functions …"



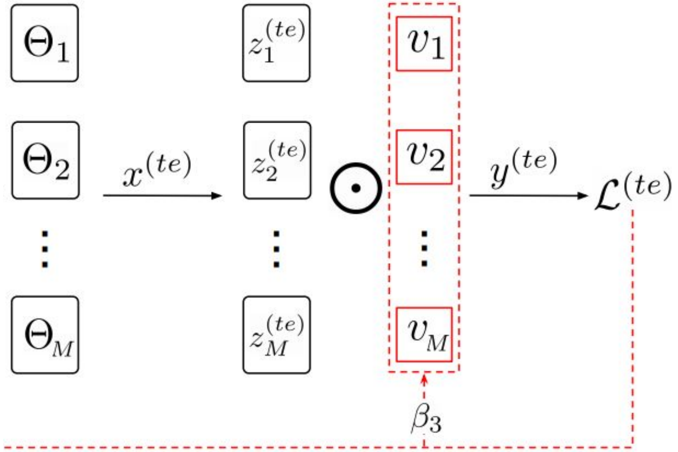
How to use?

Yaoyao Liu, et al. An Ensemble of Epoch-wise Empirical Bayes for Few-Shot Learning. ECCV 2020.

Learning to combine models

“Multiple models” == “models with different architectures, learning rates, data inputs, loss functions …”

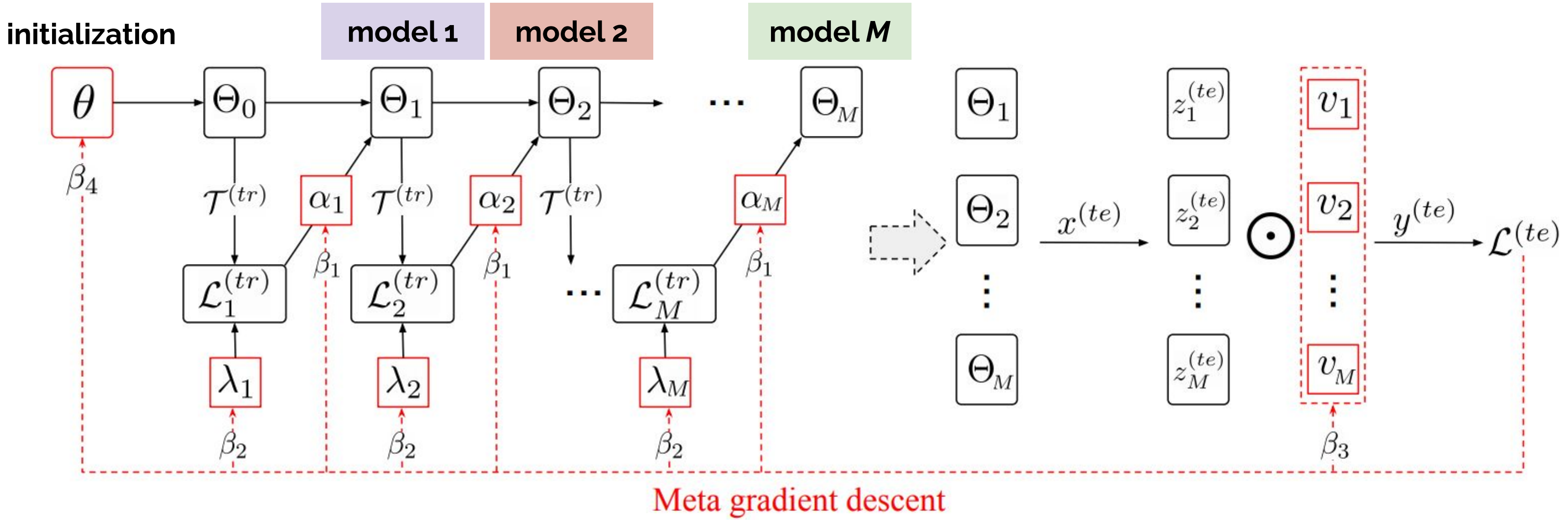
Weighted combination:



Yaoyao Liu, et al. An Ensemble of Epoch-wise Empirical Bayes for Few-Shot Learning. ECCV 2020.

