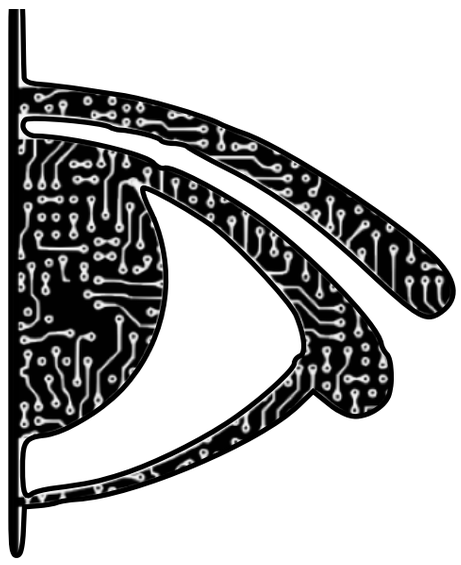
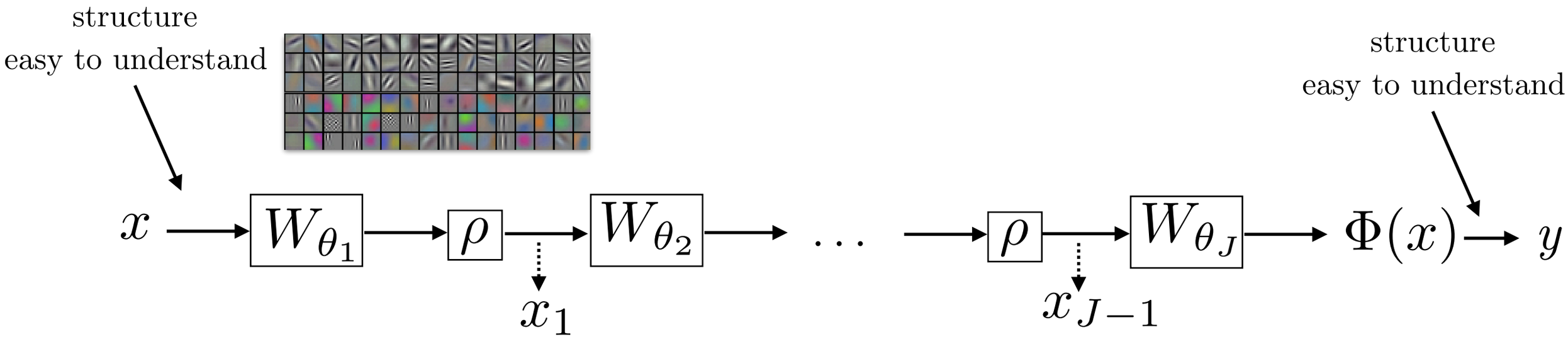
Greedy Layerwise Learning Can Scale to ImageNet

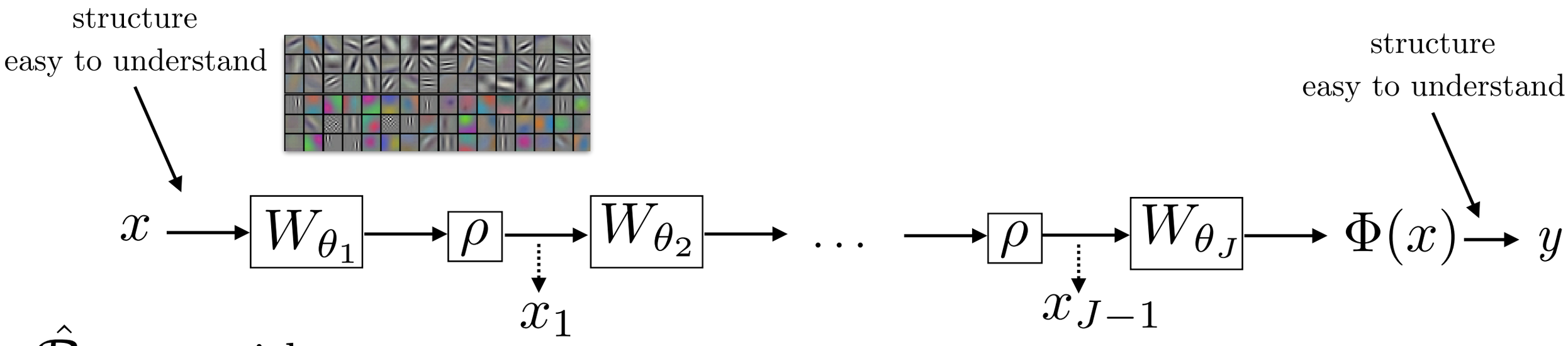
Eugene Belilovsky, Michael Eickenberg, Edouard Oyallon



Sequential Optimization



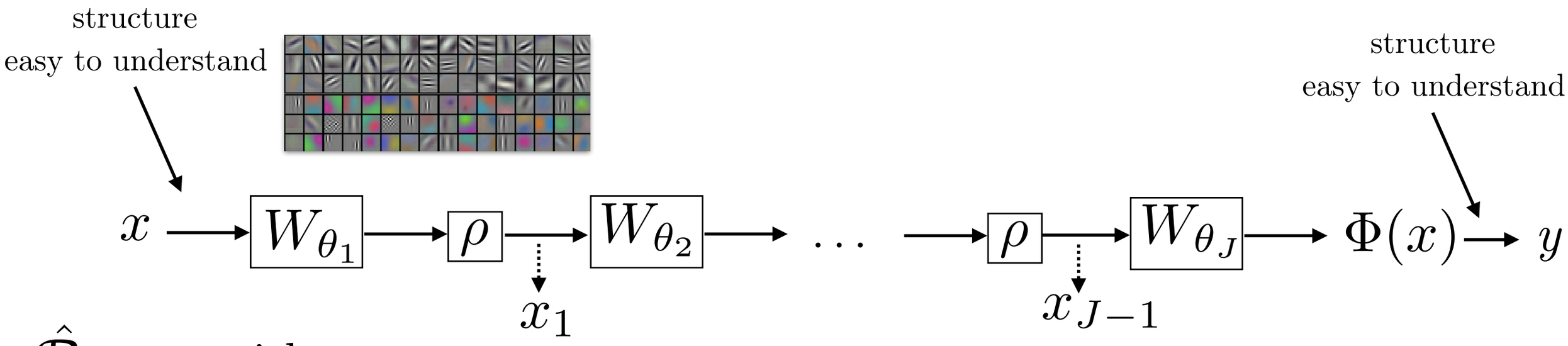
Sequential Optimization



$\hat{\mathcal{R}}$: some risk

$$\inf_{\{\theta_j\}} \hat{\mathcal{R}}(X_J, Y, \{\theta_j\}) \qquad \text{compared to} \qquad \forall j, \inf_{\theta_j} \hat{\mathcal{R}}(X_j, Y, \theta_j)$$

Sequential Optimization

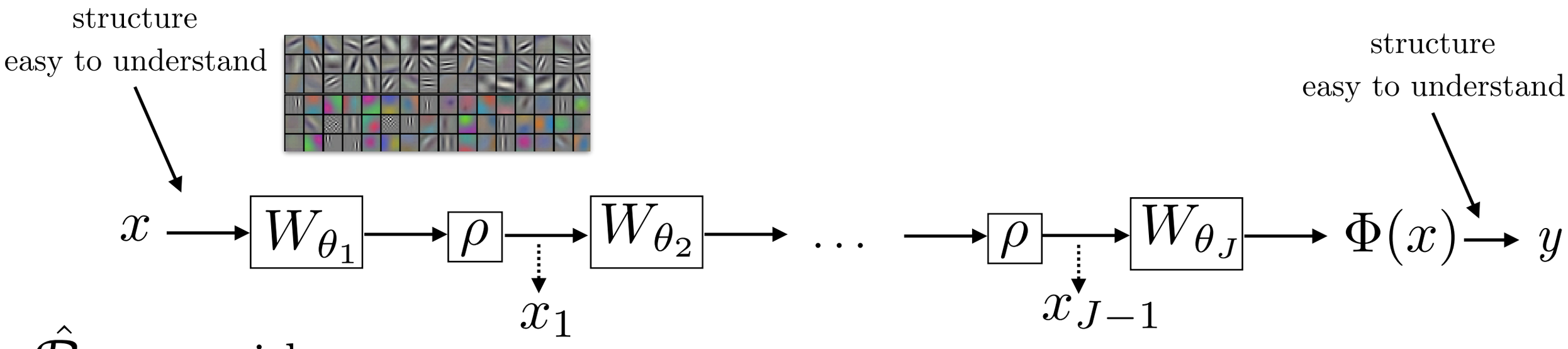


$\hat{\mathcal{R}}$: some risk

$$\inf_{\{\theta_j\}} \hat{\mathcal{R}}(X_J, Y, \{\theta_j\}) \qquad \text{compared to} \qquad \forall j, \inf_{\theta_j} \hat{\mathcal{R}}(X_j, Y, \theta_j)$$

- Goal: *removing* the joint end-to-end constraint.

