Phenomena in (Deep) Learning

COMP6258

Deep Learning Theory Research

with practical implications

Generalisation

Deep Learning Theory Research

with practical implications

DNNs and Overparametrisation

Is overparametrisation good or bad? Why?

DNNs and Overparametrisation

- ImageNet contains 1.2M training examples of size 224x224 split between 1000 classes

* UNDERSTANDING DEEP LEARNING REQUIRES RETHINKING GENERALIZATION, Zhang et. al (2016)

DNNs and Overparametrisation

- ImageNet contains 1.2M training examples of size 224x224 split between 1000 classes
- Suppose we randomise the labels of all the 1.2M training samples (each receives one of the 1000 labels at random)

* UNDERSTANDING DEEP LEARNING REQUIRES RETHINKING GENERALIZATION, Zhang et. al (2016)

DNNs and Overparametrisation

- ImageNet contains 1.2M training examples of size 224x224 split between 1000 classes
- Suppose we randomise the labels of all the 1.2M training samples (each receives one of the 1000 labels at random)
- What training accuracy do you expect a model like AlexNet to be able to achieve if left to train to convergence? What about a ResNet-18?

* UNDERSTANDING DEEP LEARNING REQUIRES RETHINKING GENERALIZATION, Zhang et. al (2016)

DNNs and Overparametrisation

- ImageNet contains 1.2M training examples of size 224x224 split between 1000 classes
- Suppose we randomise the labels of all the 1.2M training samples (each receives one of the 1000 labels at random)
- What training accuracy do you expect a model like AlexNet to be able to achieve if left to train to convergence? What about a ResNet-18?
- Why?

* UNDERSTANDING DEEP LEARNING REQUIRES RETHINKING GENERALIZATION, Zhang et. al (2016)

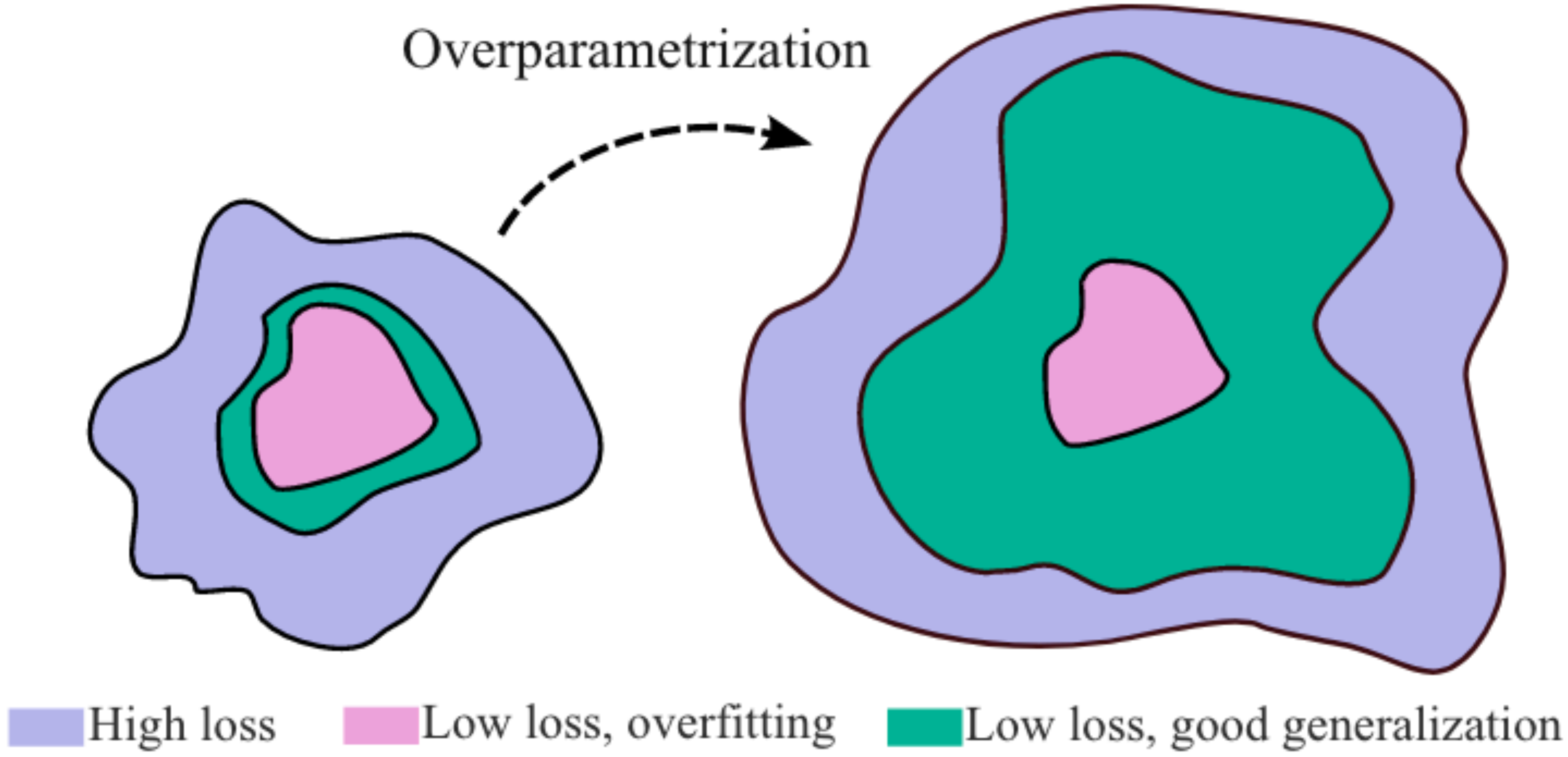
DNNs have the capacity to massively overfit.

DNNs have the capacity to massively overfit (think “memorise”).

DNNs have the capacity to massively overfit (think "memorise").

DNNs have the capacity to massively overfit
(think "memorise").
Why don't they?

More "Good" Solutions Exist

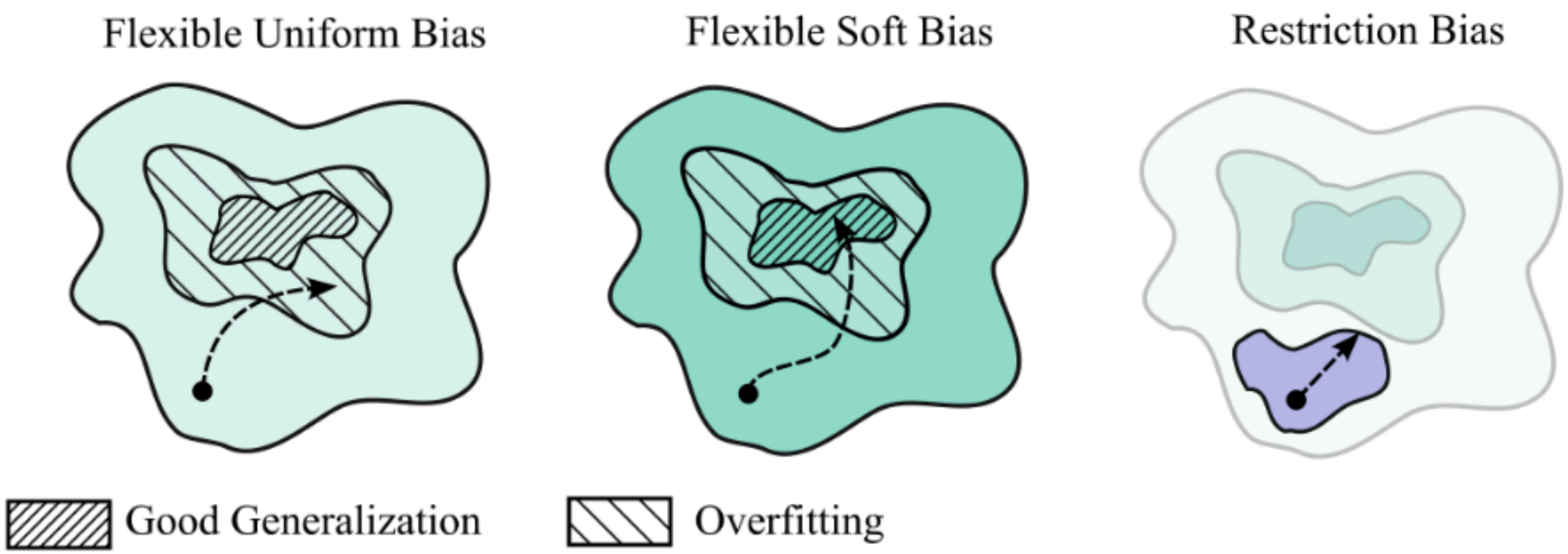


* Deep Learning is Not So Mysterious or Different, Wilson (2025)