Learning to combine **models**

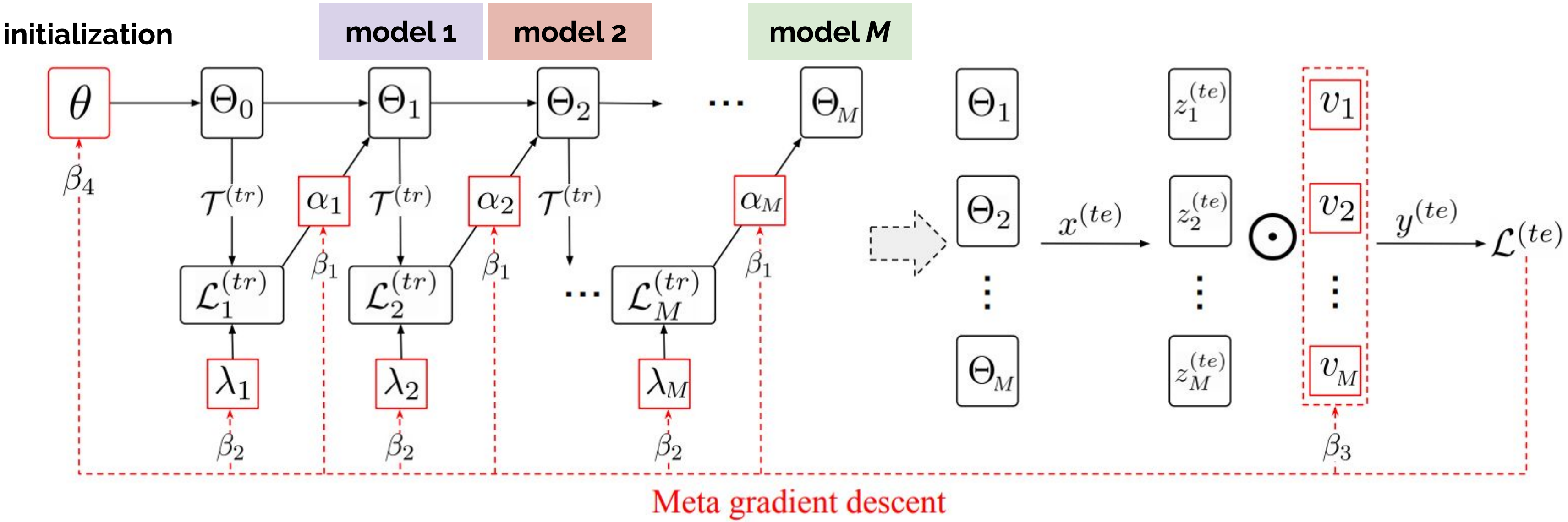
“Multiple models” == “models with different architectures, learning rates, data inputs, loss functions …”



Yaoyao Liu, et al. An Ensemble of Epoch-wise Empirical Bayes for Few-Shot Learning. ECCV 2020.

Learning to combine models

“Multiple models” == “models with different architectures, learning rates, data inputs, loss functions …”



Yaoyao Liu, et al. An Ensemble of Epoch-wise Empirical Bayes for Few-Shot Learning. ECCV 2020.

Learning to customize **&** combine

“Multiple models” == “models with different architectures, learning rates, data inputs, loss functions …”

Results on miniImageNet

Table. Classification accuracy (%). Higher is better.

	No.	Meta-learned			Accuracy	
		α	λ	v	1-shot	5-shot
	1			**E**	47.0 ± 1.8	62.0 ± 0.9
Baseline:	2			**S**	48.0 ± 1.8	62.4 ± 0.9
	3	✓		**S**	49.7 ± 1.8	64.4 ± 0.9
	4		✓	**S**	49.0 ± 1.8	63.4 ± 0.9
	5	✓	✓	**S**	49.0 ± 1.8	65.0 ± 0.9
	6			**L**	49.7 ± 1.8	65.4 ± 0.9
	7	✓		**L**	52.9 ± 1.8	65.6 ± 0.9
	8		✓	**L**	48.6 ± 1.8	64.7 ± 0.9
Ours:	LCC(Ours)	✓	✓	**L**	**54.0** ± 1.8	**65.8** ± 0.9
	“oracle” **v**			**O**	52.4 ± 1.8	64.7 ± 0.9

