ResMem:

Learn what you can and memorize the rest

Speaker: Zitong Yang

Paper link: https://arxiv.org/pdf/2302.01576.pdf

Collaborators

Michal Lukasik

Vaishnavh Nagarajan

Zonglin Li

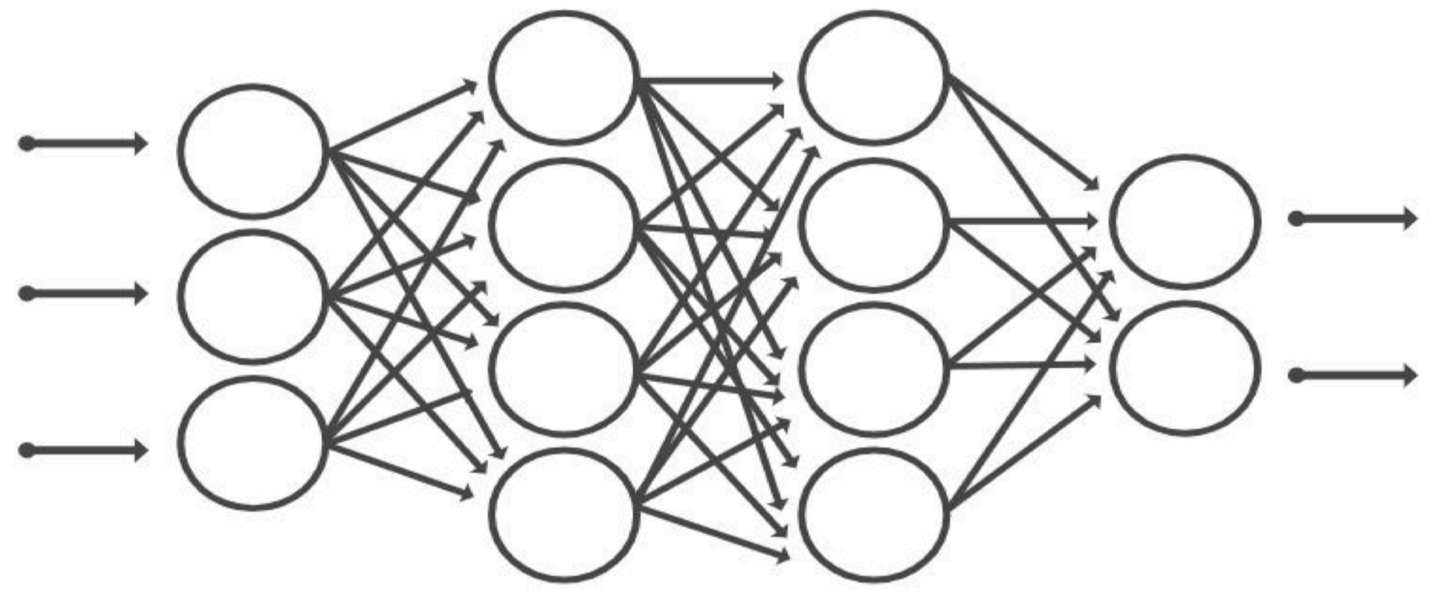
Ankit Rawat

Manzil Zaheer

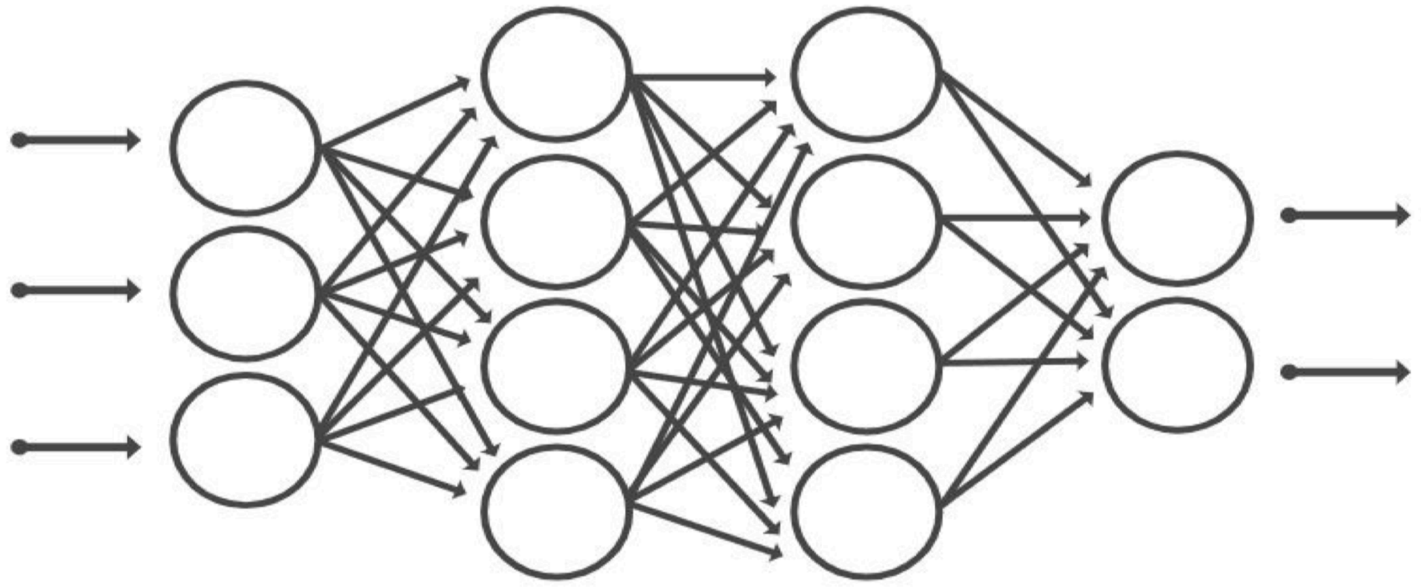
Aditya Menon

Sanjiv Kumar

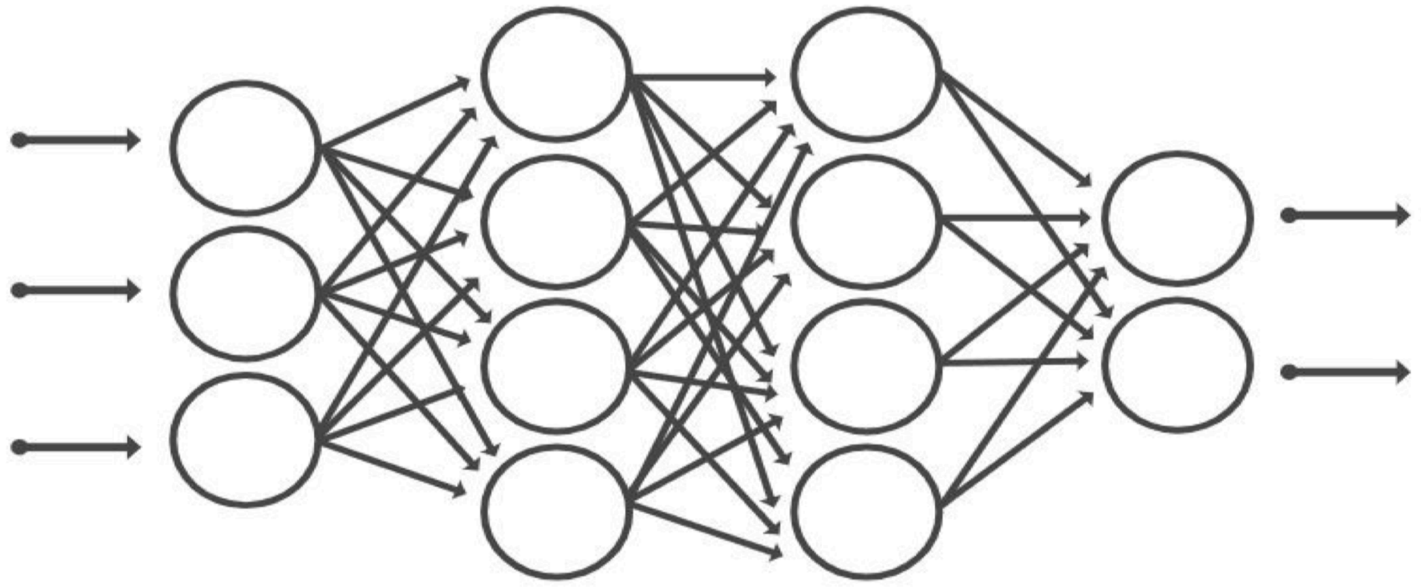
Google Research



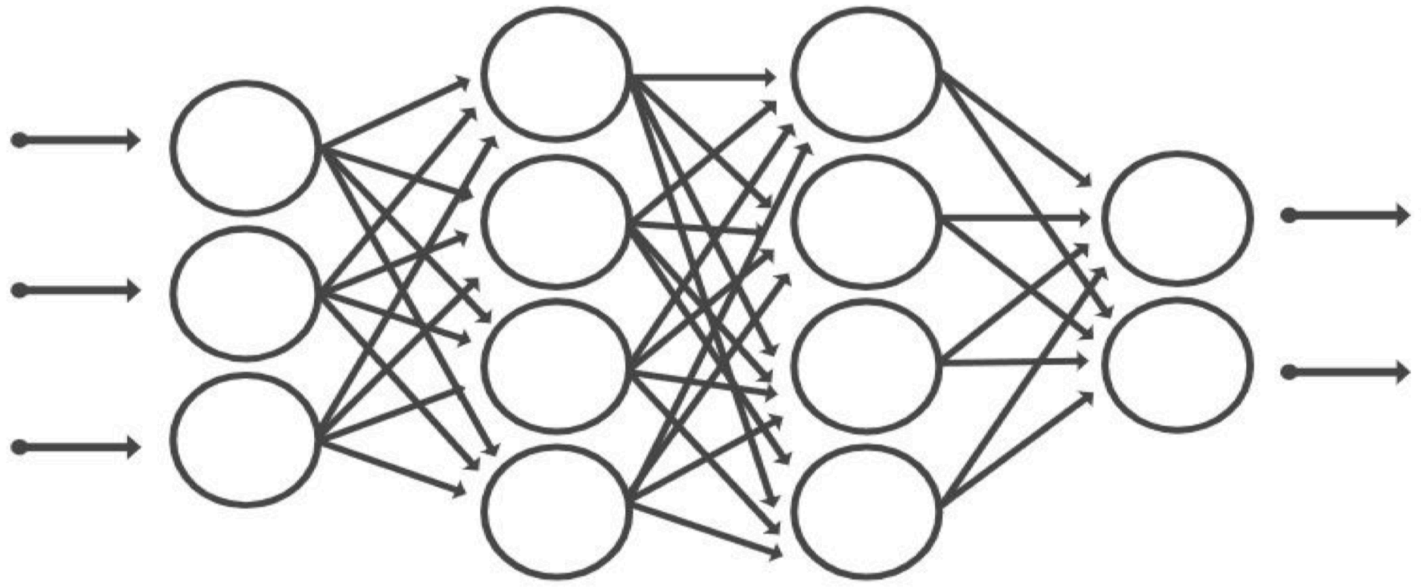
rose



rose

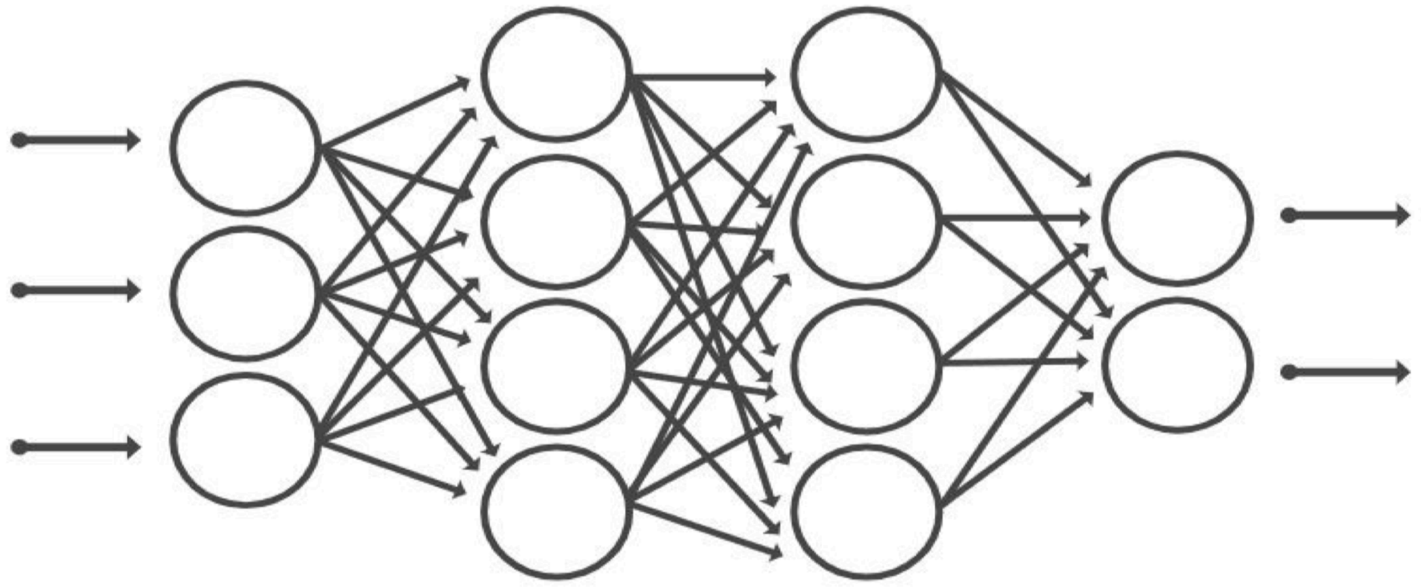


poppy



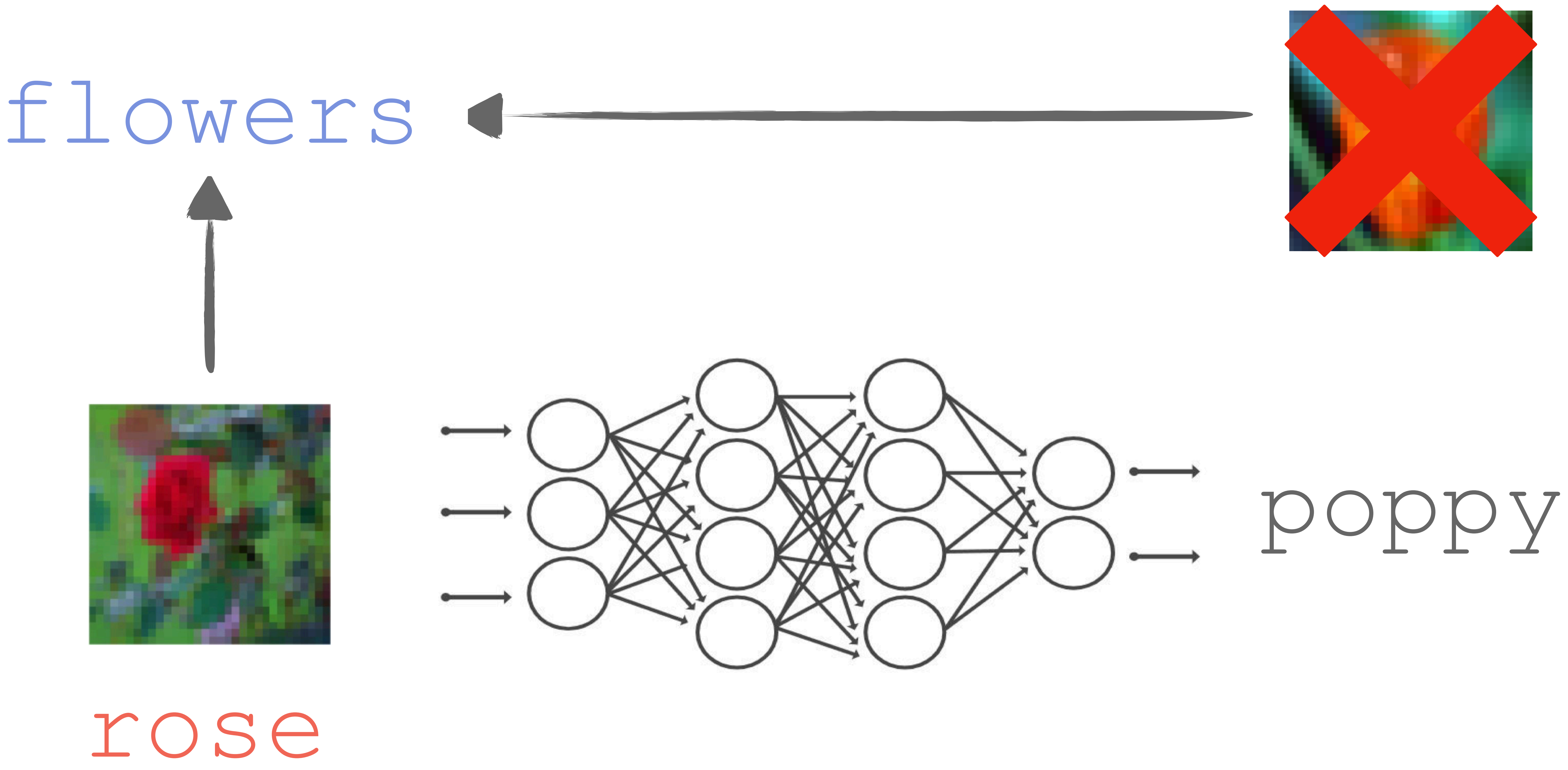
poppy

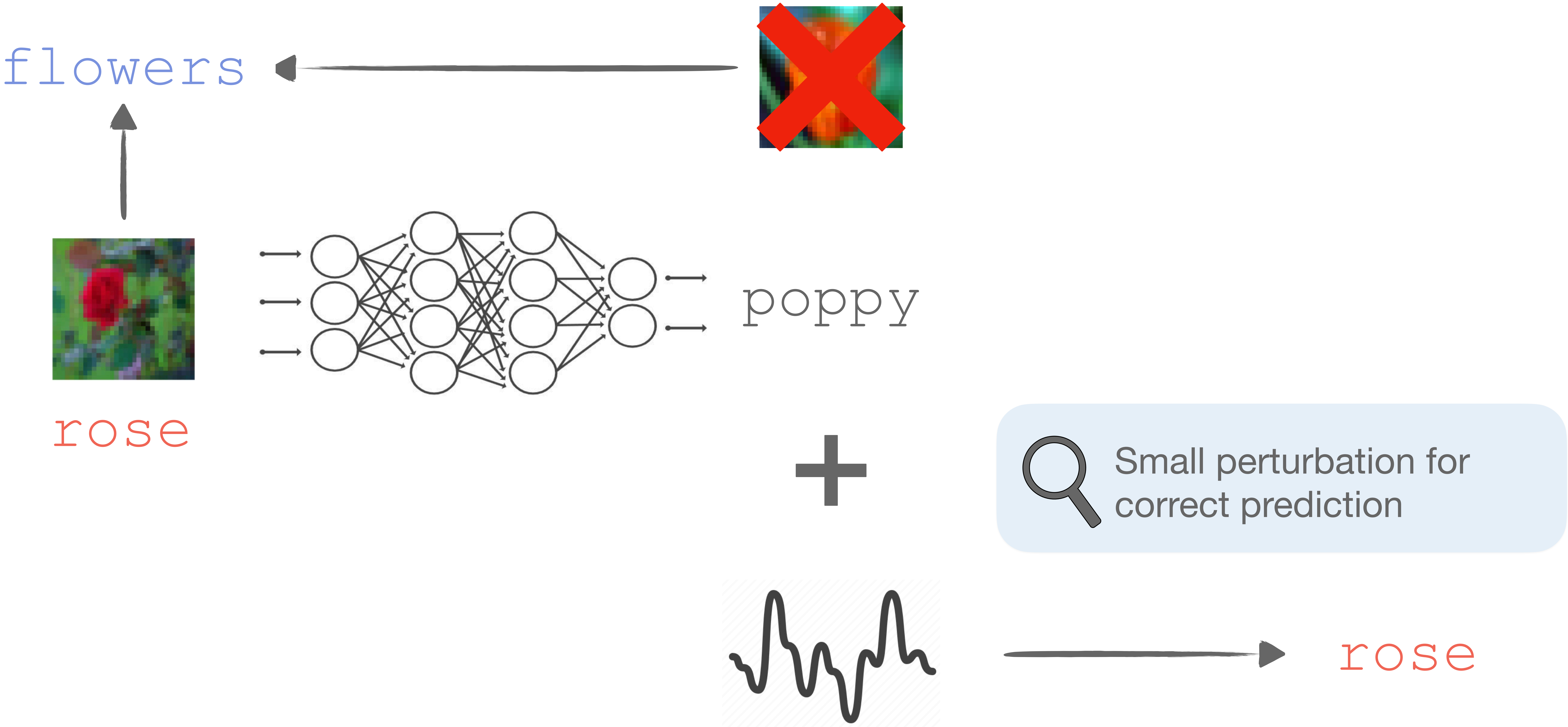
rose

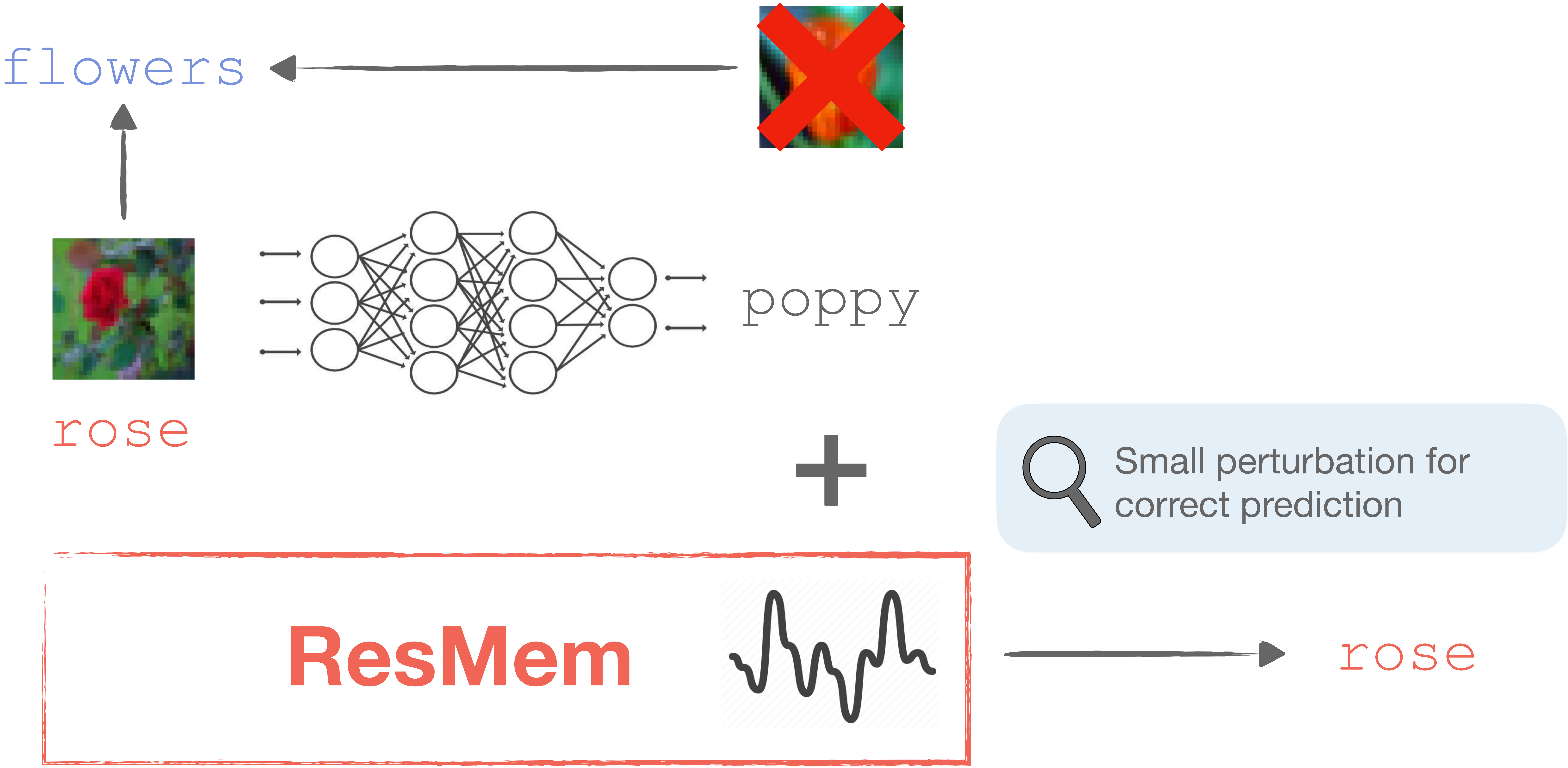


poppy

rose


flowers
rose
poppy


flowers
poppy
rose
Small perturbation for correct prediction