Sequential Optimization



$\hat{\mathcal{R}}$: some risk

$$\inf_{\{\theta_j\}} \hat{\mathcal{R}}(X_J, Y, \{\theta_j\}) \qquad \text{compared to} \qquad \forall j, \inf_{\theta_j} \hat{\mathcal{R}}(X_j, Y, \theta_j)$$

- Goal: *removing* the joint end-to-end constraint.
- Can we specify *explicitly* the objective of each individual layers? (beyond the "black-box" optimization)

Motivation

Motivation

A. shallow (1-hidden layer) NNs: approximation or optimisation guarantees are widely studied.

Ref.: Approximation and Estimation Bounds for Artificial Neural Networks, Barron 1994
Spurious Valleys in Two-layer Neural Network Optimisation Landscapes, Venturi et al.
Breaking the curse of dimensionality with convex neural networks, F Bach
Gradient Descent Learns One-hidden-layer CNN: Don't be afraid of Spurious Local Minima; Du et al, 2018

Motivation

A. <u>shallow (1-hidden layer) NNs:</u> approximation or optimisation guarantees are widely studied.

Ref.: Approximation and Estimation Bounds for Artificial Neural Networks, Barron 1994
Spurious Valleys in Two-layer Neural Network Optimisation Landscapes, Venturi et al.
Breaking the curse of dimensionality with convex neural networks, F Bach
Gradient Descent Learns One-hidden-layer CNN: Don't be afraid of Spurious Local Minima; Du et al, 2018

B. <u>Inner organization</u>: interaction between layers is not well understood.

Ref.: On the information bottleneck theory of deep learning, Saxe et al

Motivation

A. <u>shallow (1-hidden layer) NNs:</u> approximation or optimisation guarantees are widely studied.

Ref.: Approximation and Estimation Bounds for Artificial Neural Networks, Barron 1994
Spurious Valleys in Two-layer Neural Network Optimisation Landscapes, Venturi et al.
Breaking the curse of dimensionality with convex neural networks, F Bach
Gradient Descent Learns One-hidden-layer CNN: Don't be afraid of Spurious Local Minima; Du et al, 2018

B. <u>Inner organization</u>: interaction between layers is not well understood.

Ref.: On the information bottleneck theory of deep learning, Saxe et al

Study of deep CNNs for (A) or (B) are limited to < 3 layers...

Ref.: Learning and Generalization in Overparametrized Neural Networks, Going beyond to Layer; Allen-Zhu et all, 2018
The power of Depth for Feedforward Neural Networks, Ronen Eldan and Ohad Shamir

Motivation

A. <u>shallow (1-hidden layer) NNs:</u> approximation or optimisation guarantees are widely studied.

Ref.: Approximation and Estimation Bounds for Artificial Neural Networks, Barron 1994
Spurious Valleys in Two-layer Neural Network Optimisation Landscapes, Venturi et al.
Breaking the curse of dimensionality with convex neural networks, F Bach
Gradient Descent Learns One-hidden-layer CNN: Don't be afraid of Spurious Local Minima; Du et al, 2018

B. <u>Inner organization</u>: interaction between layers is not well understood.

Ref.: On the information bottleneck theory of deep learning, Saxe et al

Study of deep CNNs for (A) or (B) are limited to < 3 layers...

Ref.: Learning and Generalization in Overparametrized Neural Networks, Going beyond to Layer; Allen-Zhu et all, 2018
The power of Depth for Feedforward Neural Networks, Ronen Eldan and Ohad Shamir

- Can (A) help to reveal the **structure** of (B)?

Motivation

A. <u>shallow (1-hidden layer) NNs:</u> approximation or optimisation guarantees are widely studied.

Ref.: Approximation and Estimation Bounds for Artificial Neural Networks, Barron 1994
Spurious Valleys in Two-layer Neural Network Optimisation Landscapes, Venturi et al.
Breaking the curse of dimensionality with convex neural networks, F Bach
Gradient Descent Learns One-hidden-layer CNN: Don't be afraid of Spurious Local Minima; Du et al, 2018

B. <u>Inner organization</u>: interaction between layers is not well understood.

Ref.: On the information bottleneck theory of deep learning, Saxe et al

Study of deep CNNs for (A) or (B) are limited to < 3 layers...

Ref.: Learning and Generalization in Overparametrized Neural Networks, Going beyond to Layer; Allen-Zhu et all, 2018
The power of Depth for Feedforward Neural Networks, Ronen Eldan and Ohad Shamir

- Can (A) help to reveal the **structure** of (B)?

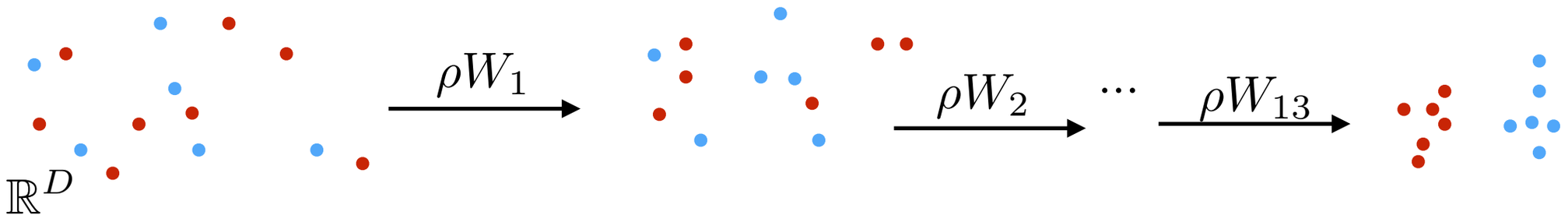
- Have some of the above references a chance to **scale numerically**?

(e.g., can theory work in practice?)

Empirical observation: Progressive separability

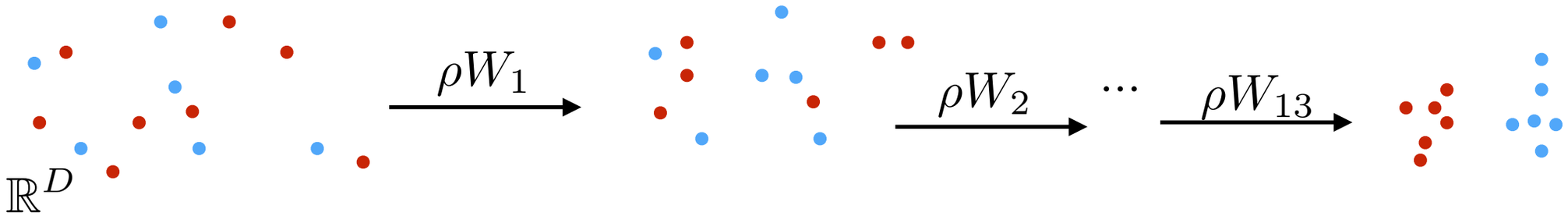
Empirical observation: Progressive separability