Inductive Biases

- Informally:
  - What “can” be expressed
  - What is “likely” to be expressed

Inductive Biases



* Deep Learning is Not So Mysterious or Different, Wilson (2025)

Inductive Biases

Bias towards Low-complexity Solutions

- “Simple” feature is less predictive of the label
- “Complex” feature is more predictive

Bias towards Low-complexity Solutions

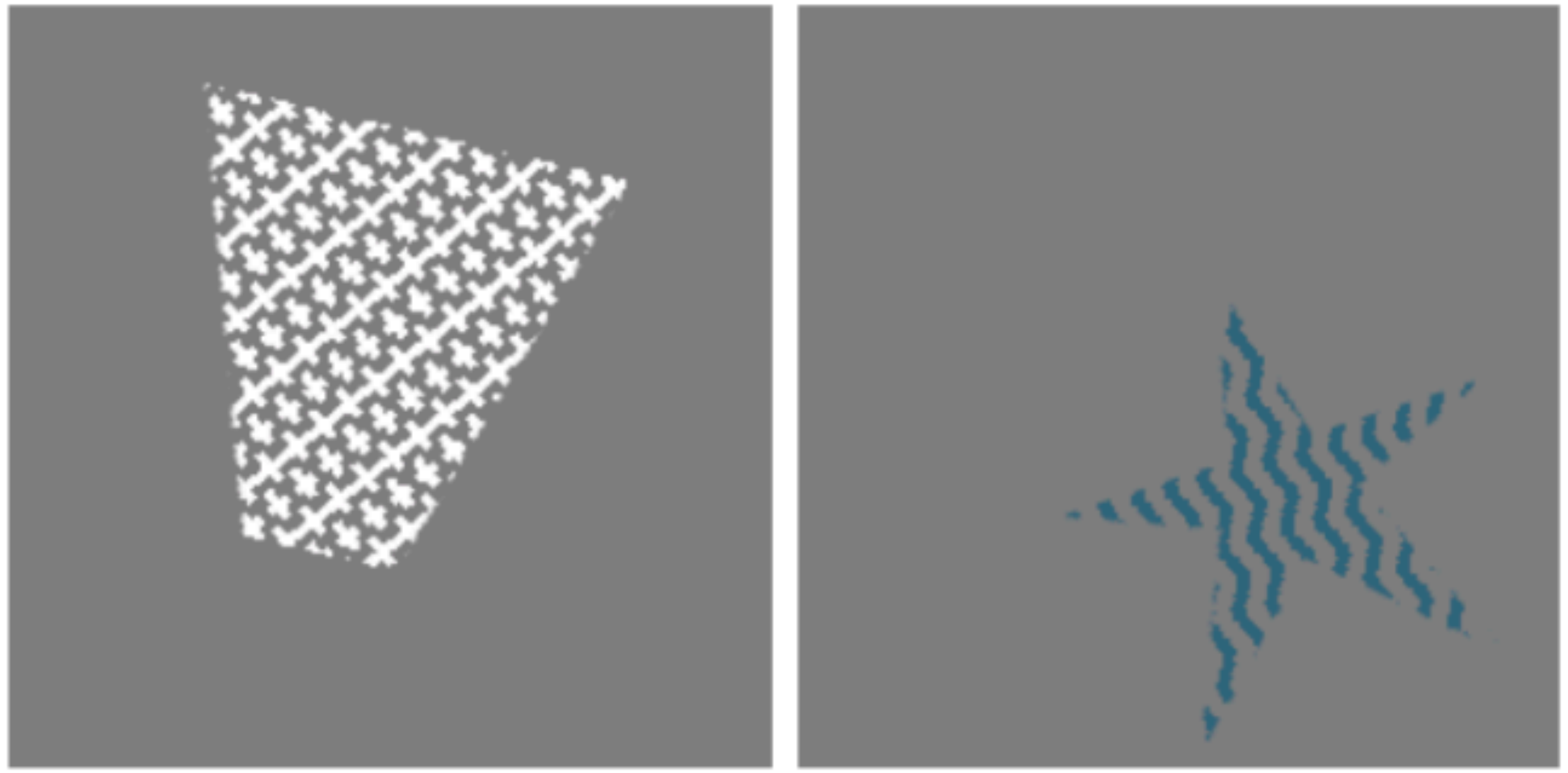
- “Simple” feature is less predictive of the label
- “Complex” feature is more predictive

- Models tend to sacrifice performance over solution complexity*

* What shapes feature representations? Exploring datasets, architectures, and training. Hermann et. al, (2020)