Baseline: using a **S**ingle model

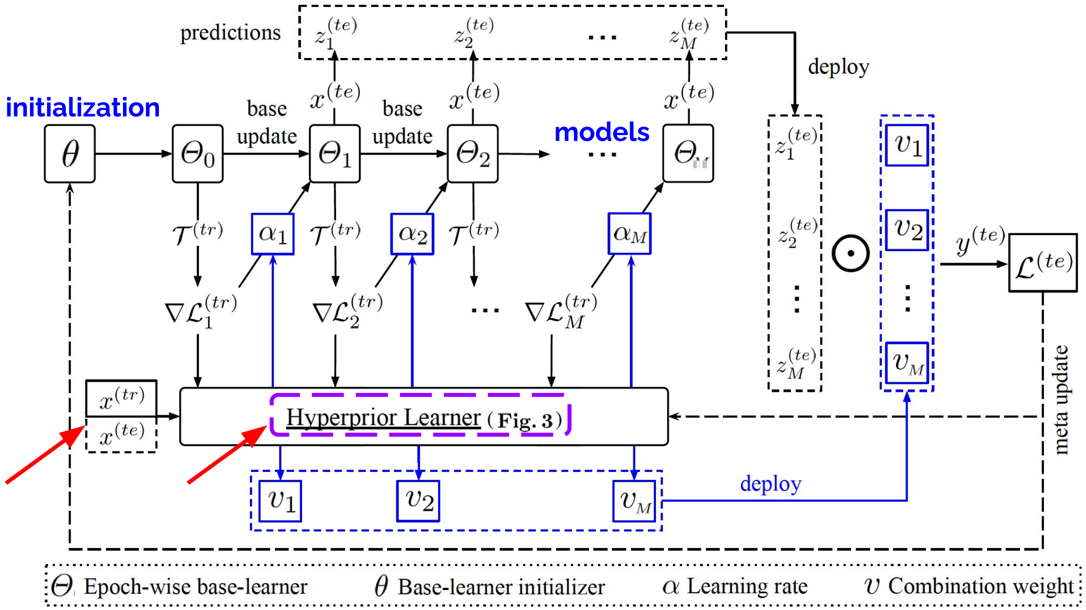
Improvements:

miniImageNet: **6.0%(1-shot)** and 3.4%(5-shot)

LCC greatly surpasses baseline, and performs better in the lower-shot setting.

Ours: Yaoyao Liu, et al. An Ensemble of Epoch-wise Empirical Bayes for Few-Shot Learning. ECCV 2020.
Baseline: Chelsea Finn, et al. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. ICML 2017.

An Ensemble of Epoch-wise Empirical Bayes



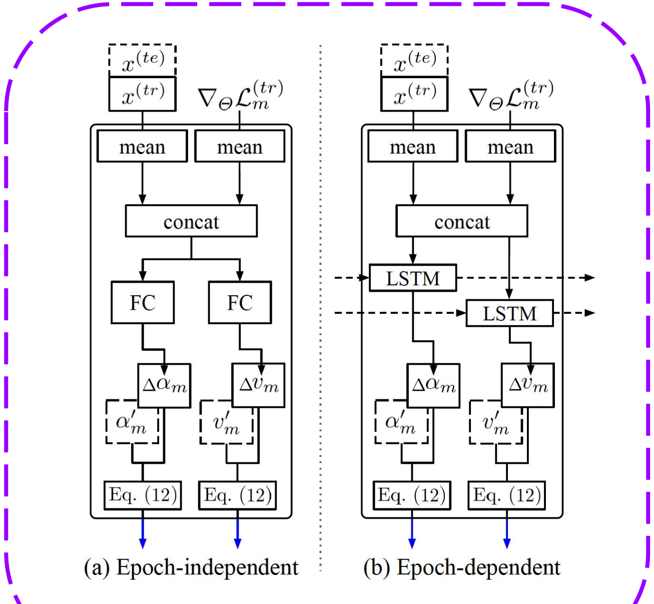


Fig. 3. Two options of hyperprior learner

$$\alpha_m = \lambda_1 \alpha'_m + (1 - \lambda_1)\Delta\alpha, \ v_m = \lambda_2 v'_m + (1 - \lambda_2)\Delta v \quad (12)$$

Yaoyao Liu, et al. An Ensemble of Epoch-wise Empirical Bayes for Few-Shot Learning. ECCV 2020.