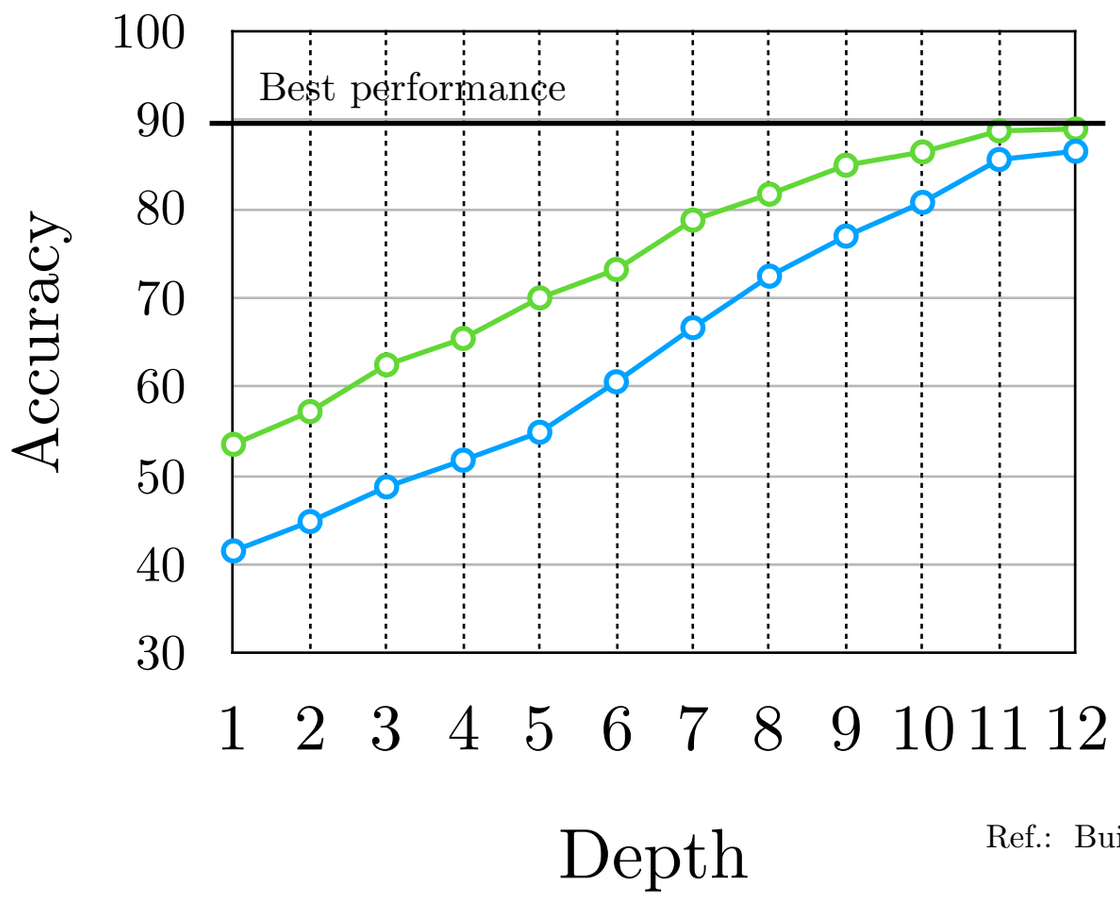
- Typical CNN exhibits a progressive contraction & separation, w.r.t. the depth:

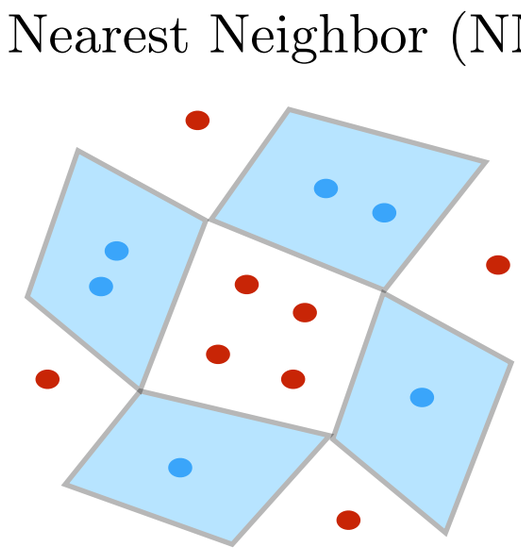


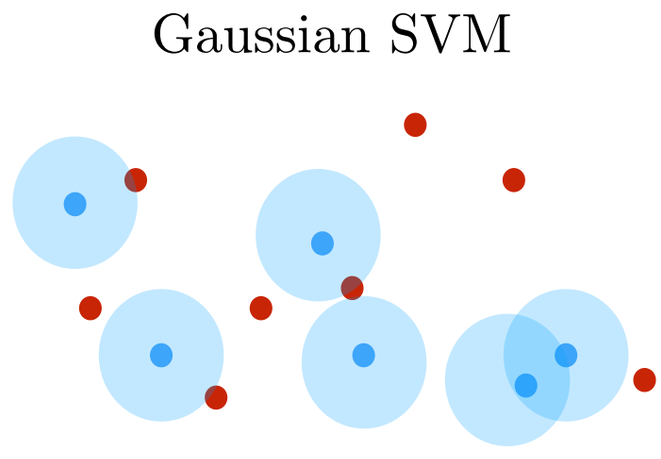
Empirical observation: Progressive separability

- Typical CNN exhibits a progressive contraction & separation, w.r.t. the depth:







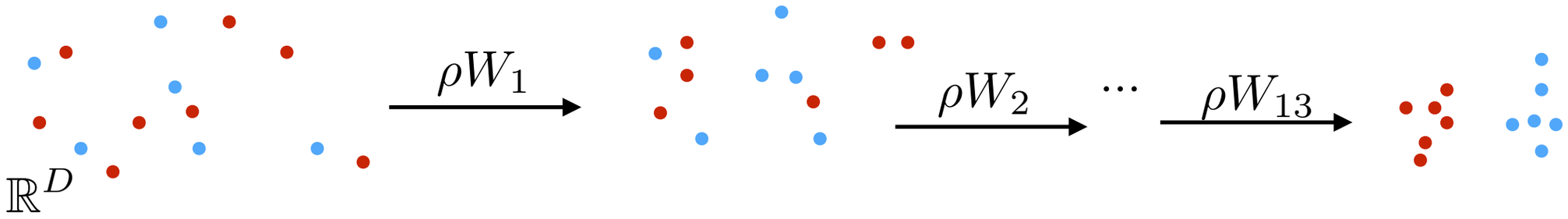


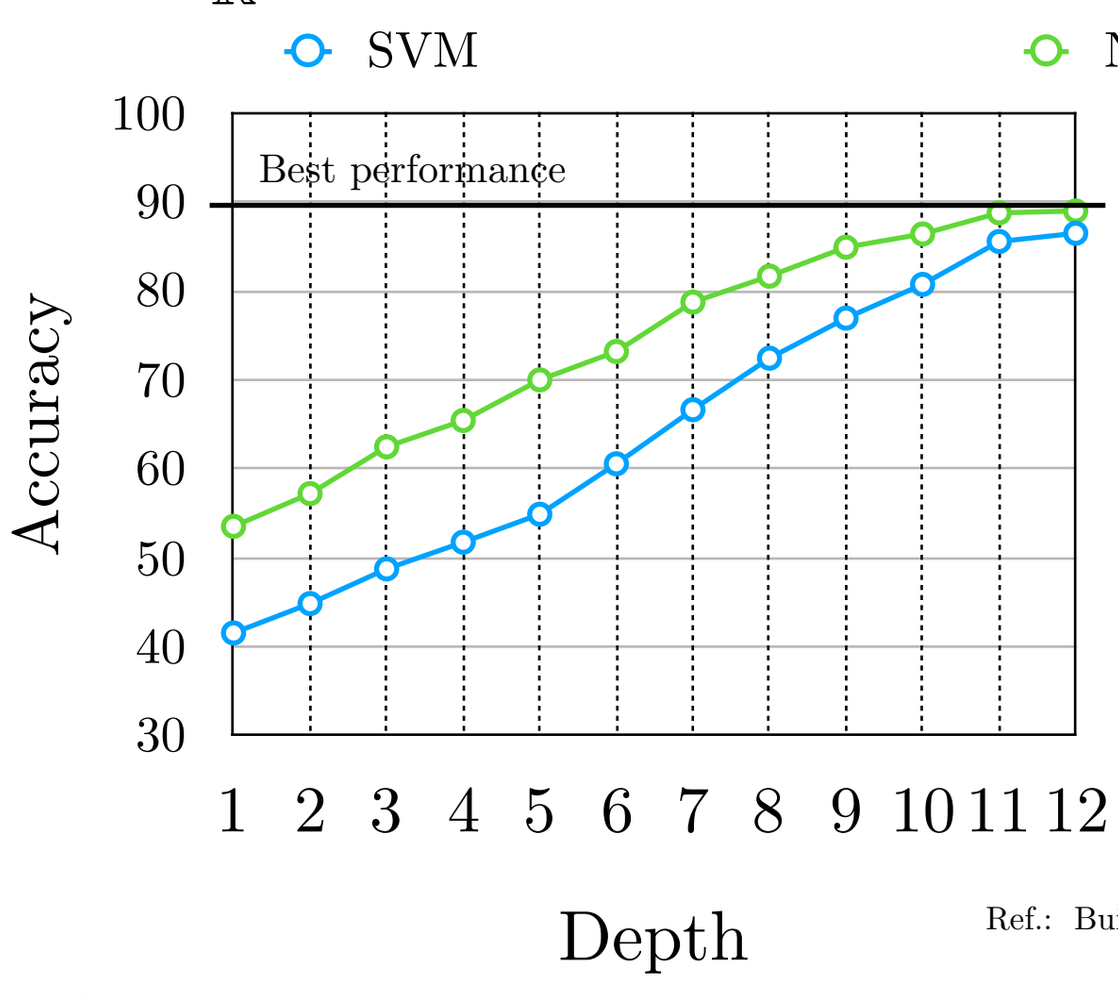
Localised classifiers

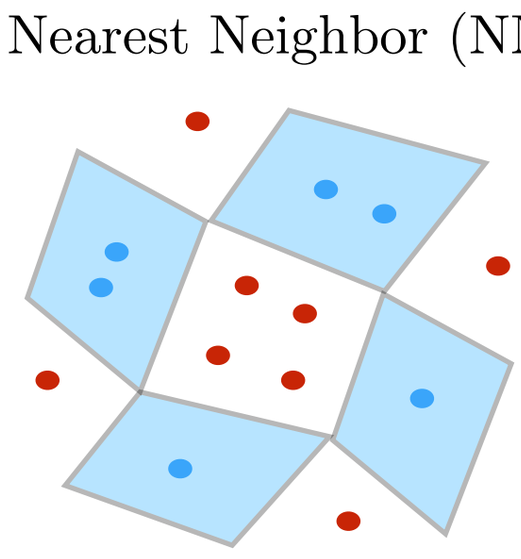
Ref.: Building a Regular Decision Boundary with Deep Networks, EO