rose


flowers
poppy
rose
Small perturbation for correct prediction
ResMem
rose

Background: Memorization $\Longleftrightarrow$ Generalization

- **[Memorization $\Longrightarrow$ Generalization]** SoTA large neural networks generalizes well, despite memorizing the training data. *(Zhang et al., ‘17; Bartlett et al., ‘20; Kapalan et al., ‘20)*

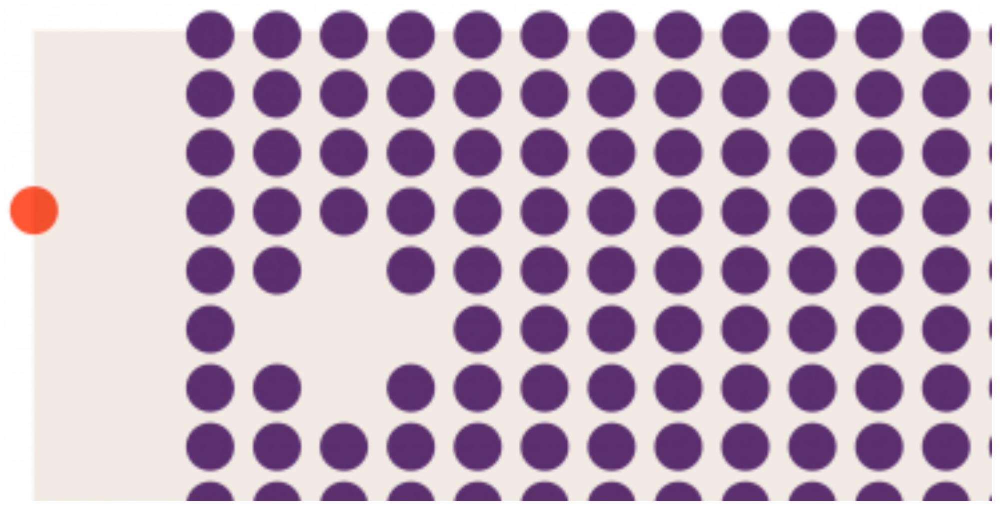
Background: Memorization $\Longleftrightarrow$ Generalization

- **[Memorization $\Longrightarrow$ Generalization]** SoTA large neural networks generalizes well, despite memorizing the training data. *(Zhang et al., '17; Bartlett et al., '20; Kapalan et al., '20)*