Bias towards Low-complexity Solutions

Trifeature

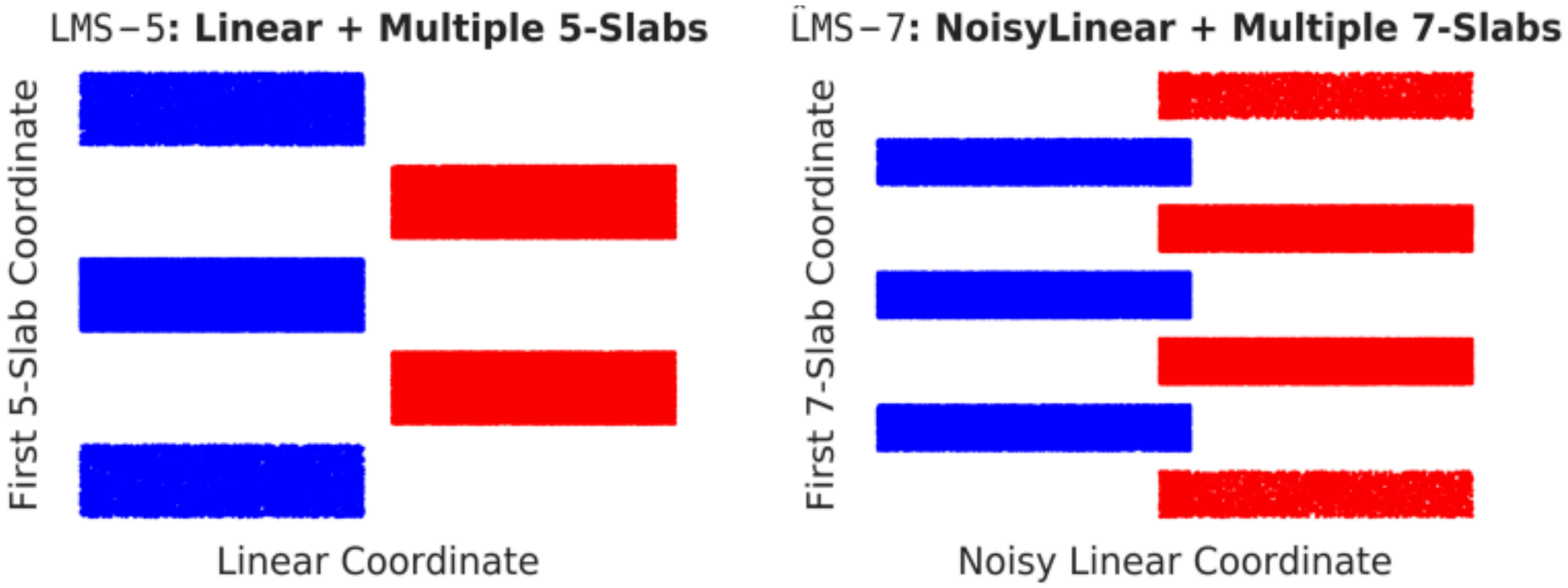


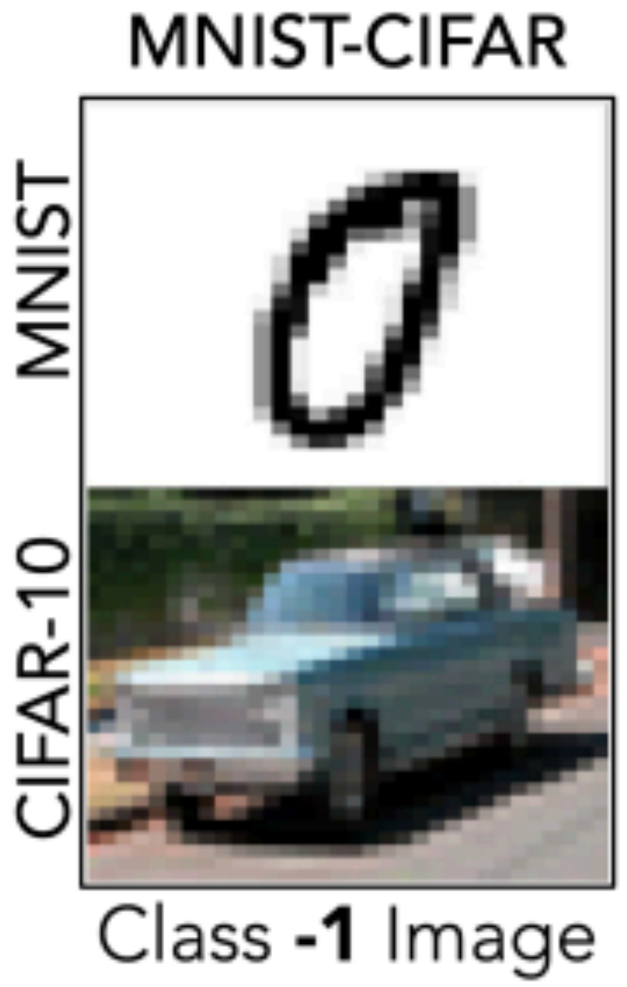
shape: trapezoid
color: white
texture: plus

shape: star
color: ocean
texture: zigzag

* What shapes feature representations? Exploring datasets, architectures, and training. Hermann et. al, (2020)

Bias towards Low-complexity Solutions





* The Pitfalls of Simplicity Bias in Neural Networks. Shah et. al, (2020)

Bias towards Low-complexity Solutions

- Some authors make optimiser-specific arguments
- There is no clear threshold after which the model switches to learning the more “complex” solutions
- Nor is it clear how complexity is defined sometimes

Bias towards Low-complexity Solutions

What can go wrong?

Bias towards Low-complexity Solutions

- In the literature you will find this phenomenon under names such as Simplicity Bias, Shortcut Learning, Gradient starvation.
- Typically seen as something undesirable because of the effect on OOD generalisation
- Although it is used to justify IID generalisation
- How to find the balance is still an open research question

AI Alignment

AI Alignment

AI Alignment

Shipwreck detection?

Back to Inductive Biases

Bias towards Low-complexity Solutions

- “Simple” feature is less predictive of the label
- “Complex” feature is more predictive

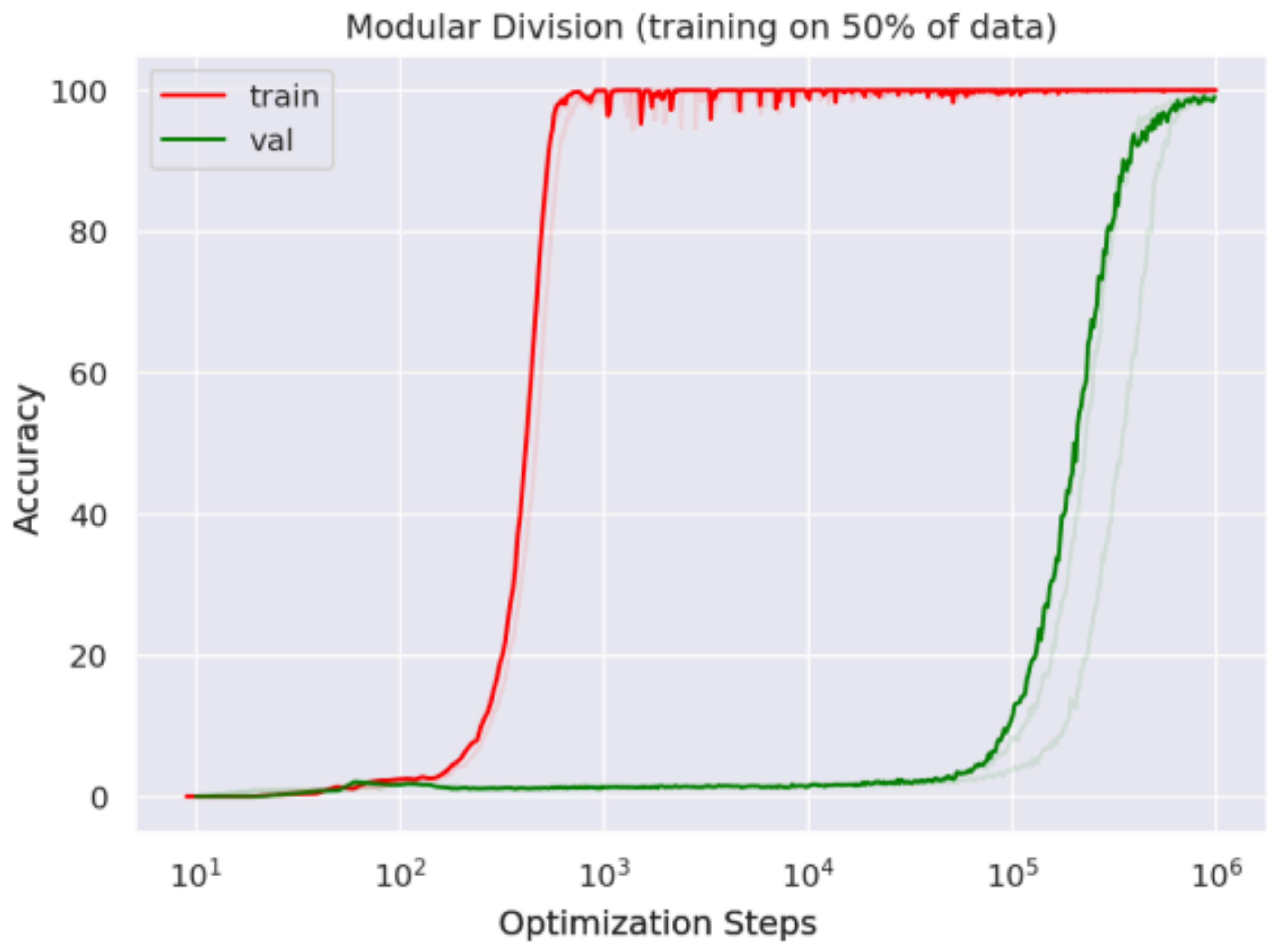
- Models tend to sacrifice performance over solution complexity*

* What shapes feature representations? Exploring datasets, architectures, and training. Hermann et. al, (2020)

What if we just didn't train long enough?

Grokking

Task: division mod 97



* Grokking: Generalization beyond overfitting on small algorithmic datasets. Power et. al, (2022)

What if we just didn’t train long enough?