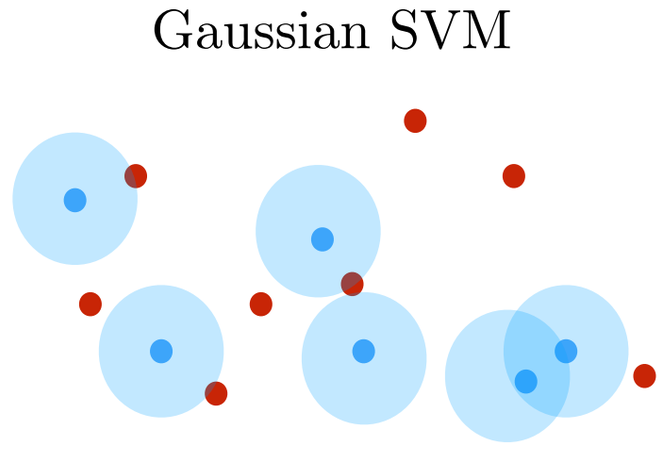
Empirical observation: Progressive separability

- Typical CNN exhibits a progressive contraction & separation, w.r.t. the depth:









Localised classifiers

Ref.: Building a Regular Decision Boundary with Deep Networks, EO

- Can we reciprocally impose this property layerwise?

Explicit greedy layer objective

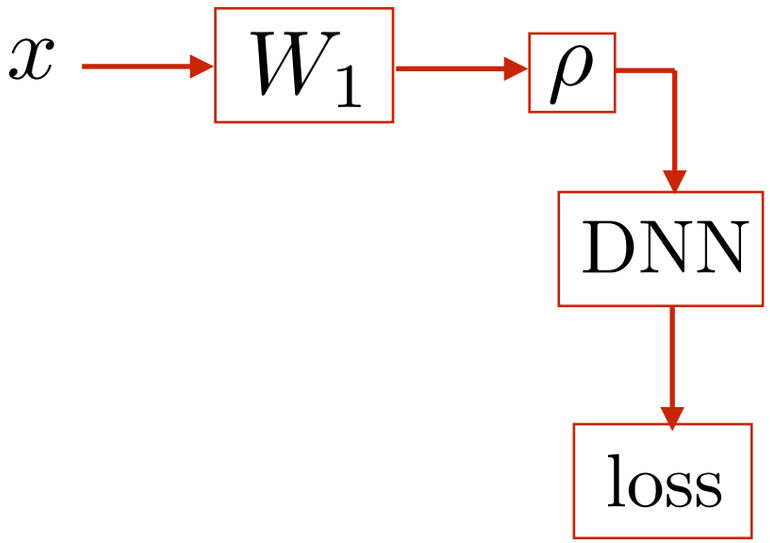
Simply train the CNN layer-per-layer via back-prop...

Frozen layers
Trainable layers

$$\boxed{\text{Let } k = \text{depth(DNN)} + 1}$$

Explicit greedy layer objective

Simply train the CNN layer-per-layer via back-prop...

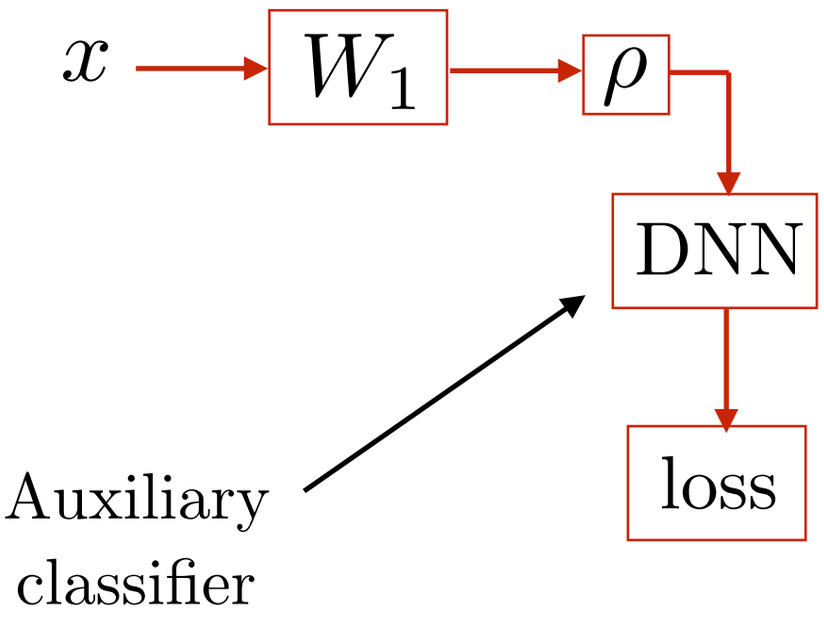


Frozen layers
Trainable layers

$$\boxed{\text{Let } k = \text{depth(DNN)} + 1}$$

Explicit greedy layer objective

Simply train the CNN layer-per-layer via back-prop...

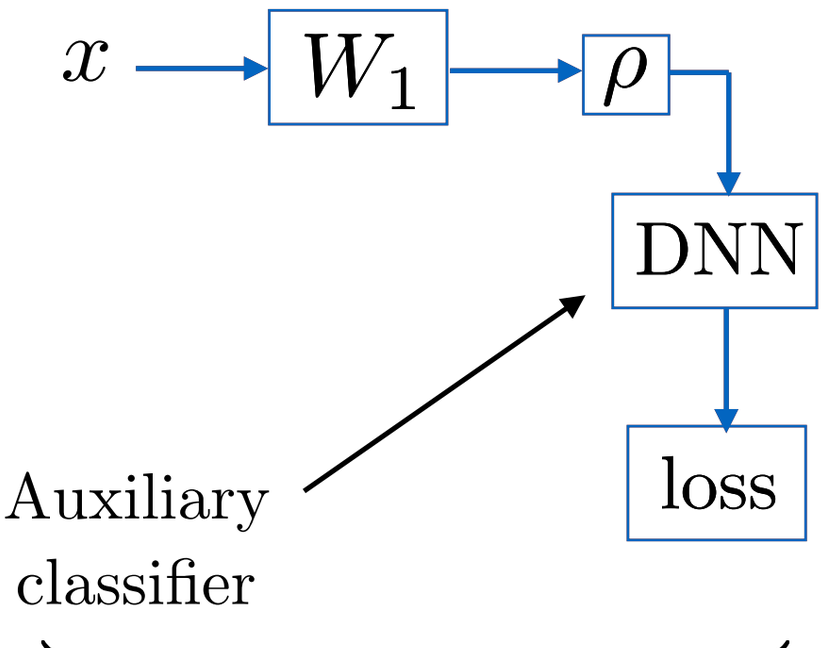


Frozen layers
Trainable layers

Let $k = \text{depth(DNN)} + 1$

Explicit greedy layer objective

Simply train the CNN layer-per-layer via back-prop...