- **[Generalization $\Longrightarrow$ Memorization]** More surprisingly, memorization can be <u>necessary</u> for generalization. *(Feldman et al., '19; Feldman et al., '20; Chen et al., '20)*

Background: Memorization $\Longleftrightarrow$ Generalization

- [Memorization $\Longrightarrow$ Generalization] SoTA large neural networks generalizes well, despite memorizing the training data. *(Zhang et al., '17; Bartlett et al., '20; Kapalan et al., '20)*

- [Generalization $\Longrightarrow$ Memorization] More surprisingly, memorization can be necessary for generalization. *(Feldman et al., '19; Feldman et al., '20; Chen et al., '20)*

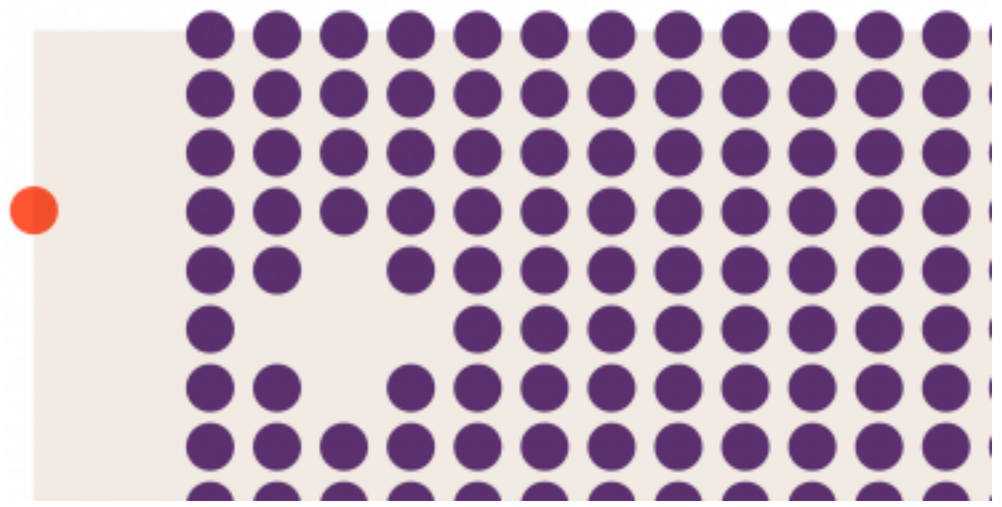


Computationally intractable to distinguish ***outliers*** *and* ***rare examples****.*