An Ensemble of Epoch-wise Empirical Bayes

(**add-on** contributions)

Three popular few-shot learning benchmarks

No.	Method	Hyperprior	Learning	*mini*ImageNet 1-shot	*mini*ImageNet 5-shot	*tiered*ImageNet 1-shot	*tiered*ImageNet 5-shot	FC100 1-shot	FC100 5-shot
2019 SOTA: 1	MTL [70]	–	Ind.	63.4	80.1	69.1	84.2	43.7	60.1
2	MTL+E^{3}BM	FC	Ind.	64.3	80.9	69.8	84.6	44.8	60.5
3	MTL+E^{3}BM	FC	Tra.	64.7	80.7	69.7	84.9	44.7	**60.6**
4	MTL+E^{3}BM	LSTM	Ind.	64.3	81.0	70.0	**85.0**	45.0	60.4
5	MTL+E^{3}BM	LSTM	Tra.	64.5	**81.1**	70.2	**85.3**	45.1	**60.6**
2020 SOTA: 6	SIB [25]	–	Tra.	70.0	79.2	72.9	82.8	45.2	55.9
7	SIB+E^{3}BM	FC	Tra.	**71.3**	81.0	**75.2**	83.8	**45.8**	56.3
8	SIB+E^{3}BM	LSTM	Tra.	**71.4**	**81.2**	**75.6**	84.3	**46.0**	57.1

E3BM: Yaoyao Liu, et al. An Ensemble of Epoch-wise Empirical Bayes for Few-Shot Learning. ECCV 2020.
MTL: Qianru Sun, et al. Meta Transfer Learning for Few-Shot Learning. CVPR 2019.
SIB: Shell Xu Hu, et al. Empirical Bayes Transductive Meta-Learning with Synthetic Gradients. ICLR 2020.

My recent works - Learning to learn
for different capabilities in computer vision tasks

Learning to **transfer** “memory” --- “experiences from other large-scale tasks”

Learning to **extract** “data” --- “images potentially useful for future training”

Learning to **combine** “models” --- “trained network parameters”

Other capabilities?

My recent works - Learning to learn
for different capabilities in computer vision tasks

Learning to **transfer** "memory" --- "experiences from other large-scale tasks"

Learning to **extract** "data" --- "images potentially useful for future training"

Learning to **combine** "models" --- "trained network parameters"

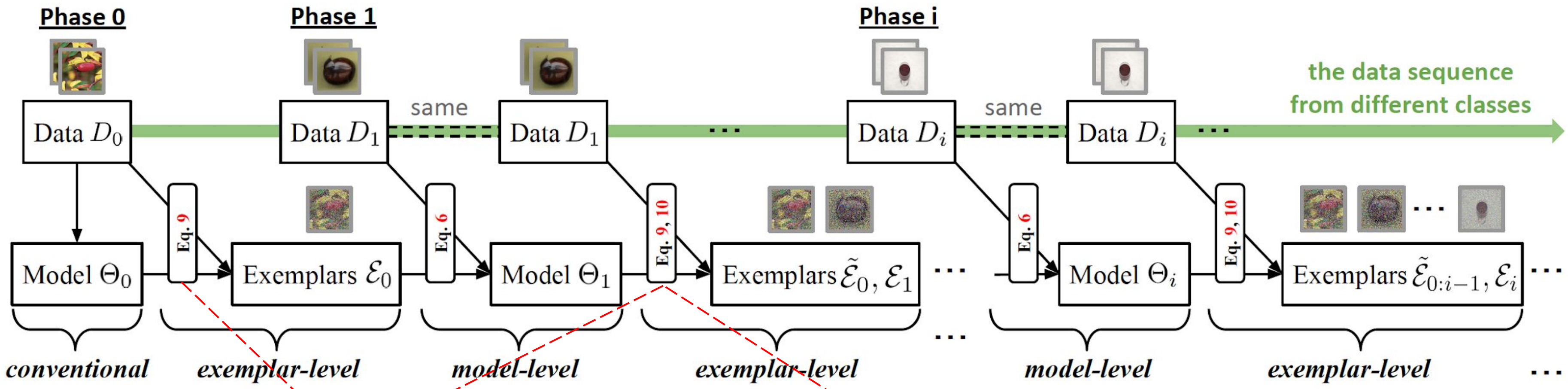
Other capabilities?