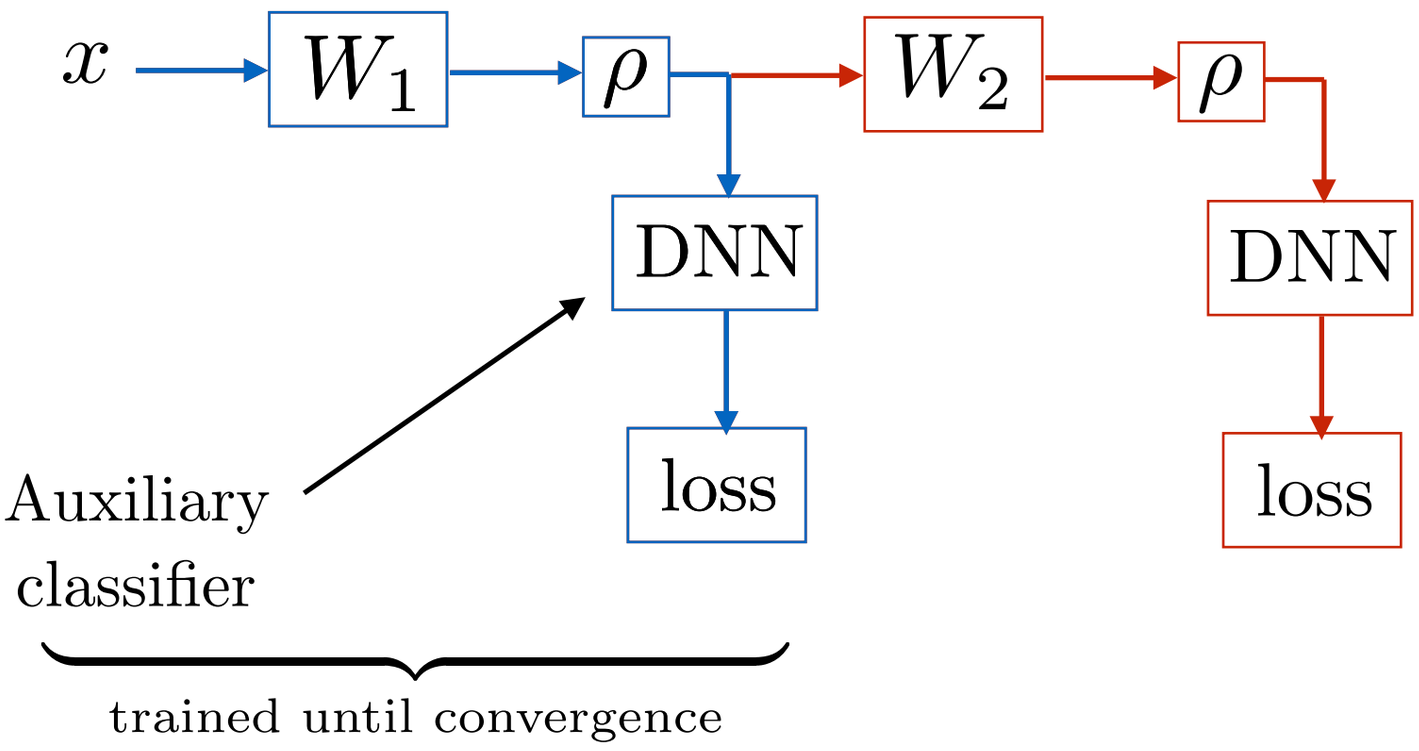


Frozen layers
Trainable layers

Let $k = \text{depth(DNN)} + 1$

Explicit greedy layer objective

Simply train the CNN layer-per-layer via back-prop...

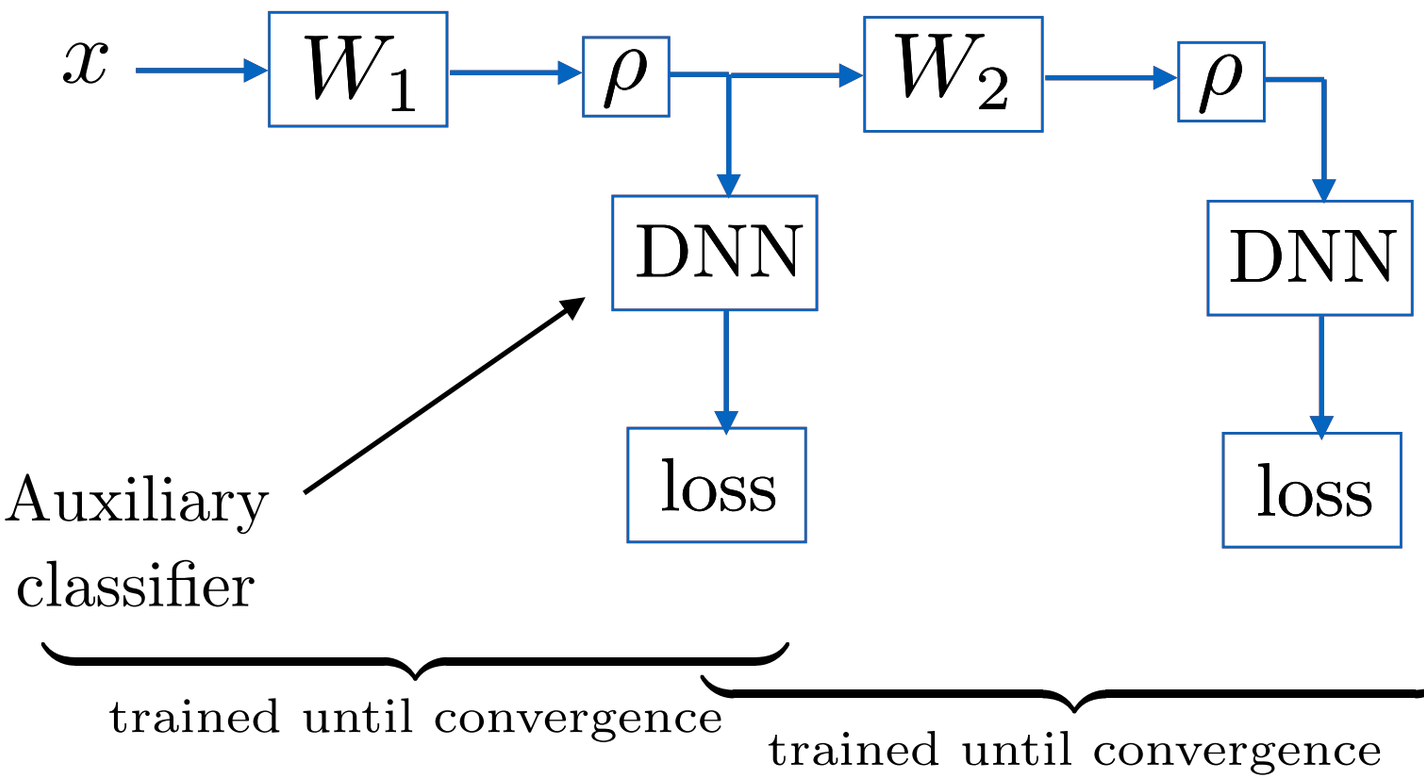


Frozen layers
Trainable layers

Let $k = \text{depth(DNN)} + 1$

Explicit greedy layer objective

Simply train the CNN layer-per-layer via back-prop...