Background: Memorization $\Longleftrightarrow$ Generalization

- **[Memorization $\Longrightarrow$ Generalization]** SoTA large neural networks generalizes well, despite memorizing the training data. *(Zhang et al., '17; Bartlett et al., '20; Kapalan et al., '20)*

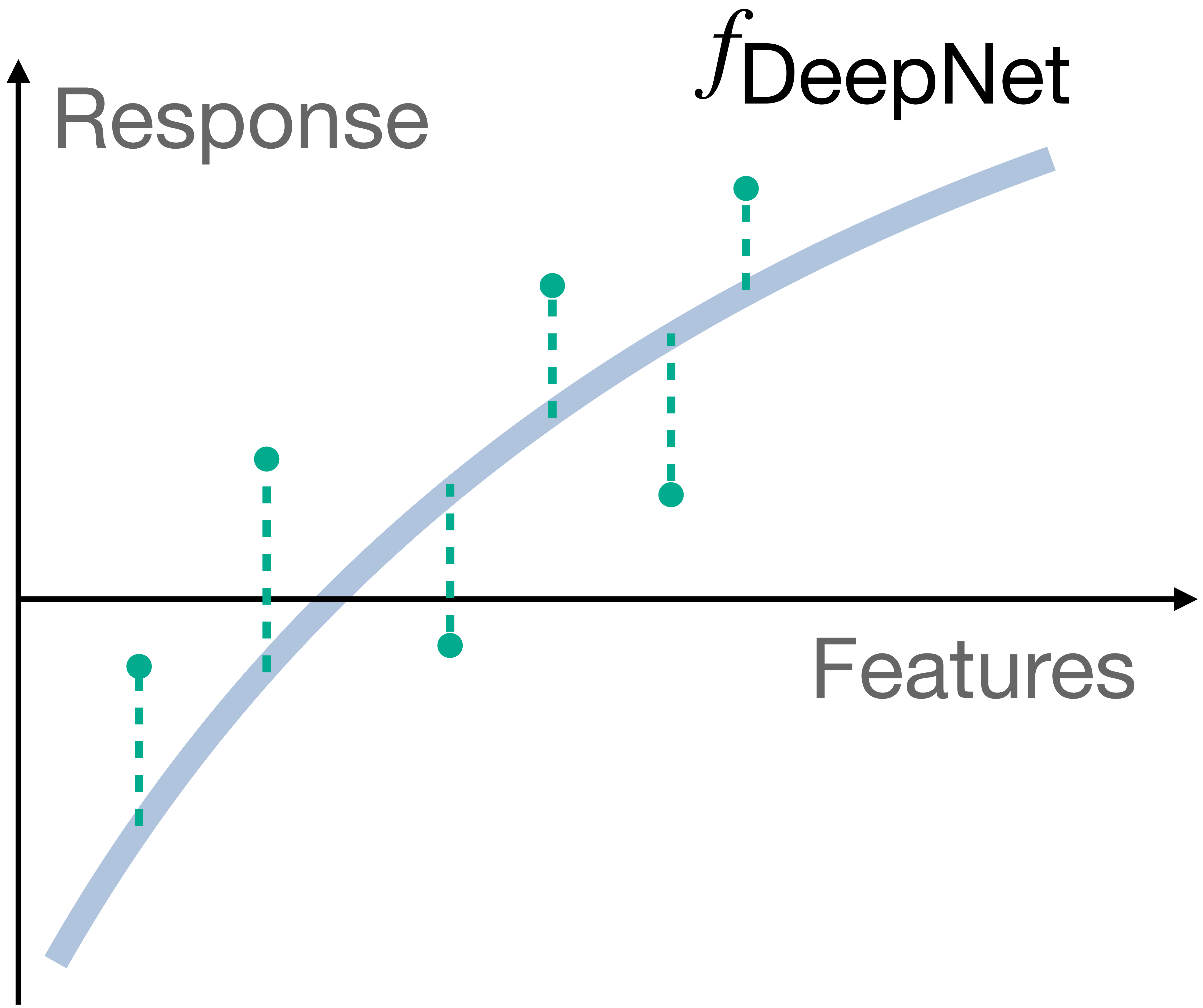
- **[Generalization $\Longrightarrow$ Memorization]** More surprisingly, memorization can be necessary for generalization. *(Feldman et al., '19; Feldman et al., '20; Chen et al., '20)*



Computationally intractable to distinguish ***outliers*** *and* ***rare examples****.*

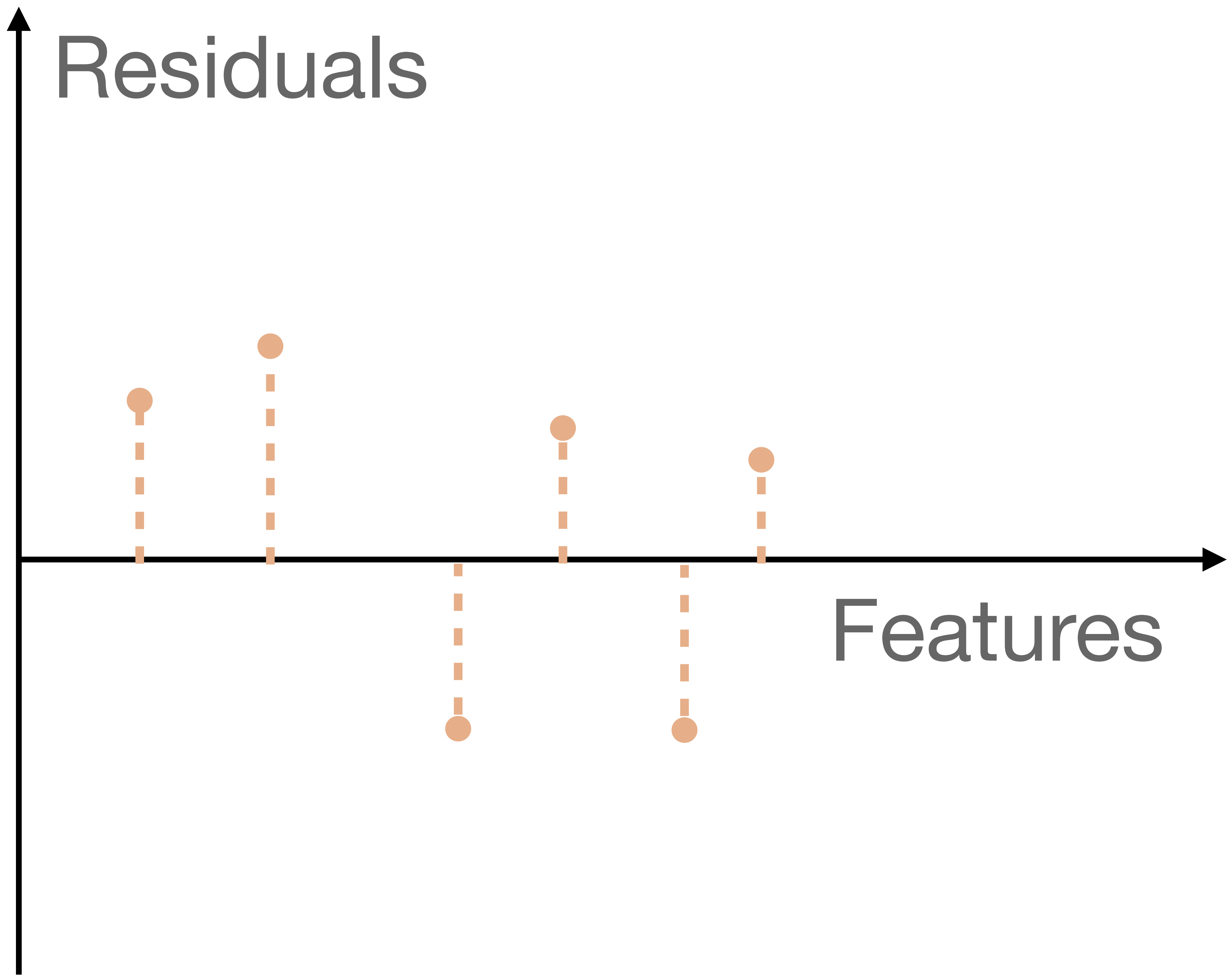
- **Question:** Explicit memorization for generalization?

ResMem: residual memorization



- **[Step 1]** Train a neural network f_{DeepNet} as usual.