Still an open research question

Catastrophic Forgetting

Catastrophic Forgetting

- When trained on new things, model performance drops on the old things

Catastrophic Forgetting

- When trained on new things, model performance drops on the old things
  - Sample level
  - Class level
  - Task level

Catastrophic Forgetting

- When trained on new things, model performance drops on the old things
- Storing data might be expensive (or breaking privacy constraints)

Catastrophic Forgetting

- When trained on new things, model performance drops on the old things
- Storing data might be expensive (or breaking privacy constraints)
- Different settings considered in the literature

Catastrophic Forgetting

- When trained on new things, model performance drops on the old things
- Storing data might be expensive (or breaking privacy constraints)
- Different settings considered in the literature
- Continual Learning, Lifelong learning

Catastrophic Forgetting

- When trained on new things, model performance drops on the old things
- Storing data might be expensive (or breaking privacy constraints)
- Different settings considered in the literature
- Continual Learning, Lifelong learning
- Links to simplicity bias (learning dynamics and diversification)

Tunnel Effect

*The Tunnel Effect: Building Data Representations in Deep Neural Networks. Masarczyk et. al, (2023)

Tunnel Effect - Context

(According to the paper)

What is currently known about representations' dependence on layer depth?

*The Tunnel Effect: Building Data Representations in Deep Neural Networks. Masarczyk et. al, (2023)

Tunnel Effect - Context

(According to the paper)

What is currently known about representations' dependence on layer depth?

Layer level

"Networks learn to use layers in the hierarchy by extracting more complex features than the layers before"

Network level

"network depth exponentially enhances capacity* (…) but overparameterized neural networks tend to simplify representations with increasing depth"

*The Tunnel Effect: Building Data Representations in Deep Neural Networks. Masarczyk et. al, (2023)

Tunnel Effect - Context

(According to the paper)

What is currently known about representations' dependence on layer depth?

"So which one is it?"

(Starting point of the paper)

Layer level

"Networks learn to use layers in the hierarchy by extracting more complex features than the layers before"

Network level

"network depth exponentially enhances capacity* (…) but overparameterized neural networks tend to simplify representations with increasing depth"

*The Tunnel Effect: Building Data Representations in Deep Neural Networks. Masarczyk et. al, (2023)

Tunnel Effect

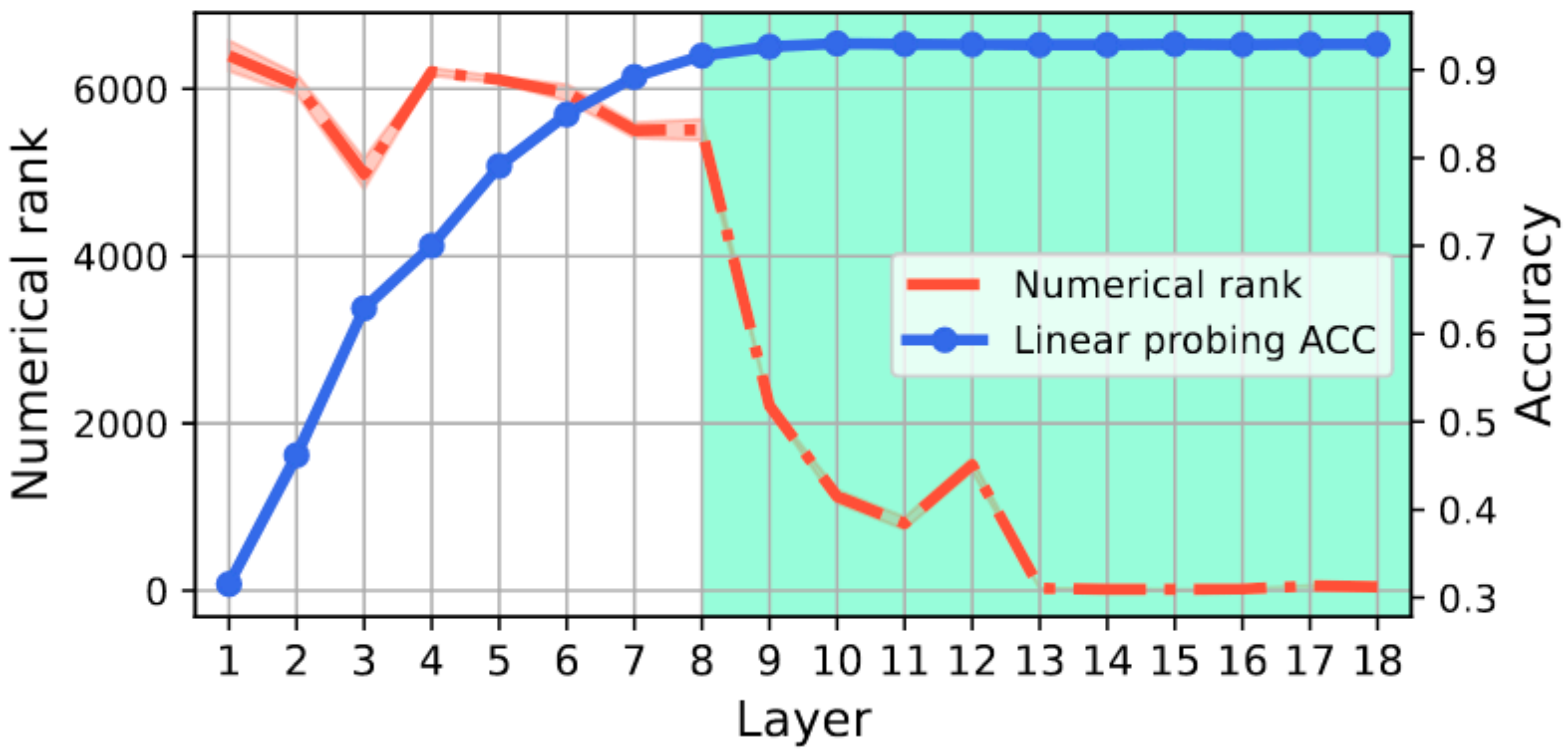


Figure 1: **The tunnel effect** for VGG19 trained on the CIFAR-10. In the tunnel (shaded area), the performance of linear probes attached to each layer saturates (blue line), and the representations rank is steeply reduced (red dashed line).

*The Tunnel Effect: Building Data Representations in Deep Neural Networks. Masarczyk et. al, (2023)