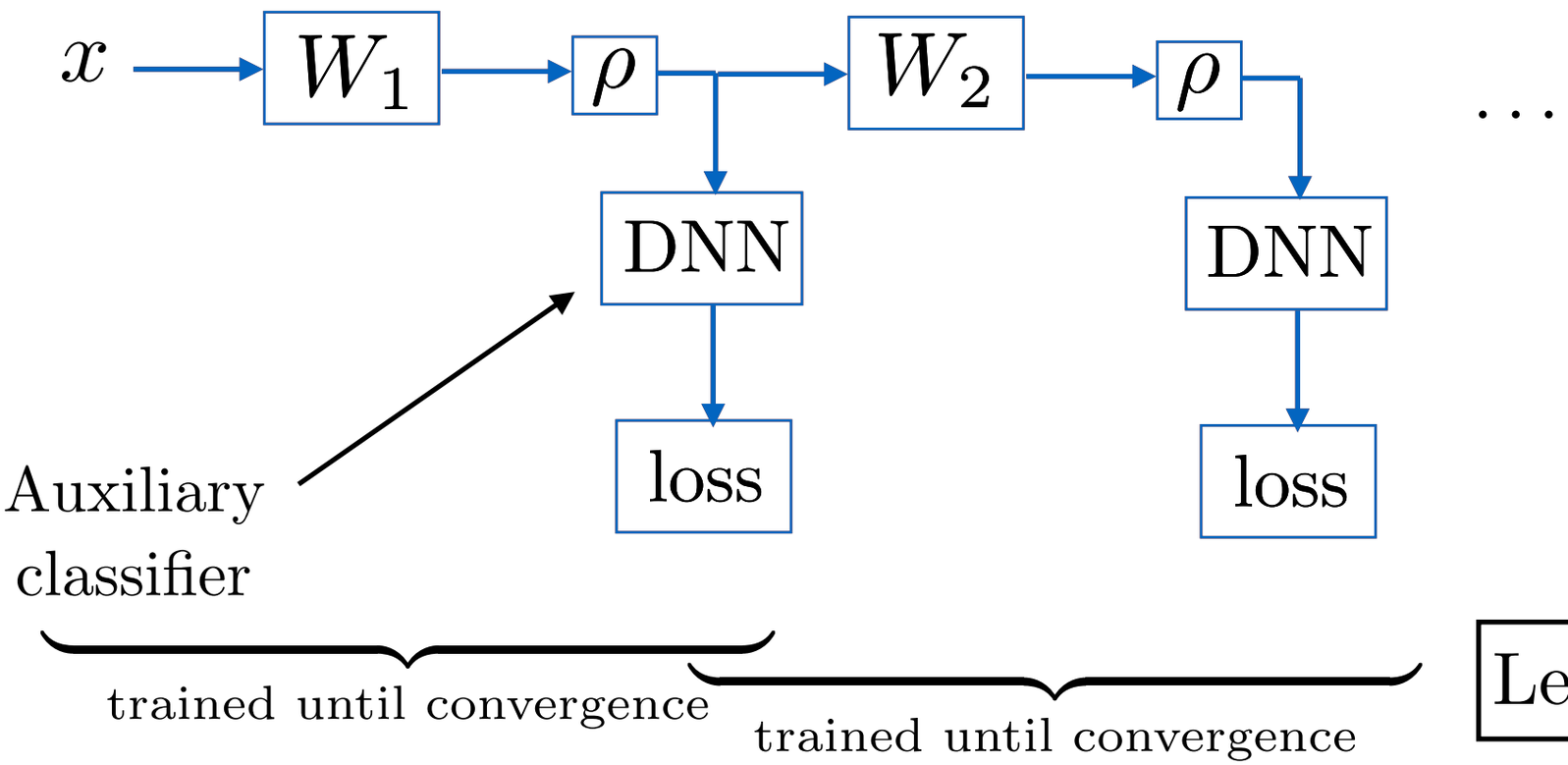


Frozen layers
Trainable layers

Let $k = \text{depth(DNN)} + 1$

Explicit greedy layer objective

Simply train the CNN layer-per-layer via back-prop...

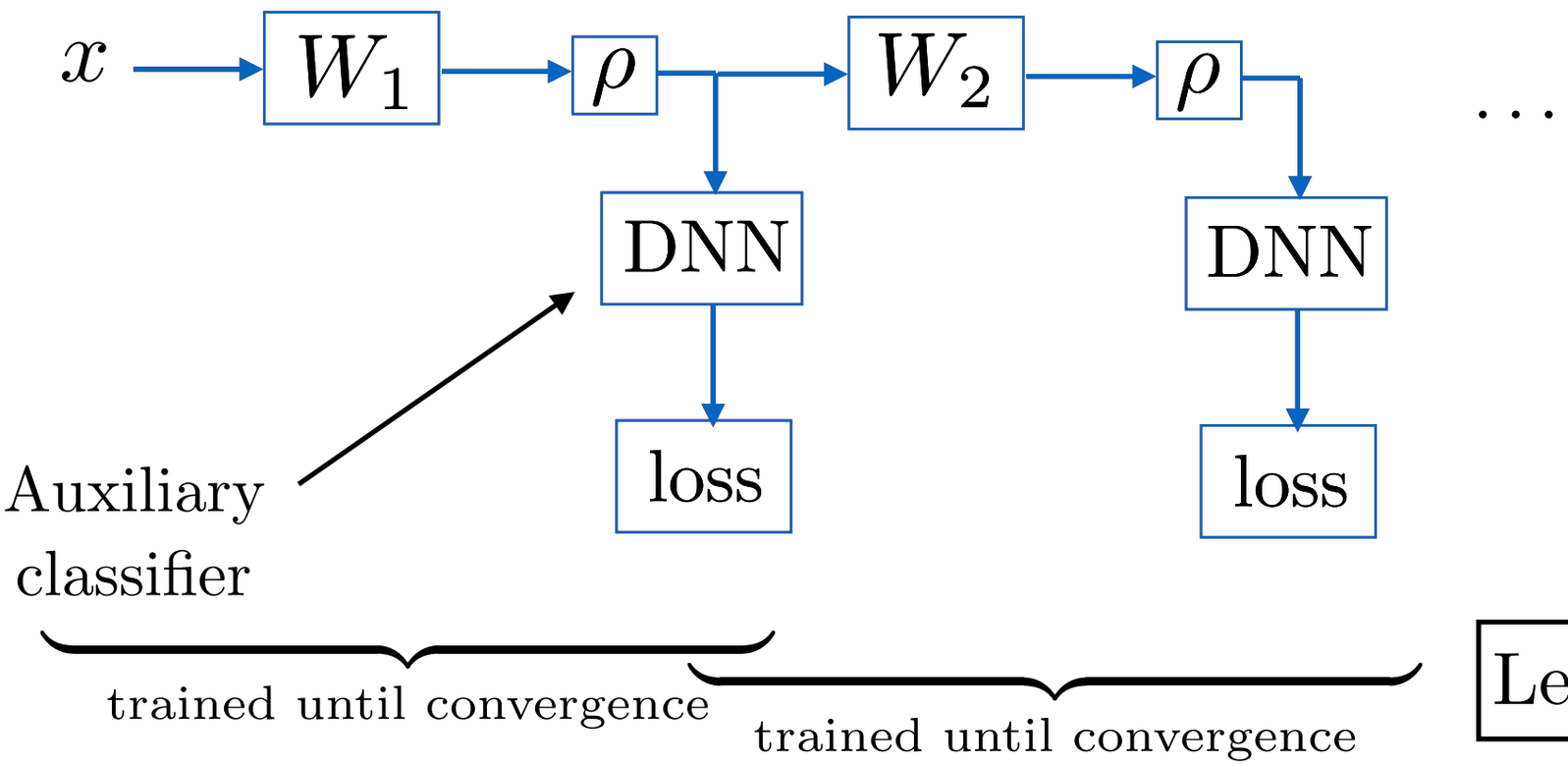


Frozen layers
Trainable layers

Let $k = \text{depth(DNN)} + 1$

Explicit greedy layer objective

Simply train the CNN layer-per-layer via back-prop...