ResMem: residual memorization

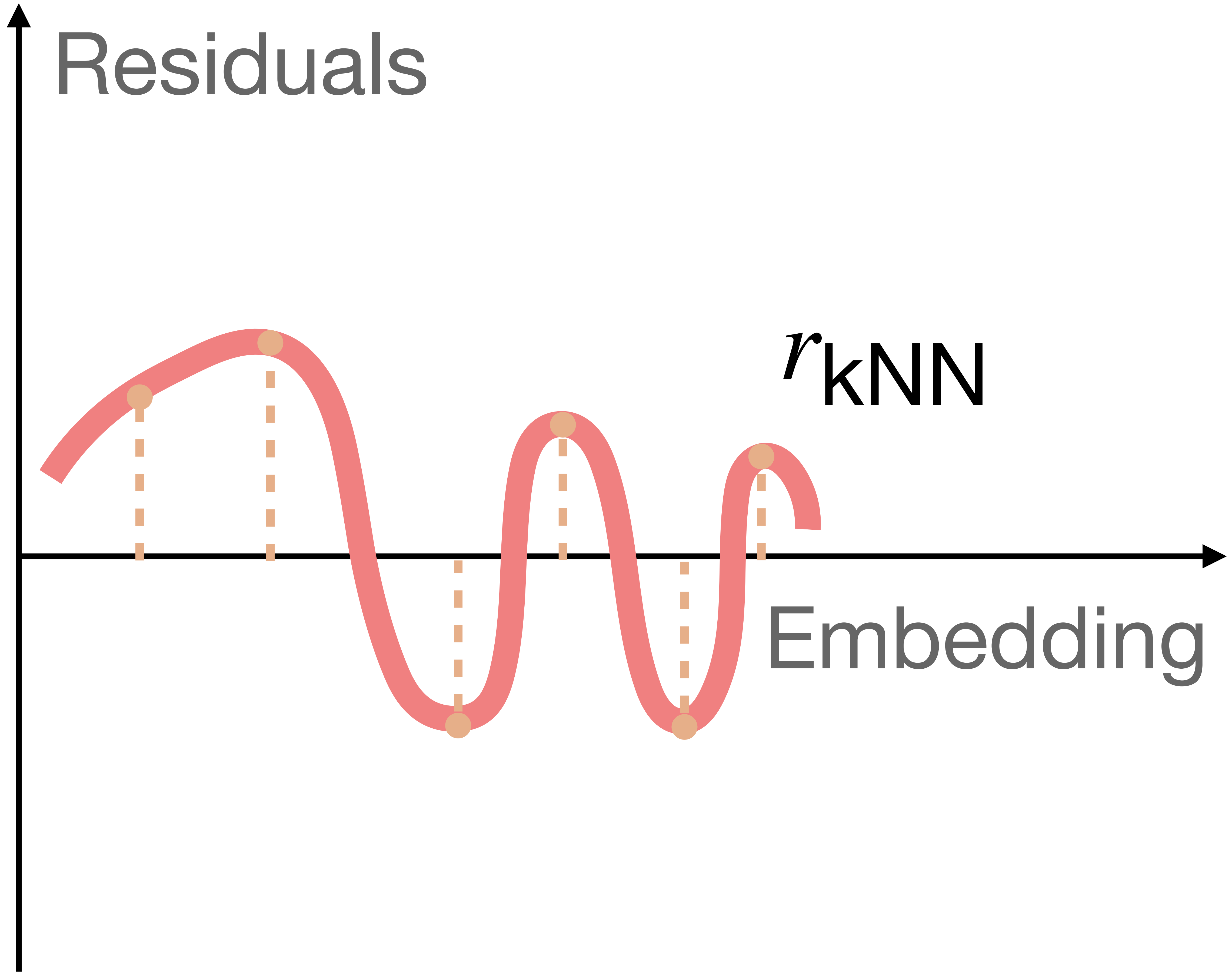


- [Step 1] Train a neural network f_{DeepNet} as usual.

- **[Step 2]** Compute the residuals. For classification,

$$r_i = \text{onehot}(y_i) - \text{softmax}\left(f_{\text{DeepNet}}(x_i)\right).$$

ResMem: residual memorization



- [Step 1] Train a neural network f_{DeepNet} as usual.

- [Step 2] Compute the residuals. For classification,

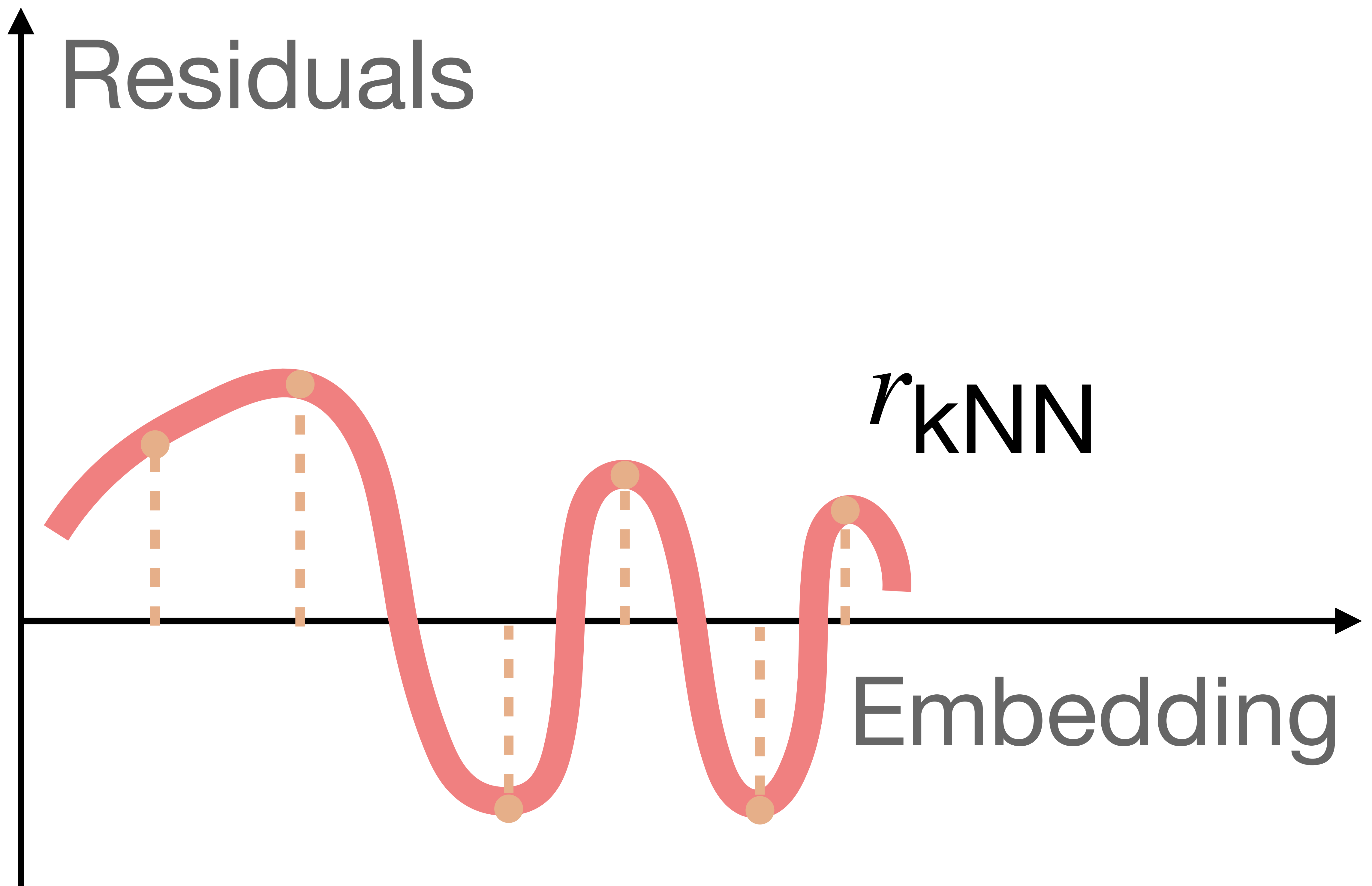
$$r_i = \text{onehot}(y_i) - \text{softmax}\left(f_{\text{DeepNet}}(x_i)\right).$$

- **[Step 3]** Memorize residuals using a k-NN

$$r_{\text{kNN}}(\tilde{x}) = \sum_i r_i \cdot w_i\,,$$

where $w_i \sim -\|\phi(\tilde{x}) - \phi(x_i)\|$ is computed from the intermediate layer of the neural network.

ResMem: residual memorization



- [Step 1] Train a neural network f_{DeepNet} as usual.

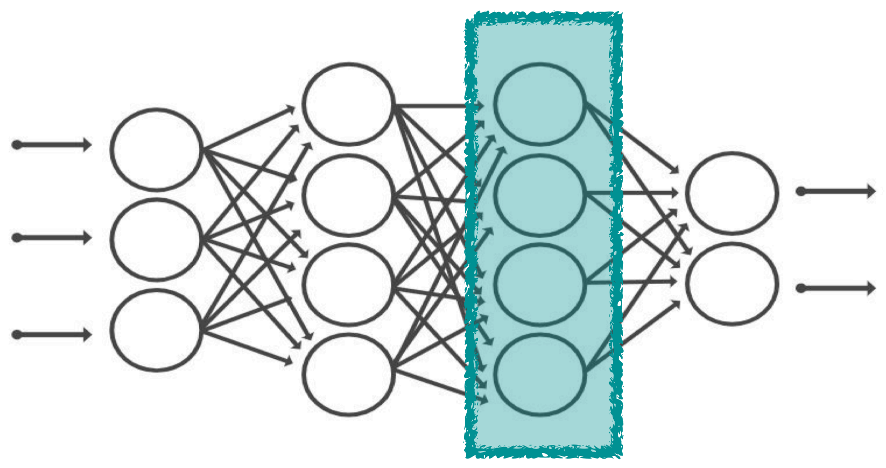
- [Step 2] Compute the residuals. For classification,

$$r_i = \text{onehot}(y_i) - \text{softmax}\left(f_{\text{DeepNet}}(x_i)\right).$$