Tunnel Effect

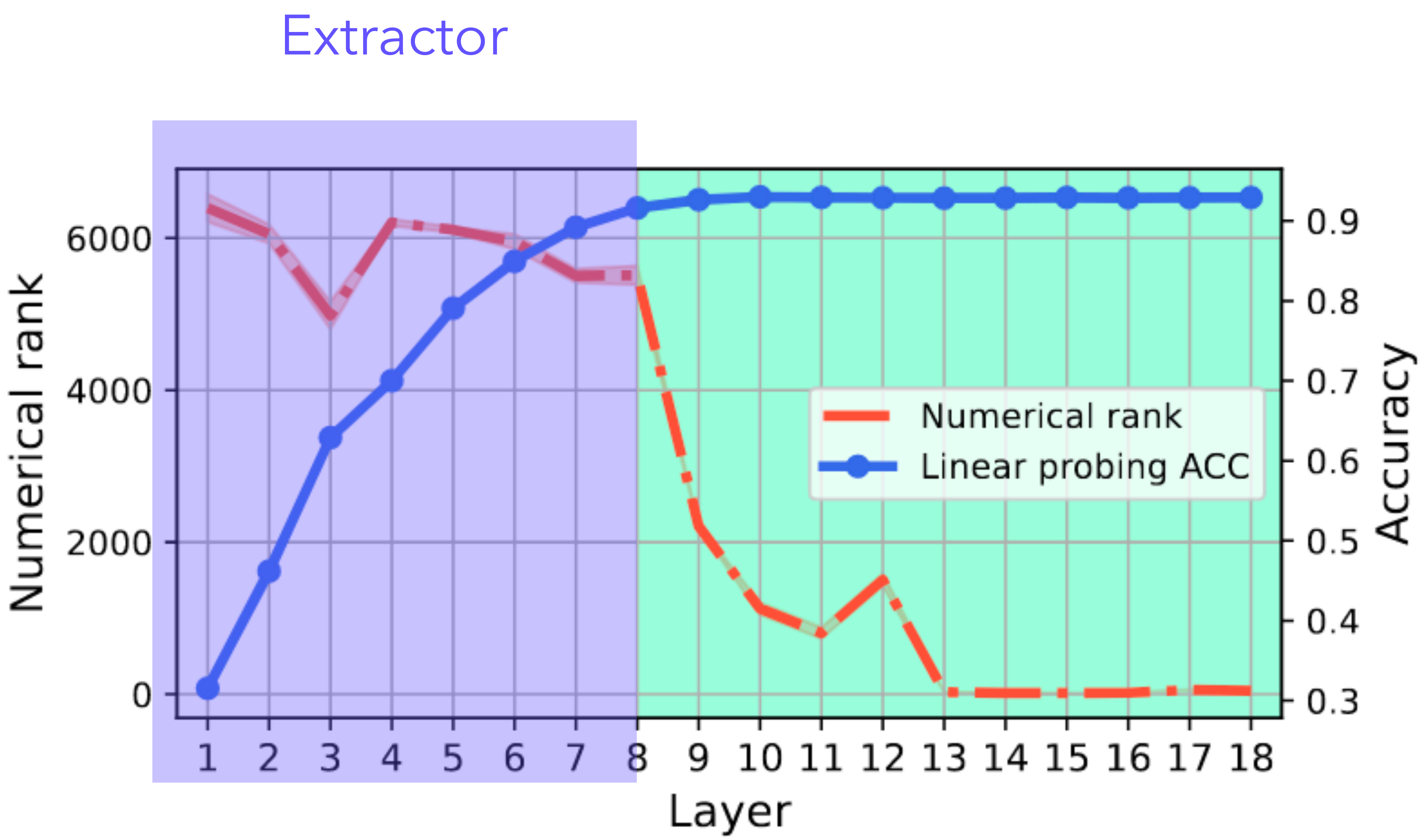


Figure 1: **The tunnel effect** for VGG19 trained on the CIFAR-10. In the tunnel (shaded area), the performance of linear probes attached to each layer saturates (blue line), and the representations rank is steeply reduced (red dashed line).

*The Tunnel Effect: Building Data Representations in Deep Neural Networks. Masarczyk et. al, (2023)

Tunnel Effect

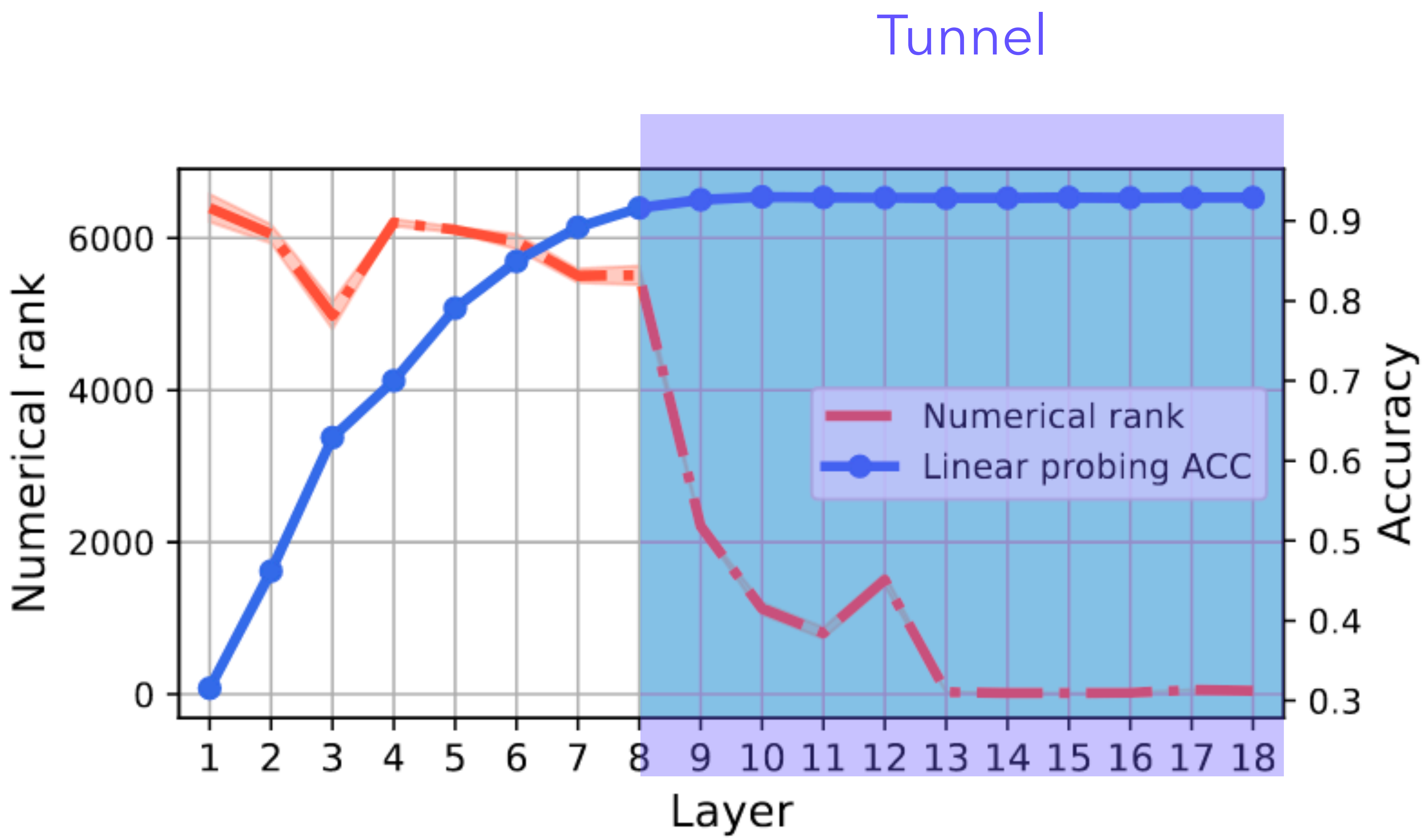


Figure 1: **The tunnel effect** for VGG19 trained on the CIFAR-10. In the tunnel (shaded area), the performance of linear probes attached to each layer saturates (blue line), and the representations rank is steeply reduced (red dashed line).

*The Tunnel Effect: Building Data Representations in Deep Neural Networks. Masarczyk et. al, (2023)

Tunnel Effect

"95% (or 98%) of final accuracy"