Learning to compress **the training data (for incremental learning)**

"Mnemonics Training: Multi-Class Incremental Learning without Forgetting" CVPR 2020 Oral

Learning to compress **data (Mnemonics)**

Idea: Before discarding the training data of the i-th phase, compress them to a small representative $\mathcal{E}_i$



$$\mathcal{E}_i \leftarrow \mathcal{E}_i - \beta_1 \nabla_{\mathcal{E}} \mathcal{L}_c\big(\Theta_i'(\mathcal{E}_i); D_i\big)$$

Eq. 9

Eq. 10

$$\mathcal{E}^{\mathbf{A}}_{0:i-1} \leftarrow \mathcal{E}^{\mathbf{A}}_{0:i-1} - \beta_2 \nabla_{\mathcal{E}^{\mathbf{A}}} \mathcal{L}_c\big(\Theta_i'(\mathcal{E}^{\mathbf{A}}_{0:i-1}); \mathcal{E}^{\mathbf{B}}_{0:i-1}\big)$$

$$\mathcal{E}^{\mathbf{B}}_{0:i-1} \leftarrow \mathcal{E}^{\mathbf{B}}_{0:i-1} - \beta_2 \nabla_{\mathcal{E}^{\mathbf{B}}} \mathcal{L}_c\big(\Theta_i'(\mathcal{E}^{\mathbf{B}}_{0:i-1}); \mathcal{E}^{\mathbf{A}}_{0:i-1}\big)$$

Yaoyao Liu, et al. Mnemonics Training: Multi-Class Incremental Learning without Forgetting. CVPR 2020 Oral Presentation.

My recent works - Learning to learn

for different capabilities in computer vision tasks

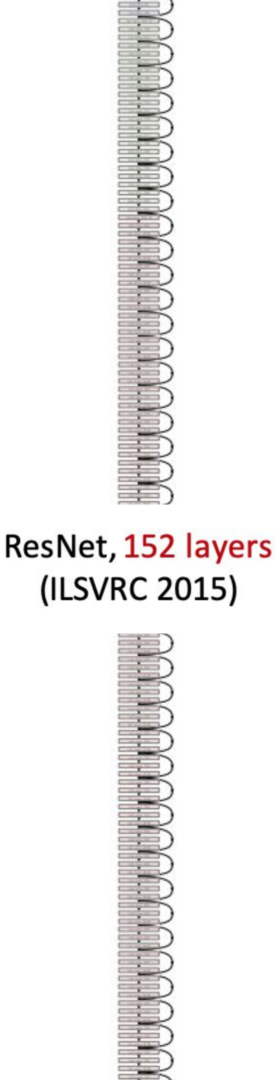
Learning to **transfer** “memory” --- “experiences from other large-scale tasks”

Learning to **extract** “data” --- “images potentially useful for future training”

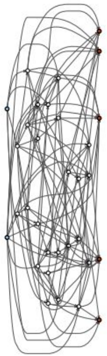
Learning to **combine** “models” --- “trained network parameters”

future work?