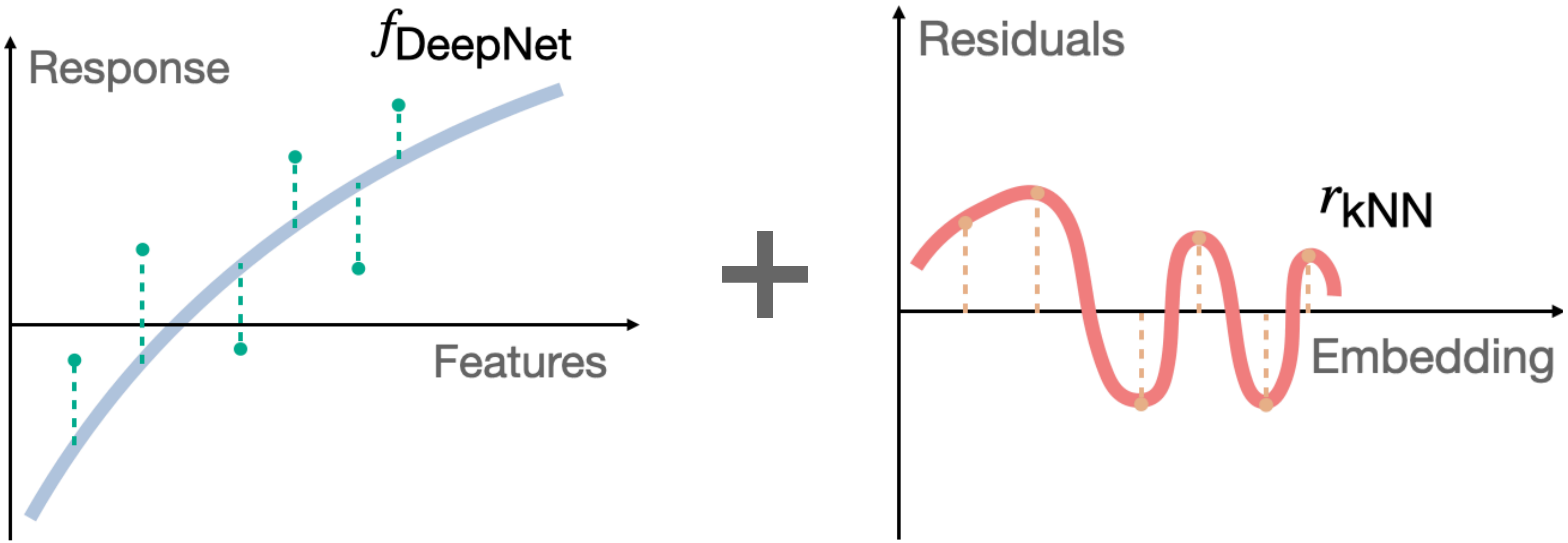
- **[Step 3]** Memorize residuals using a k-NN

$$r_{\text{kNN}}(\widetilde{x}) = \sum_i r_i \cdot w_i \,,$$

where $w_i \sim -\|\phi(\tilde{x}) - \phi(x_i)\|$ is computed from the intermediate layer of the neural network.



ResMem: residual memorization



$$\text{ResMem prediction} = f_{\text{DeepNet}} + r_{\text{kNN}}$$

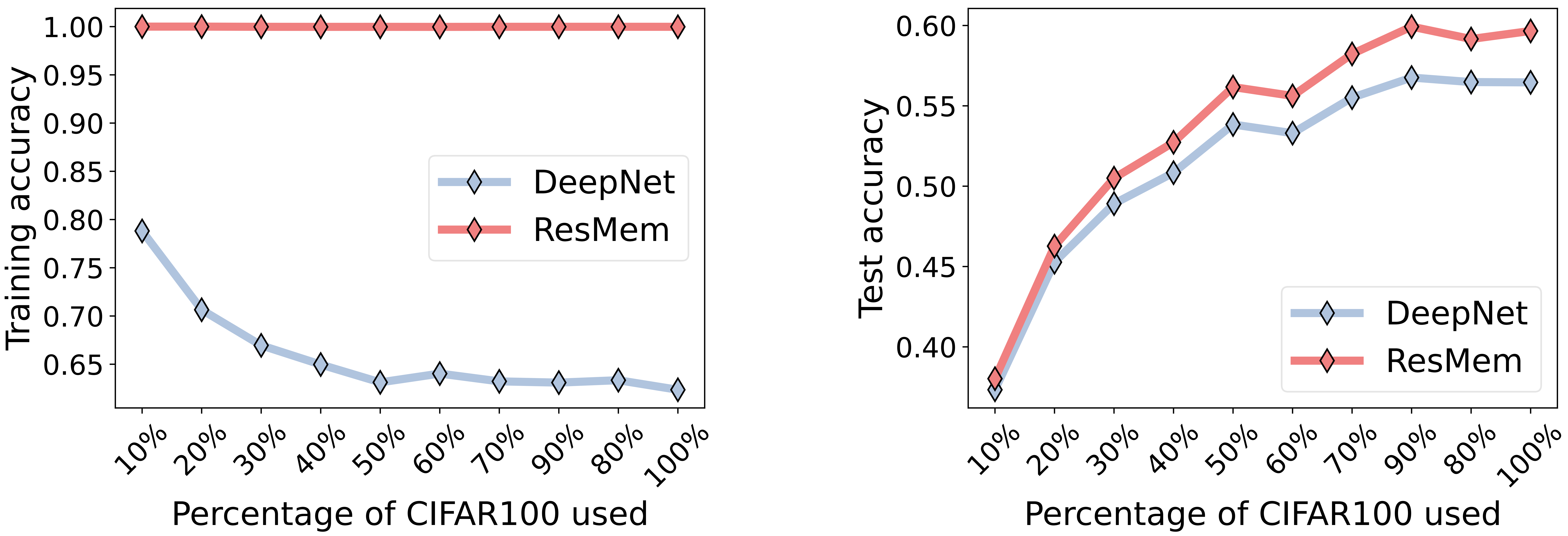
Empirical results:

Dataset	Architecture	Test accuracy		
		DeepNet	ResMem	Gain

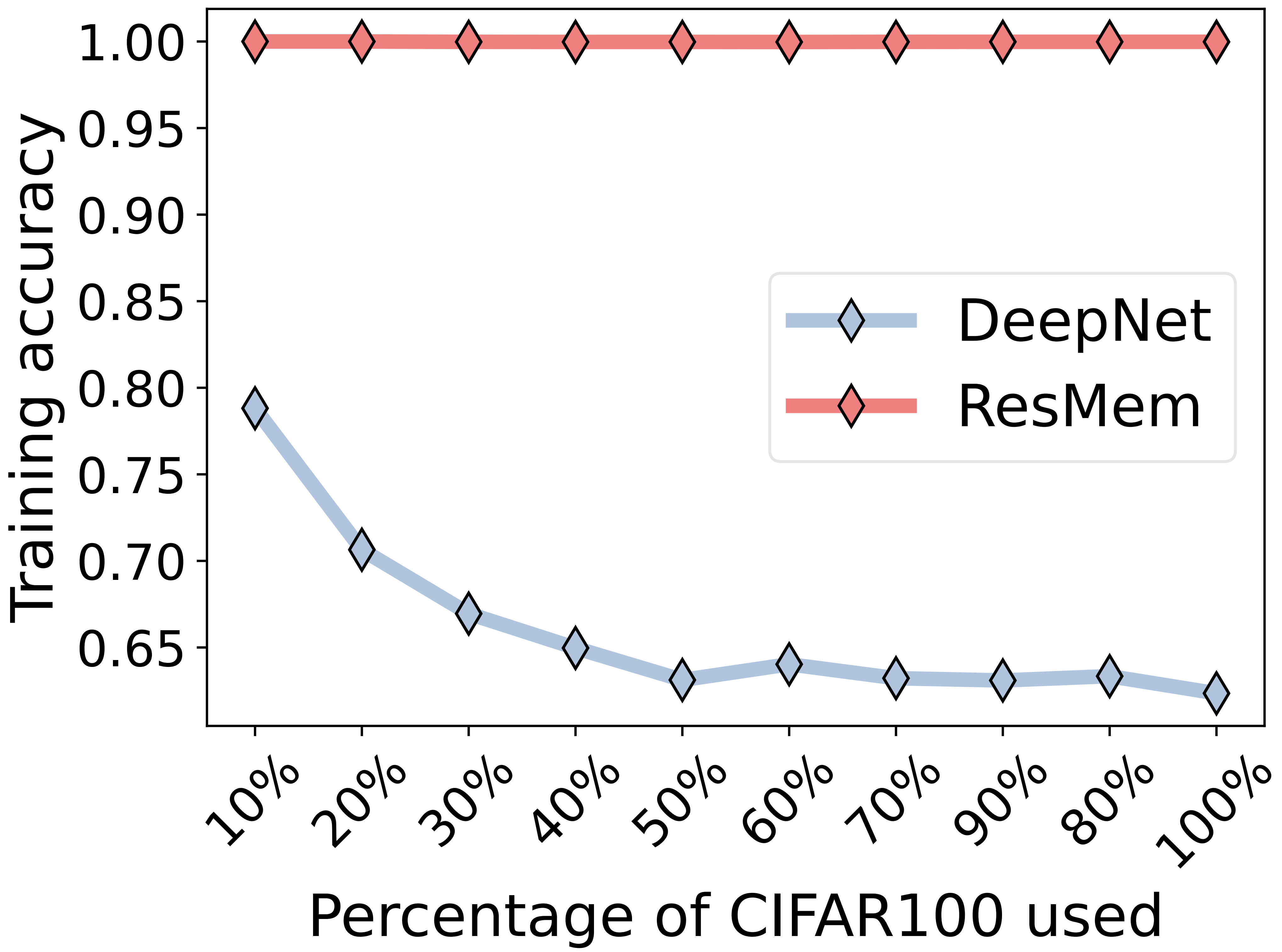
Empirical results: simple CIFAR experiment