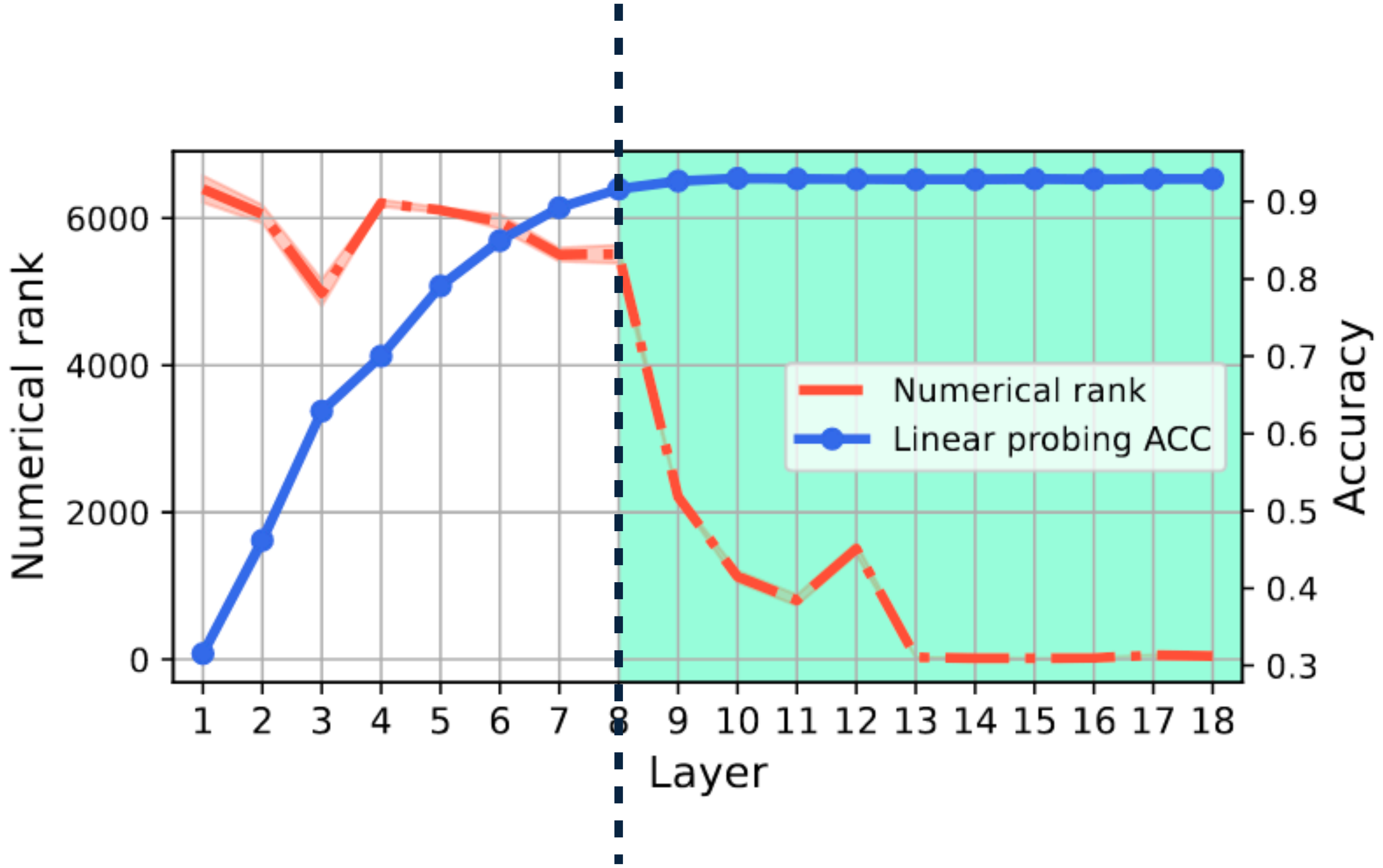


Figure 1: **The tunnel effect** for VGG19 trained on the CIFAR-10. In the tunnel (shaded area), the performance of linear probes attached to each layer saturates (blue line), and the representations rank is steeply reduced (red dashed line).

*The Tunnel Effect: Building Data Representations in Deep Neural Networks. Masarczyk et. al, (2023)

Tunnel Effect

“95% (or 98%) of final accuracy”

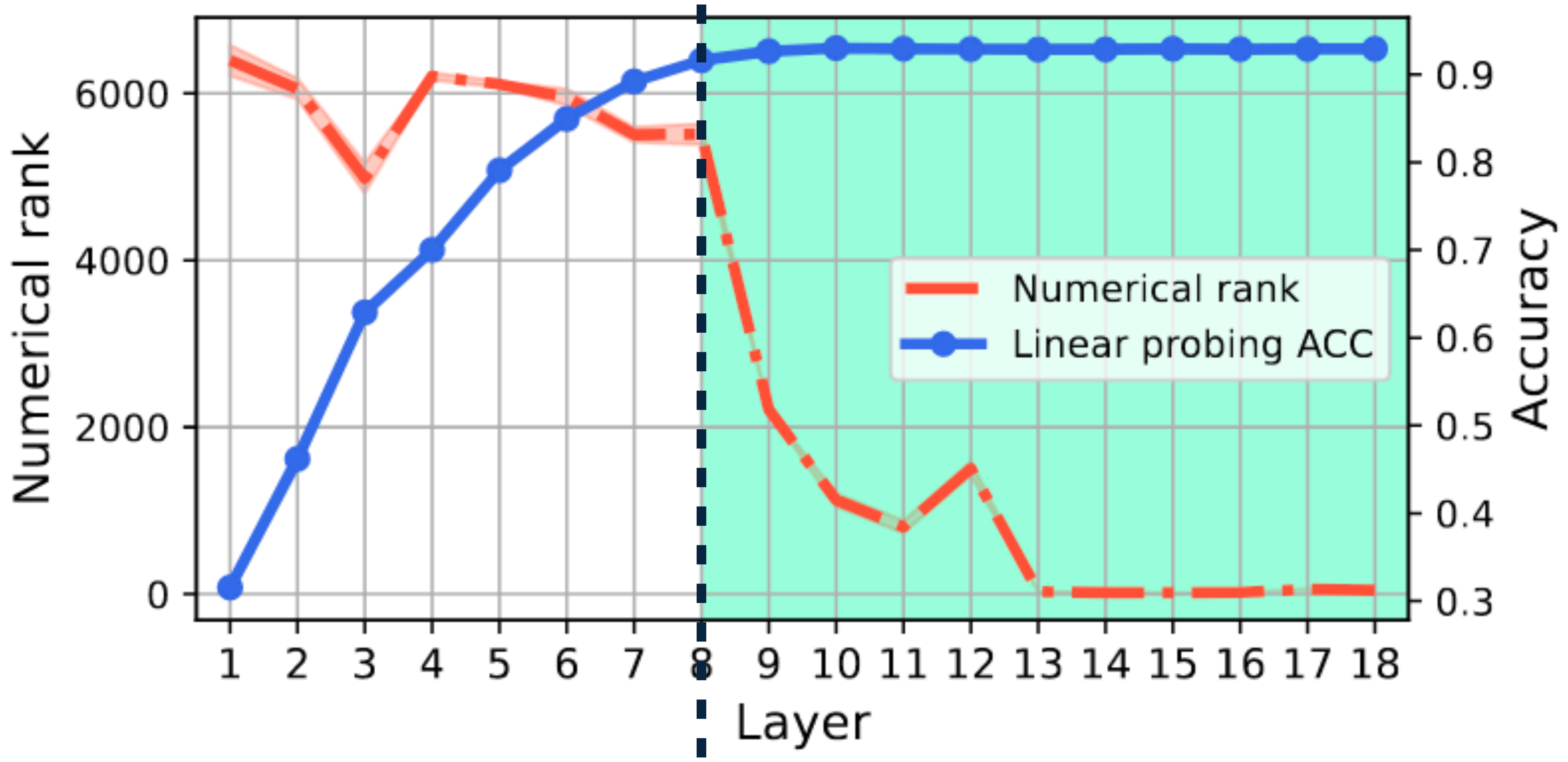


Figure 1: **The tunnel effect** for VGG19 trained on the CIFAR-10. In the tunnel (shaded area), the performance of linear probes attached to each layer saturates (blue line), and the representations rank is steeply reduced (red dashed line).

*The Tunnel Effect: Building Data Representations in Deep Neural Networks. Masarczyk et. al, (2023)

Tunnel Effect

+ linear layer
and
retrain

"95% (or 98%) of final accuracy" → "linear probing"