Architecture

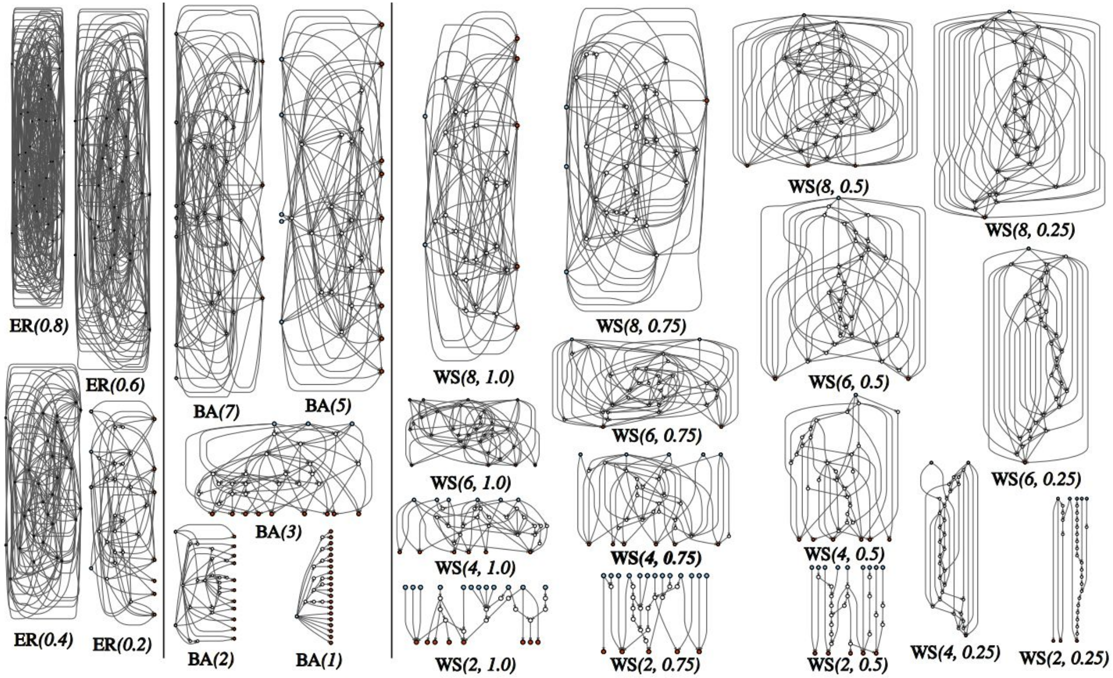


Learning to design **architectures**

Task → 

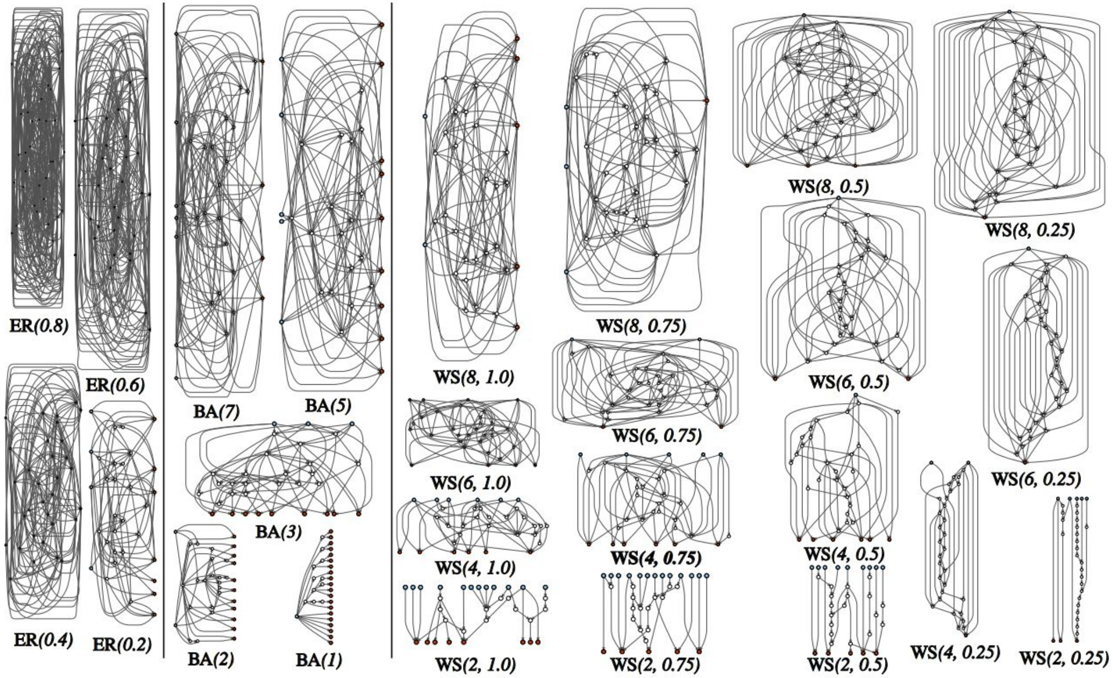
Saining Xie, Alexander Kirillov, Ross Girshick,Kaiming He. Exploring Randomly Wired Neural Networks for Image Recognition. ICCV 2019.

Learning to design **architectures**



Saining Xie, Alexander Kirillov, Ross Girshick,Kaiming He. Exploring Randomly Wired Neural Networks for Image Recognition. ICCV 2019.