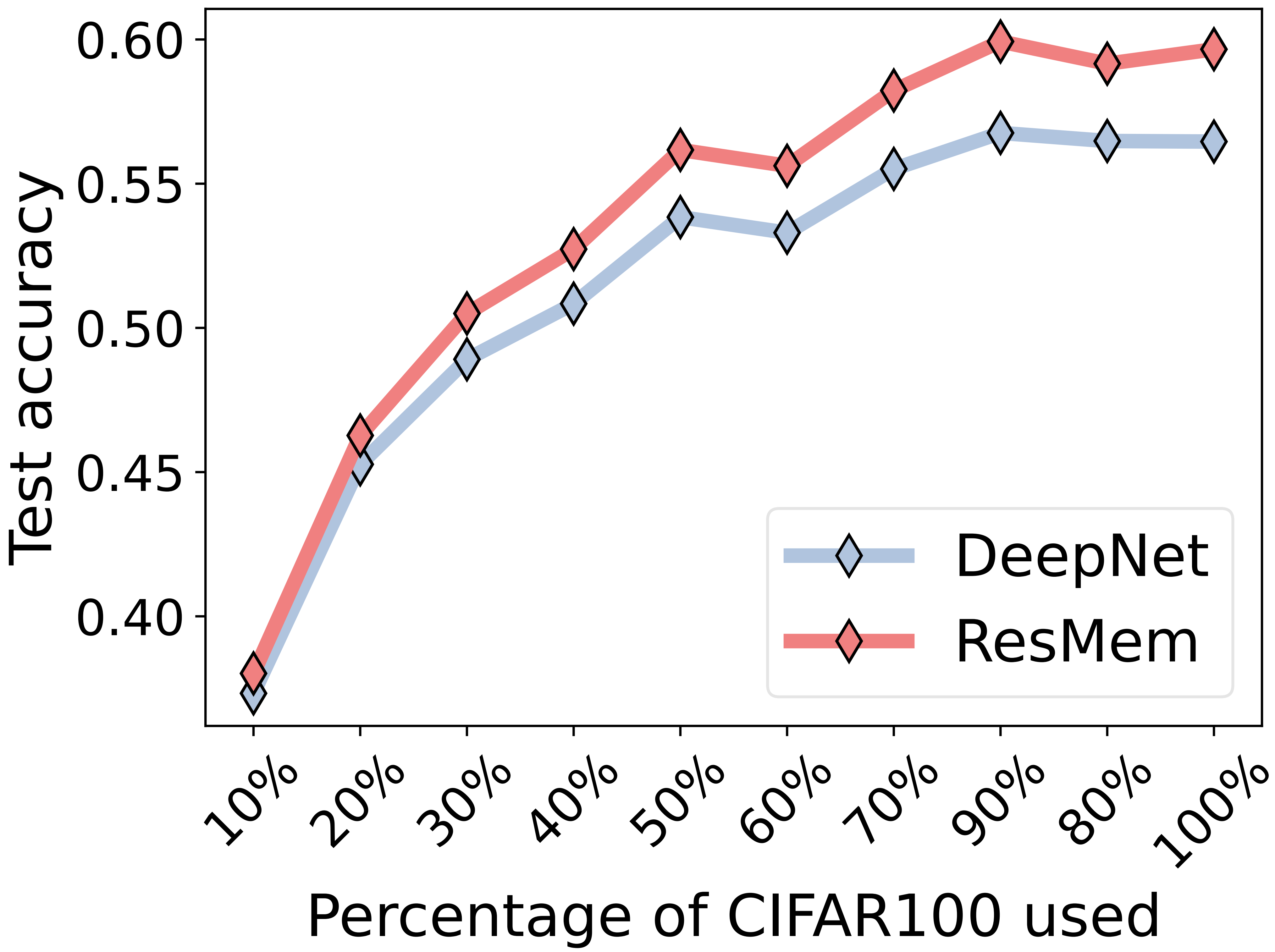
Dataset	Architecture	Test accuracy		
		DeepNet	ResMem	Gain
CIFAR100	ResNet-8	56.46%	**59.66%**	**3.20%**

When does ResMem help?



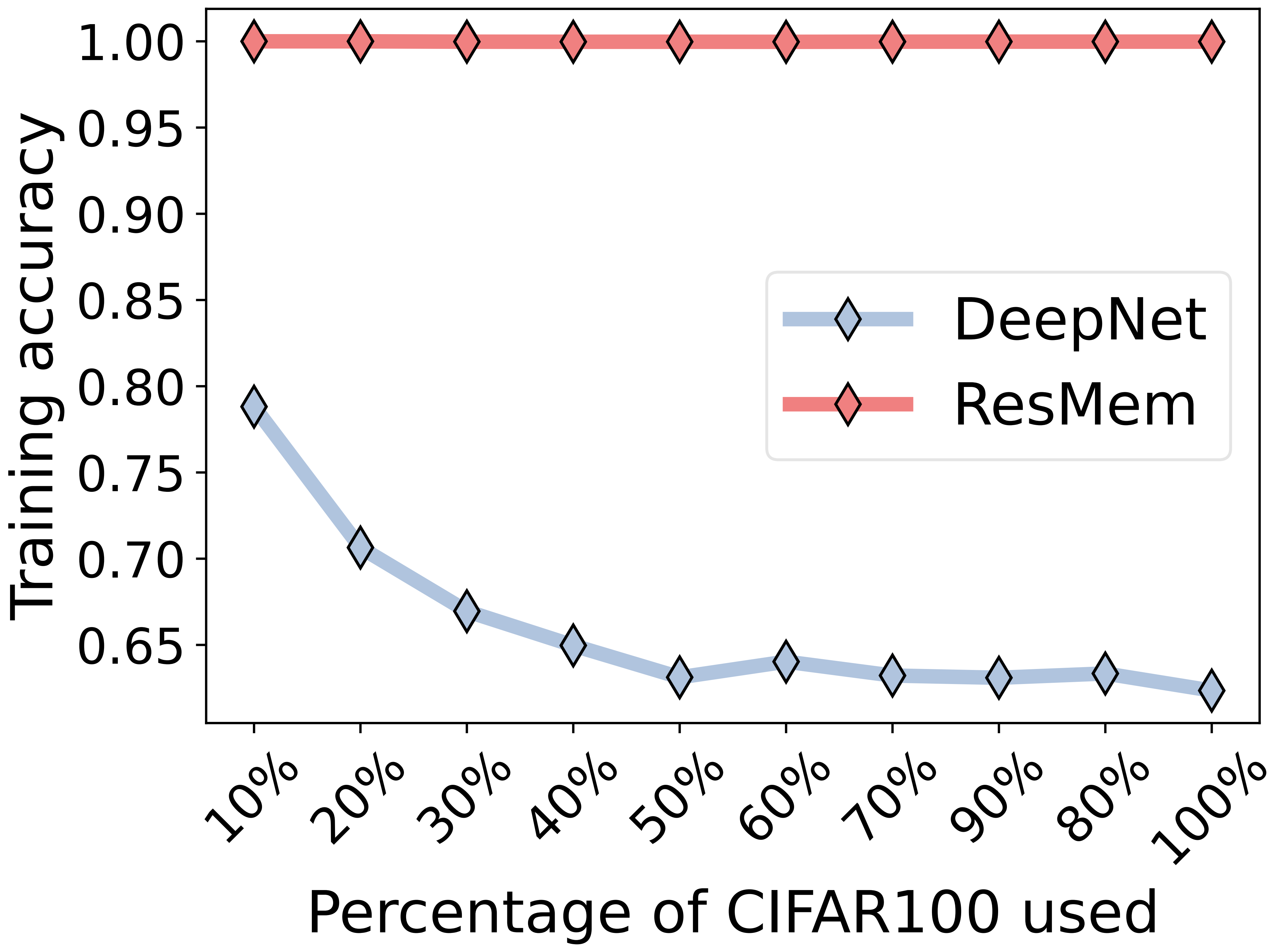
When does ResMem help?

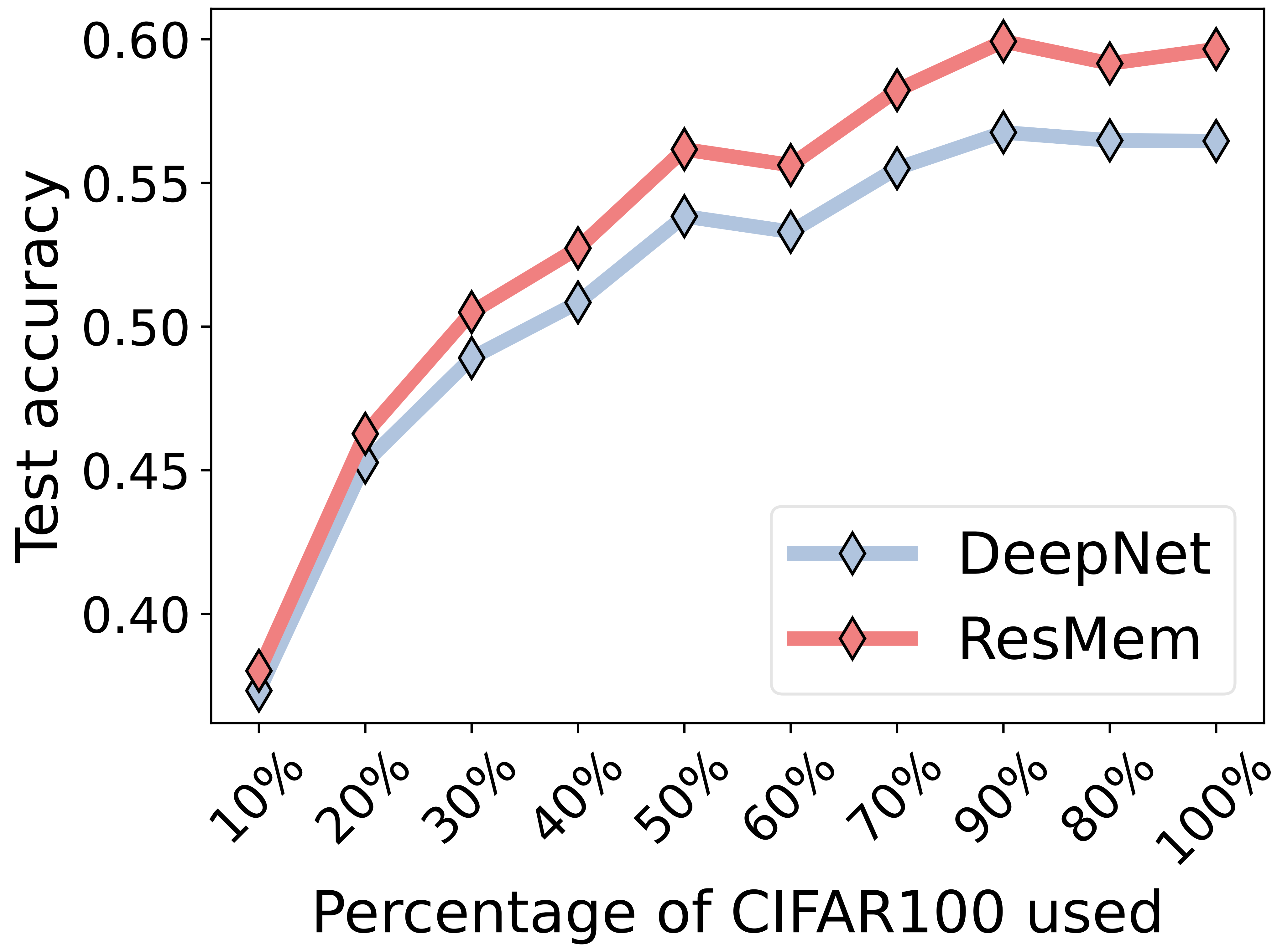




- When the **training sample is large**, ResMem is particularly effective.