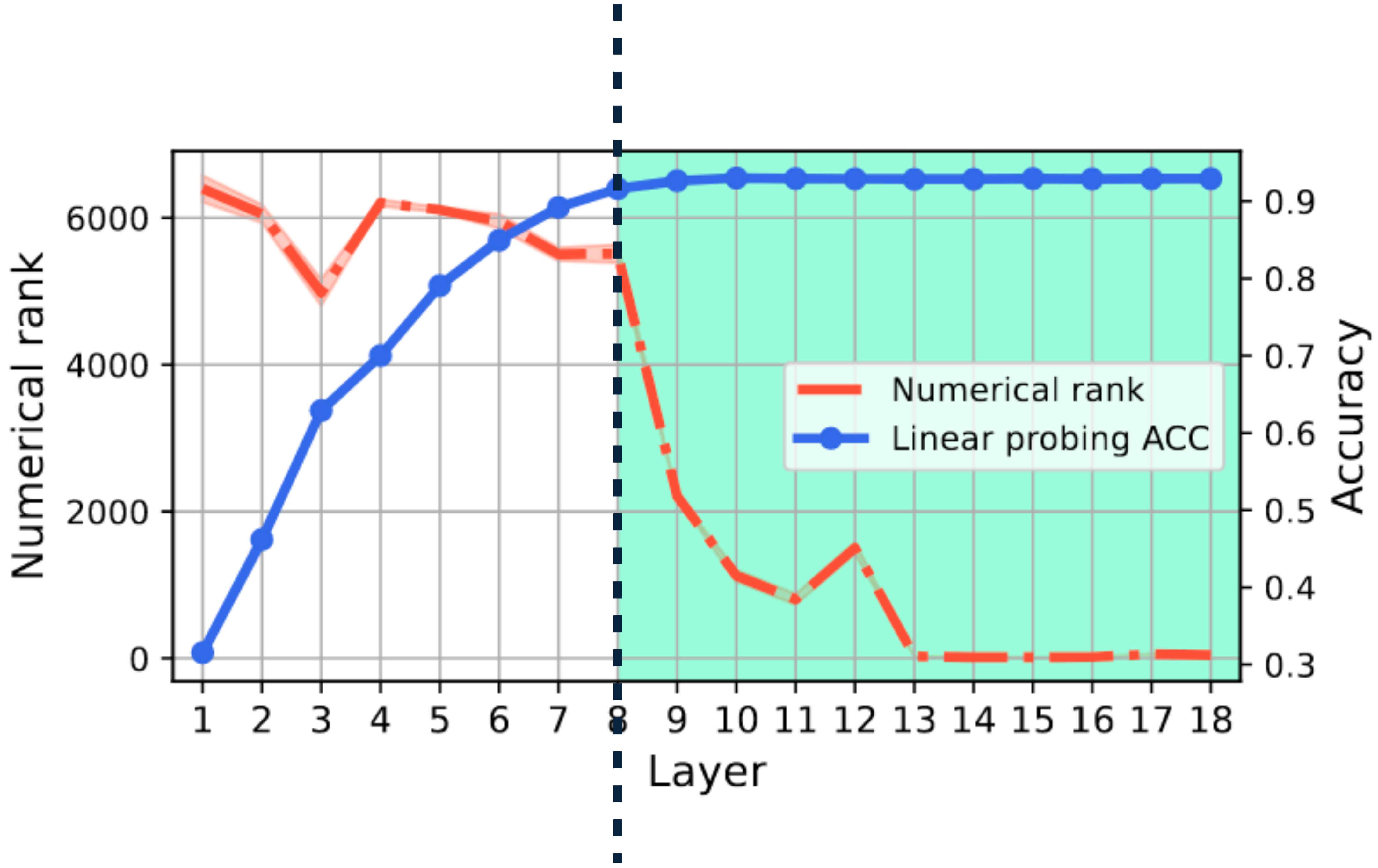


Figure 1: **The tunnel effect** for VGG19 trained on the CIFAR-10. In the tunnel (shaded area), the performance of linear probes attached to each layer saturates (blue line), and the representations rank is steeply reduced (red dashed line).

*The Tunnel Effect: Building Data Representations in Deep Neural Networks. Masarczyk et. al, (2023)

Tunnel Effect

What does the numerical rank capture?

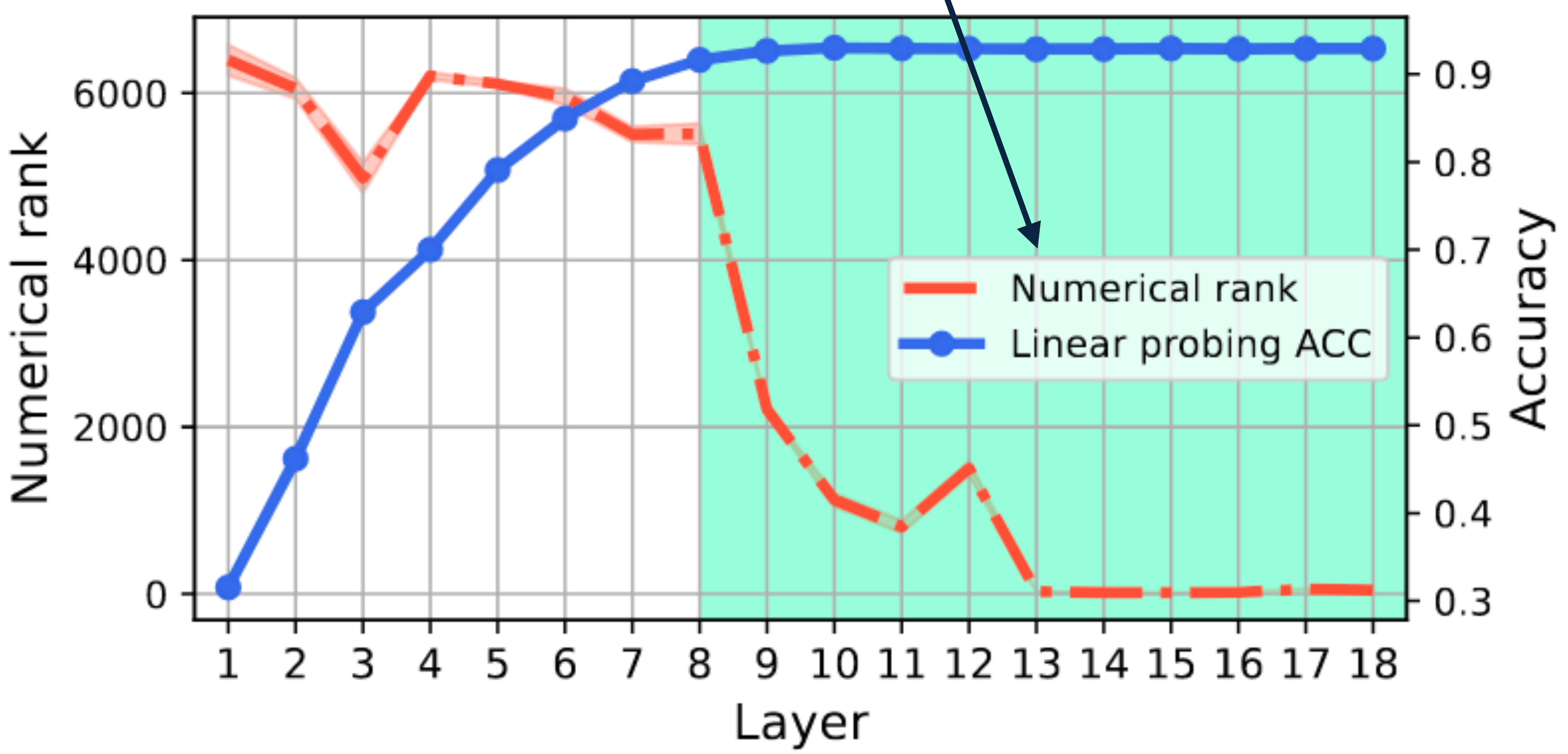


Figure 1: **The tunnel effect** for VGG19 trained on the CIFAR-10. In the tunnel (shaded area), the performance of linear probes attached to each layer saturates (blue line), and the representations rank is steeply reduced (red dashed line).