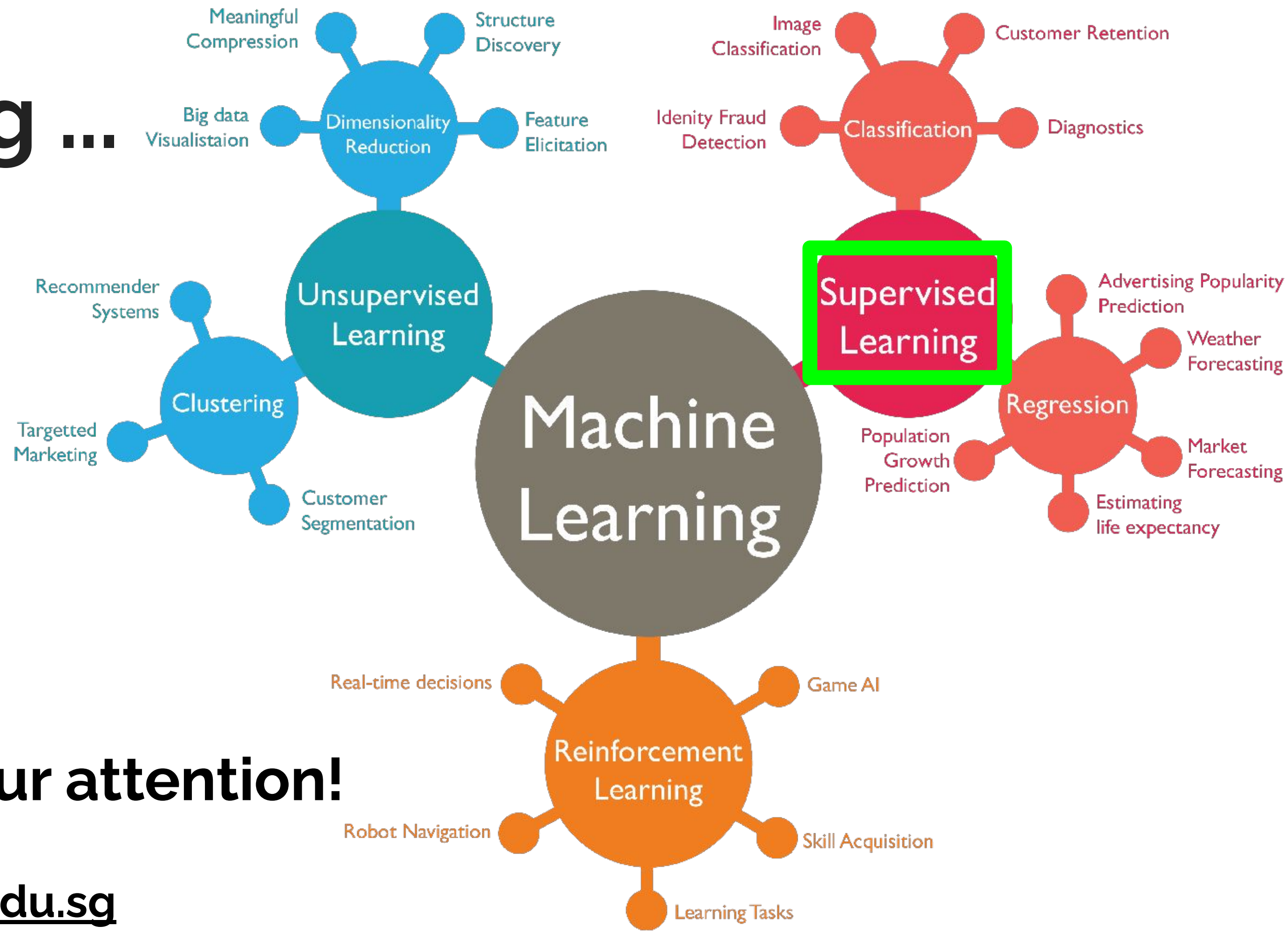
Learning to design **architectures**



Which one performs the best, given a few-shot task?

Saining Xie, Alexander Kirillov, Ross Girshick,Kaiming He. Exploring Randomly Wired Neural Networks for Image Recognition. ICCV 2019.

Learning ...



Thanks for your attention!

qianrusun.com
qianrusun@smu.edu.sg