Frozen layers
Trainable layers

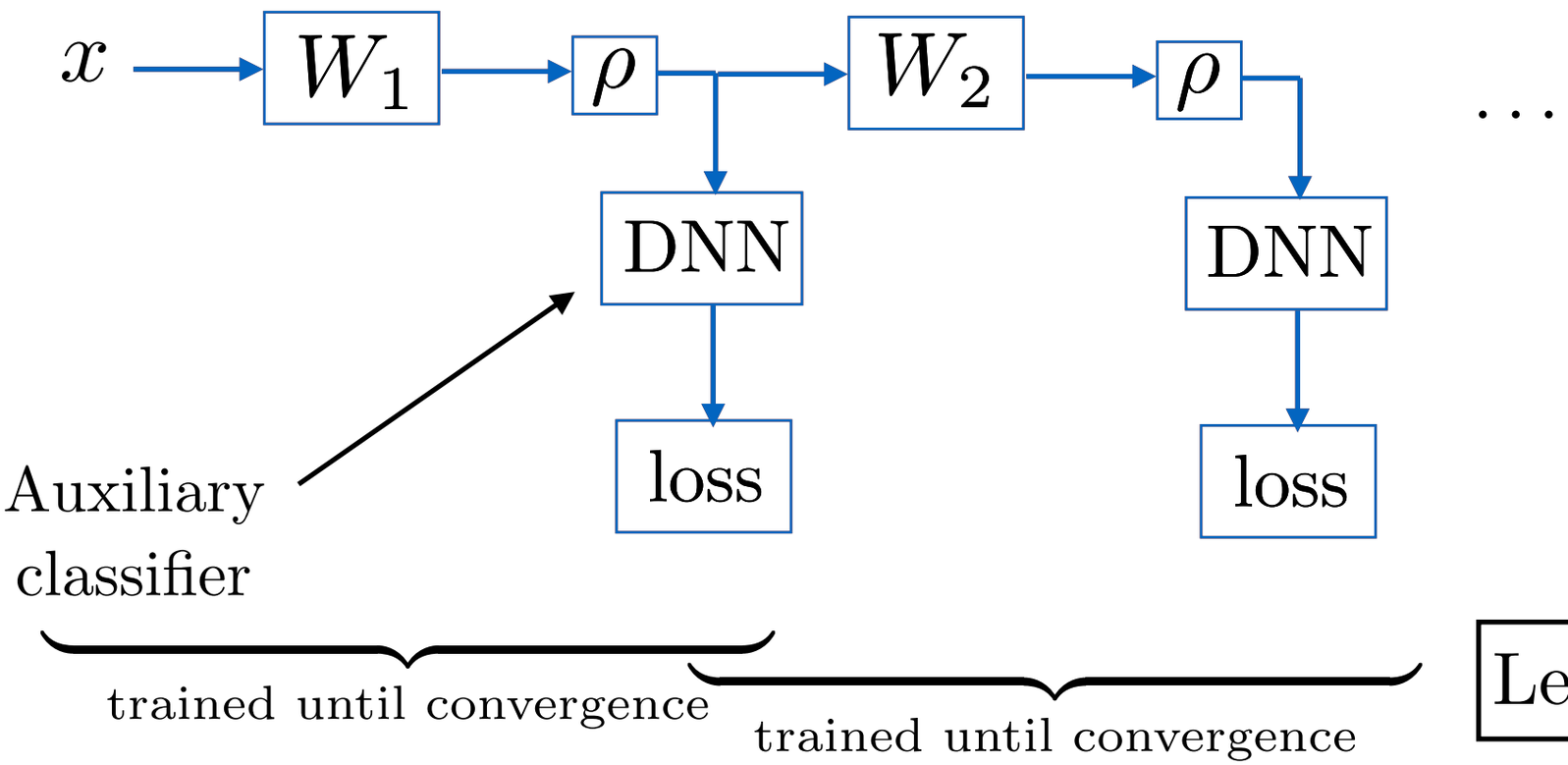
Let $k = \text{depth(DNN)} + 1$

- A very simple idea in the literature for a while...

Ref.: Learning Deep ResNet Blocks Sequentially using Boosting Theory, Huang et al, 2018
Greedy layer-wise training of Deep Networks, Bengio et al, 2006
Cybernetic predicting devices. Ivakhnenko et al 1965

Explicit greedy layer objective

Simply train the CNN layer-per-layer via back-prop...



Frozen layers
Trainable layers

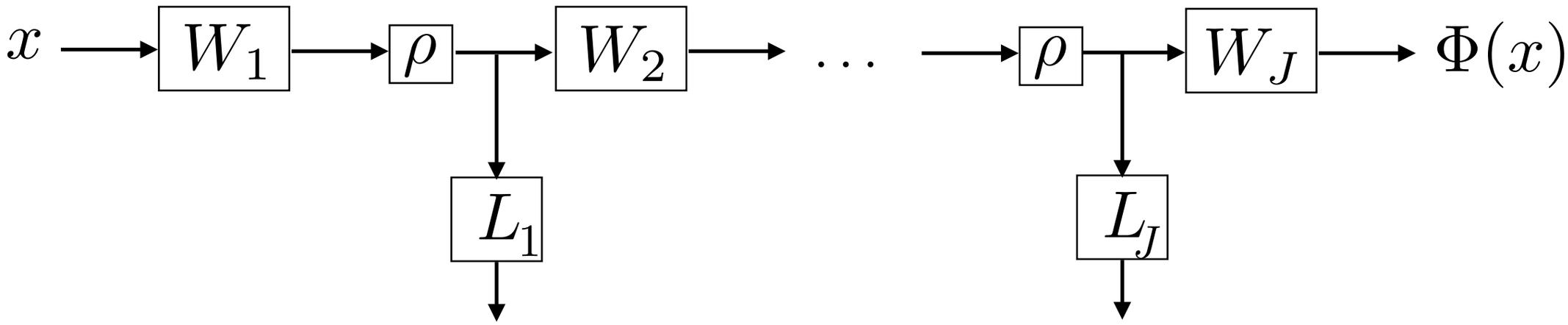
Let $k = \text{depth(DNN)} + 1$

- A very simple idea in the literature for a while...

Ref.: Learning Deep ResNet Blocks Sequentially using Boosting Theory, Huang et al, 2018
Greedy layer-wise training of Deep Networks, Bengio et al, 2006
Cybernetic predicting devices. Ivakhnenko et al 1965

- ... but it was *known to not scale* to ImageNet!

1 hidden layer $(k = 1)$



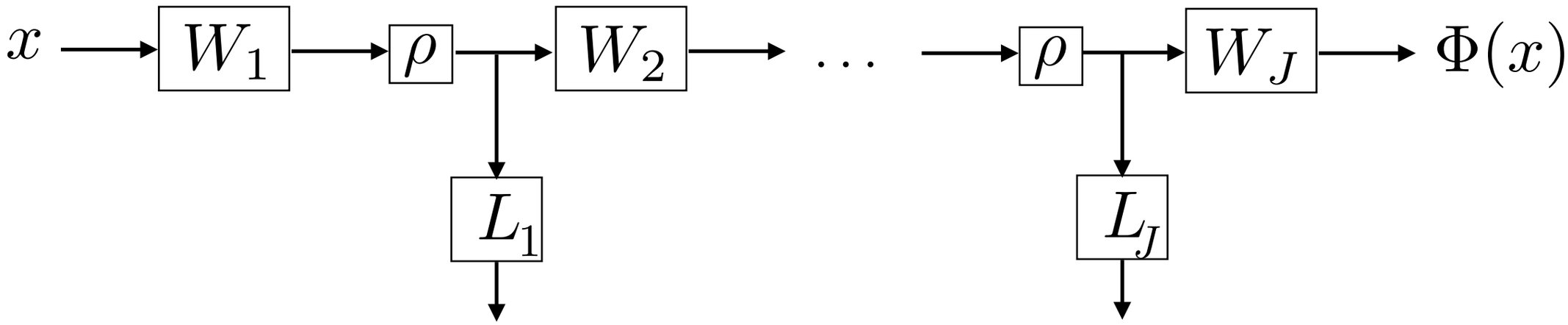
Arch/Perf. on ImageNet	Top 5
Layerwise	**79.7**
AlexNet	79.1
Handcrafted	74.2
Feedback Align(Bio plausible)	16.7

$$\Phi(x) = L\underline{\rho W}x$$

Simple to analyse

Explicit goal:
linear separability

1 hidden layer $(k = 1)$



Arch/Perf. on ImageNet	Top 5
Layerwise	**79.7**
AlexNet	79.1
Handcrafted	74.2
Feedback Align(Bio plausible)	16.7

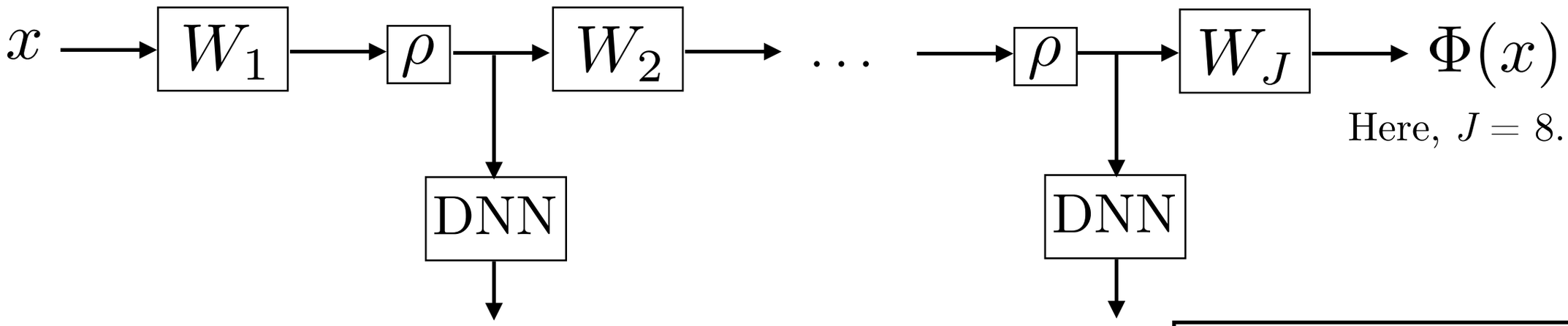
We show that linear separability, as a layer wise objective... scales!