*The Tunnel Effect: Building Data Representations in Deep Neural Networks. Masarczyk et. al, (2023)

Tunnel Effect - Claims

- “the tunnel develops early during training time”
- “compresses the representations and hinders OOD generalization”
- “its size is correlated with network capacity and dataset complexity”

*The Tunnel Effect: Building Data Representations in Deep Neural Networks. Masarczyk et. al, (2023)

Compression and OOD

- **Motivation**: “intermediate layers perform better than the penultimate ones for transfer learning”

*The Tunnel Effect: Building Data Representations in Deep Neural Networks. Masarczyk et. al, (2023)

Compression and OOD

- **Motivation**: "intermediate layers perform better than the penultimate ones for transfer learning"

- **OOD:**

CIFAR-10 $\xrightarrow[\text{probing}]{\text{linear}}$ 10-class CIFAR-100 subset

*The Tunnel Effect: Building Data Representations in Deep Neural Networks. Masarczyk et. al, (2023)

Compression and Transfer

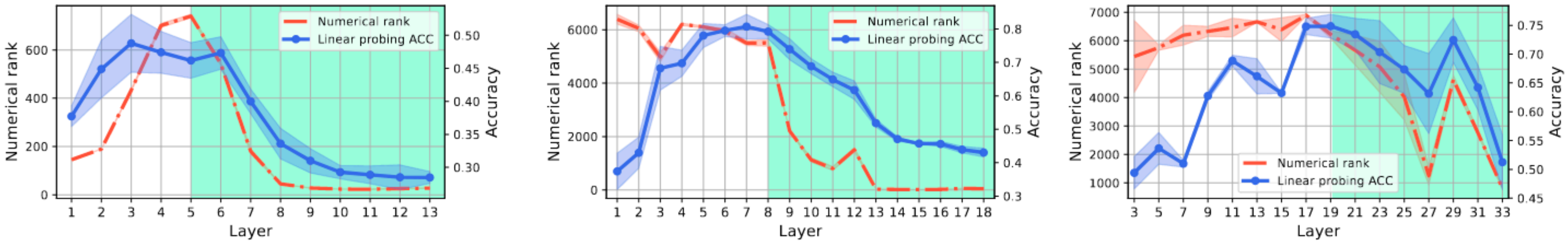
- **Motivation**: "intermediate layers perform better than the penultimate ones for transfer learning"

- **OOD:**

$$\text{CIFAR-10} \xrightarrow[\text{probing}]{\text{linear}} \text{10-class CIFAR-100 subset}$$

*The Tunnel Effect: Building Data Representations in Deep Neural Networks. Masarczyk et. al, (2023)

Compression and Transfer



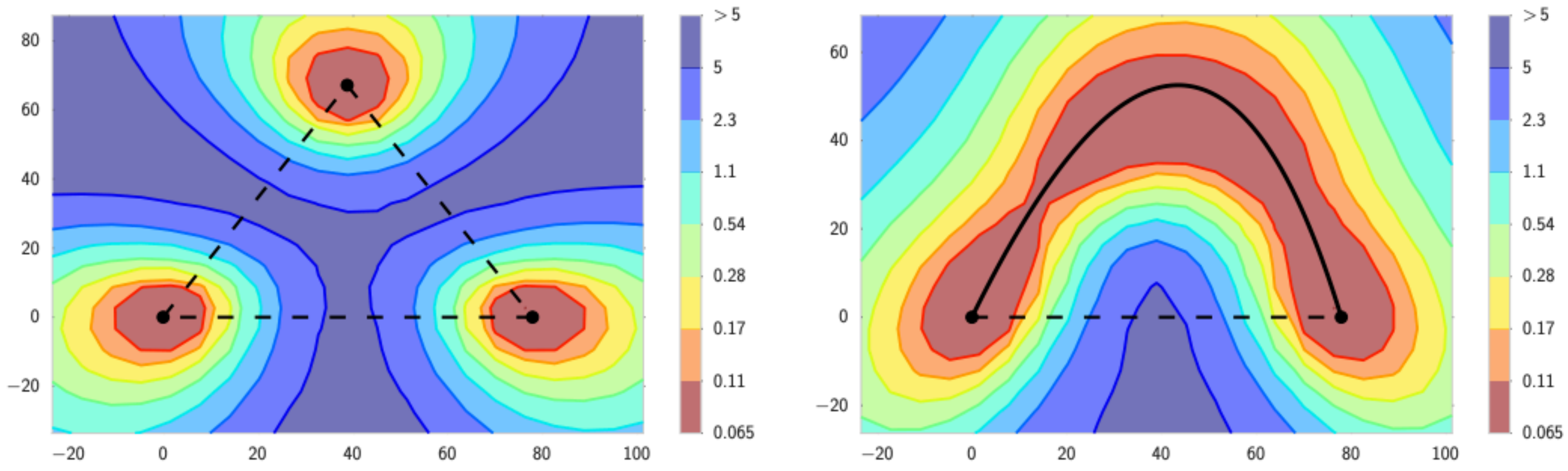
(a) MLP 12 (b) VGG-19 (c) ResNet 34

"The tunnel degrades the out-of-distribution performance correlated with the representations' numerical rank"

*The Tunnel Effect: Building Data Representations in Deep Neural Networks. Masarczyk et. al, (2023)

Mode Connectivity

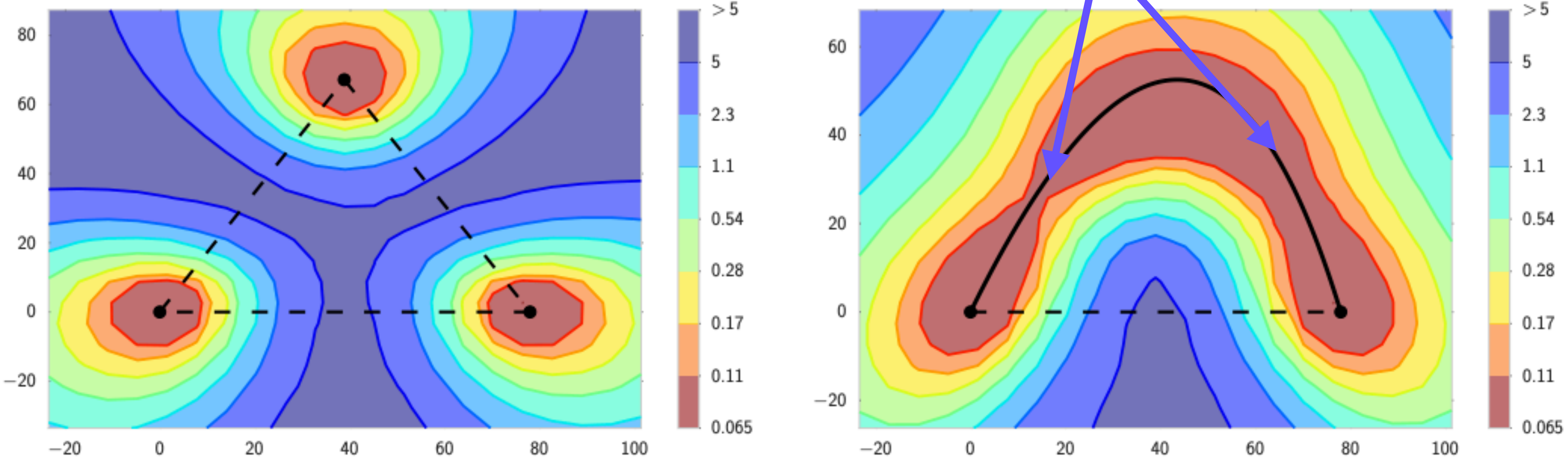
Mode Connectivity



* Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs. Garipov et. al, (2018)

Mode Connectivity

Different predictions on test samples!



* Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs. Garipov et. al, (2018)

Mode Connectivity

"It is a misnomer to even refer to the converged solutions as local optima!"

* Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs. Garipov et. al, (2018)

Take-aways

- There are many things we still don’t understand about DNNs

Take-aways

- There are many things we still don’t understand about DNNs
- but we are slowly starting to gain some intuitions and understanding.

Take-aways

- There are many things we still don’t understand about DNNs
- but we are slowly starting to gain some intuitions and understanding.
- While doing so, we also infer practical implications. What examples can you think of/remember from today’s lecture?

Take-aways

- There are many things we still don't understand about DNNs
- but we are slowly starting to gain some intuitions and understanding.
- While doing so, we also infer practical implications. What examples can you think of/remember from today's lecture?
- Many interesting phenomena are tightly linked to properties of learned representations and training dynamics

Take-aways

- There are many things we still don't understand about DNNs
- but we are slowly starting to gain some intuitions and understanding.
- While doing so, we also infer practical implications. What examples can you think of/remember from today's lecture?
- Many interesting phenomena are tightly linked to properties of learned representations and training dynamics and after finishing this module you should know the basics required to join the conversation!