$$\Phi(x) = L\underline{\rho W}x$$

Simple to analyse

Explicit goal:
linear separability

$$k > 1$$



$$k = \text{depth(DNN)} + 1$$

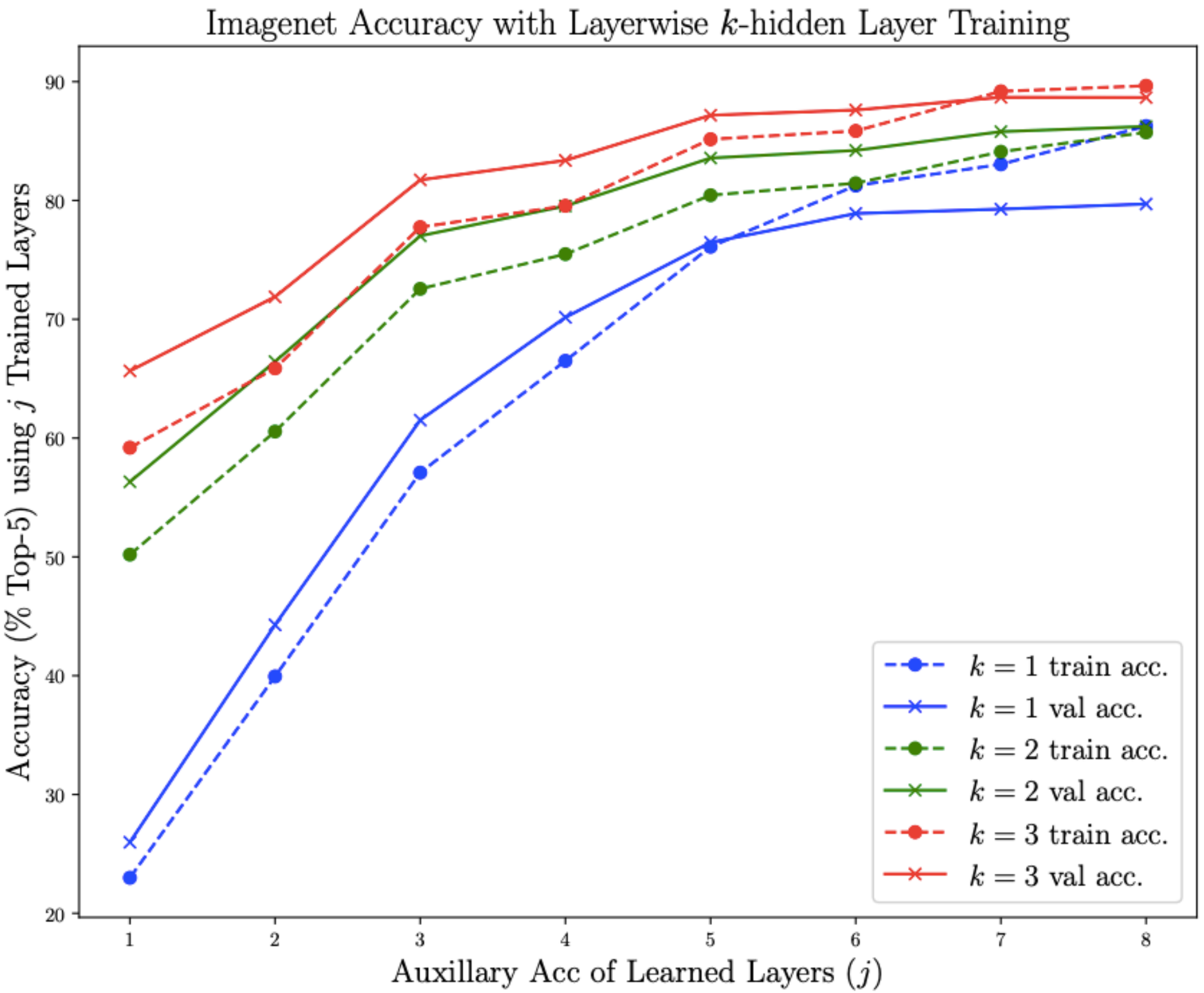
- We apply the same technique. Performance increases?!

Arch/Perf. on ImageNet	Top 5
Layerwise, $k = 2$	**86.3**
Layerwise, $k = 3$	**88.7**
State-of-the-art (152 layers)	94.1

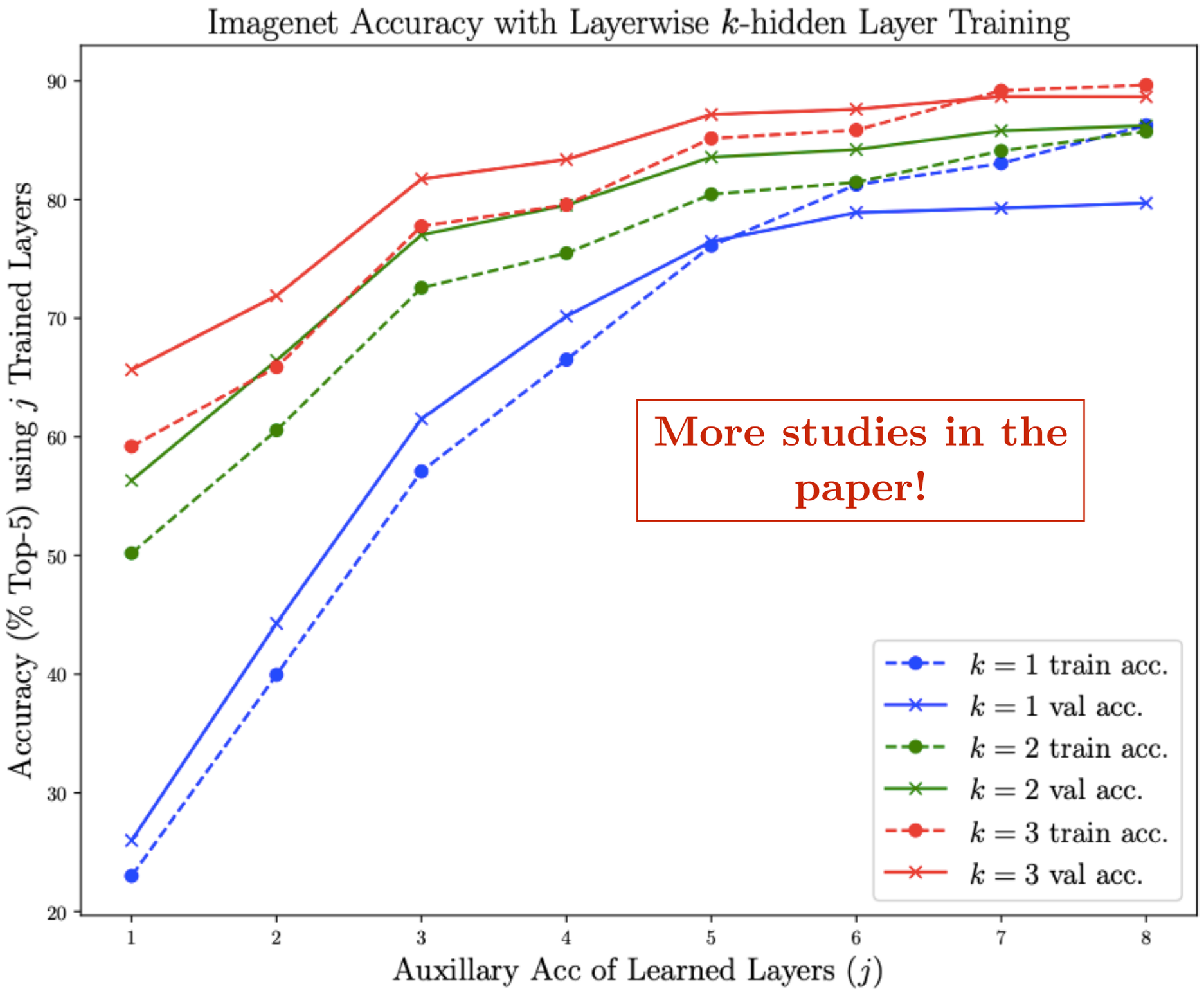
Seems to indicate that *some* depth is a key ingredient

Arch/Perf. of VGG-11 on ImageNet	Top 5
Layerwise, $k = 3$	**88.0**
End-to-end	**88.0**

Per Layer Performance



Per Layer Performance



Conclusion

Conclusion

- We demonstrate that greedy learning *scales* to ImageNet.

Conclusion

- We demonstrate that greedy learning *scales* to ImageNet.
  - Intermediary layers objective are *better specified*.

Conclusion

- We demonstrate that greedy learning *scales* to ImageNet.
  - Intermediary layers objective are *better specified*.
  - A well-understood 1-hidden layer optimisation would lead to a *numerically* successful procedure for deeper NNs.

Conclusion

- We demonstrate that greedy learning *scales* to ImageNet.
  - Intermediary layers objective are *better specified*.
  - A well-understood 1-hidden layer optimisation would lead to a *numerically* successful procedure for deeper NNs.
  - Opens many interesting questions and possibilities!
- **On going work**: decoupling layers such that they are trained in parallel.

Ref.: Decoupled Greedy Learning of CNNs, EB et al.