When does ResMem help?

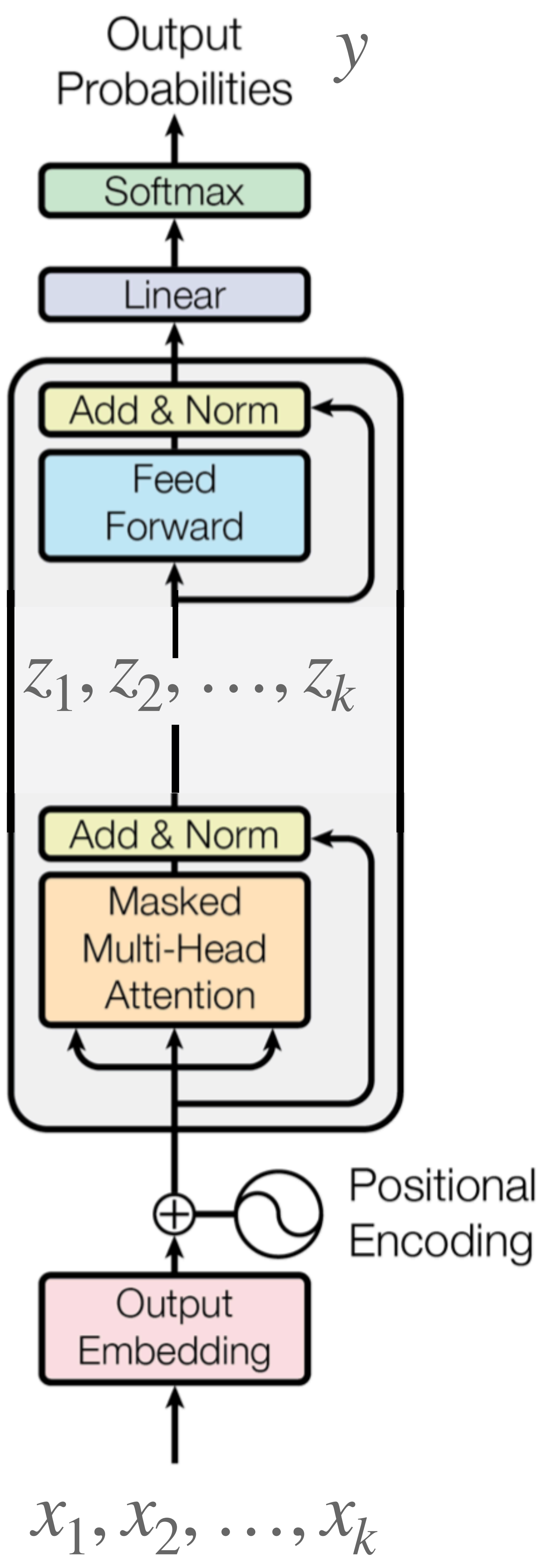




- When the **training sample is large**, ResMem is particularly effective.
- SOTA LLMs are typically trained for at most a single epoch. *(Google's PaLM '22)*

ResMem on languges models

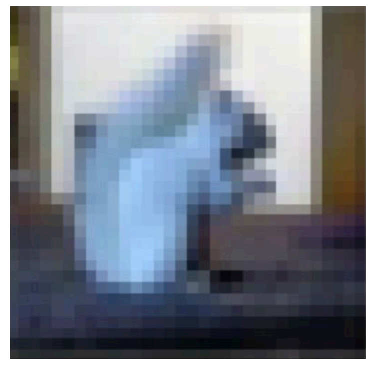
- Architecture: Decoder Only T5 *(Raffel et al. '20).*
- Dataset: C4 (text scrapped from internet).
- Residual: onehot$(x_{k+1}) - y$.
- Embedding: Post-attention, pre-feed forward z_k.
- Nearest neighbor: ScaNN *(Guo et al. '20)* for search over **1.6B** tokens.



Empirical results: an overview

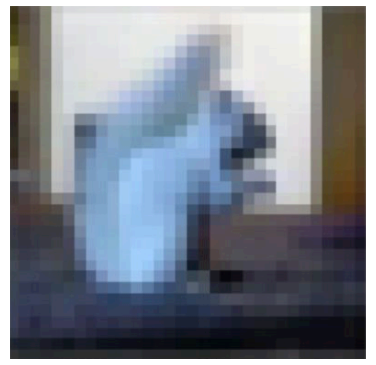
Dataset	Architecture	Test accuracy		
		DeepNet	ResMem	Gain
CIFAR100	ResNet-8	56.46%	**59.66%**	**3.20%**
C4	T5-Small	38.01%	**40.87%**	**2.86%**
C4	T5-Large	44.80%	**46.60%**	**1.80%**

Where does the improved accuracy come from?

	y^{ResMem}	y^{DeepNet}
	rose	poppy
	cup	plate
	squirrel	rabbit

	y^{ResMem}	y^{DeepNet}
...allow for plenty of headroom inside whilst still being less than 2.5m in height.	height	length
Graphic now consists of all cities with greater than 30,000 locals. Acquiring a home in Spain…	home	residence
Filmed around 7:30-8:30 a.m. on Friday, March 9, 2012.	March	June

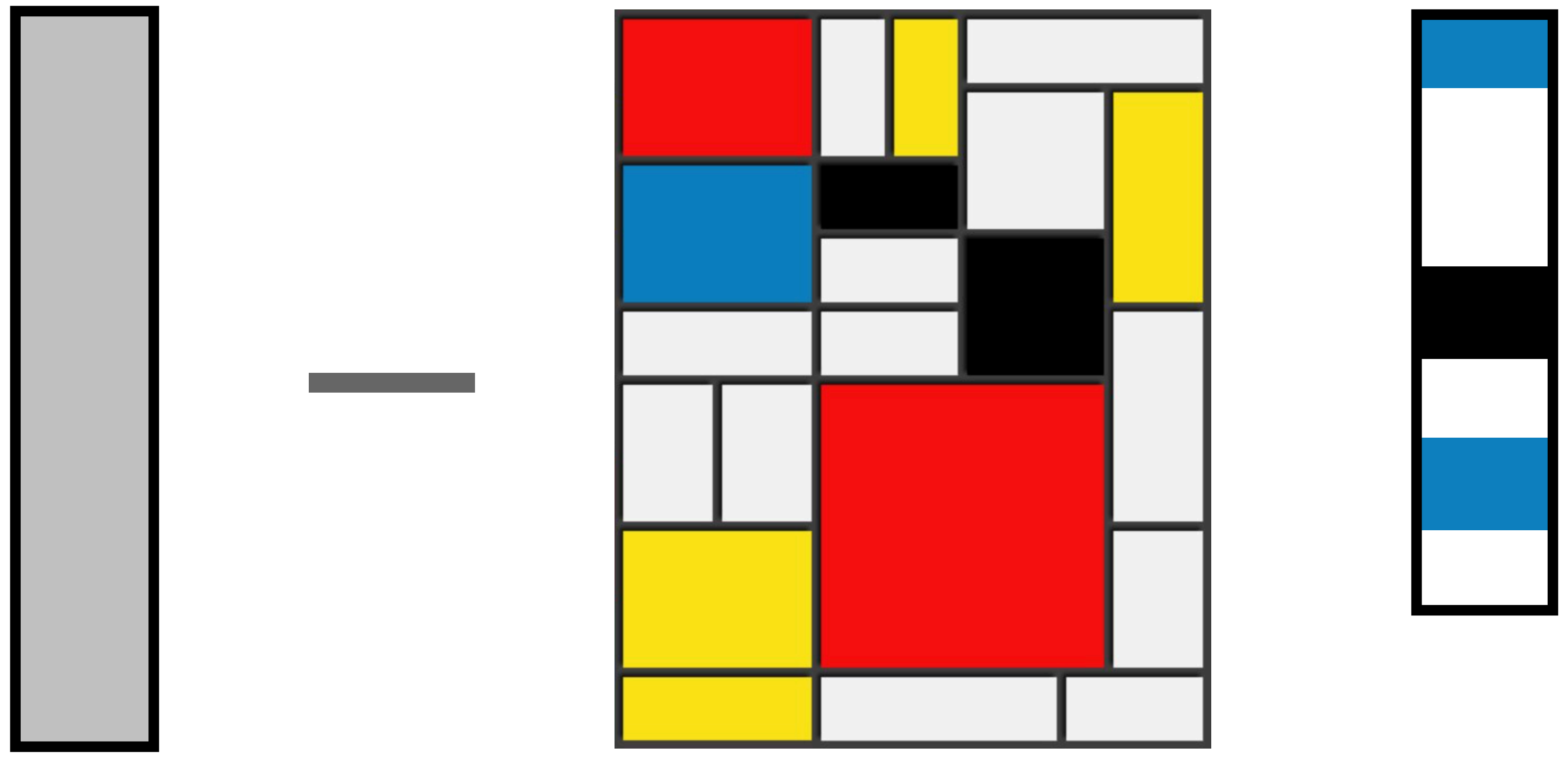
Where does the improved accuracy come from?

	y^{ResMem}	y^{DeepNet}
	rose	poppy
	cup	plate
	squirrel	rabbit

	y^{ResMem}	y^{DeepNet}
...allow for plenty of headroom inside whilst still being less than 2.5m in height.	height	length
Graphic now consists of all cities with greater than 30,000 locals. Acquiring a home in Spain…	home	residence
Filmed around 7:30-8:30 a.m. on Friday, March 9, 2012.	March	June

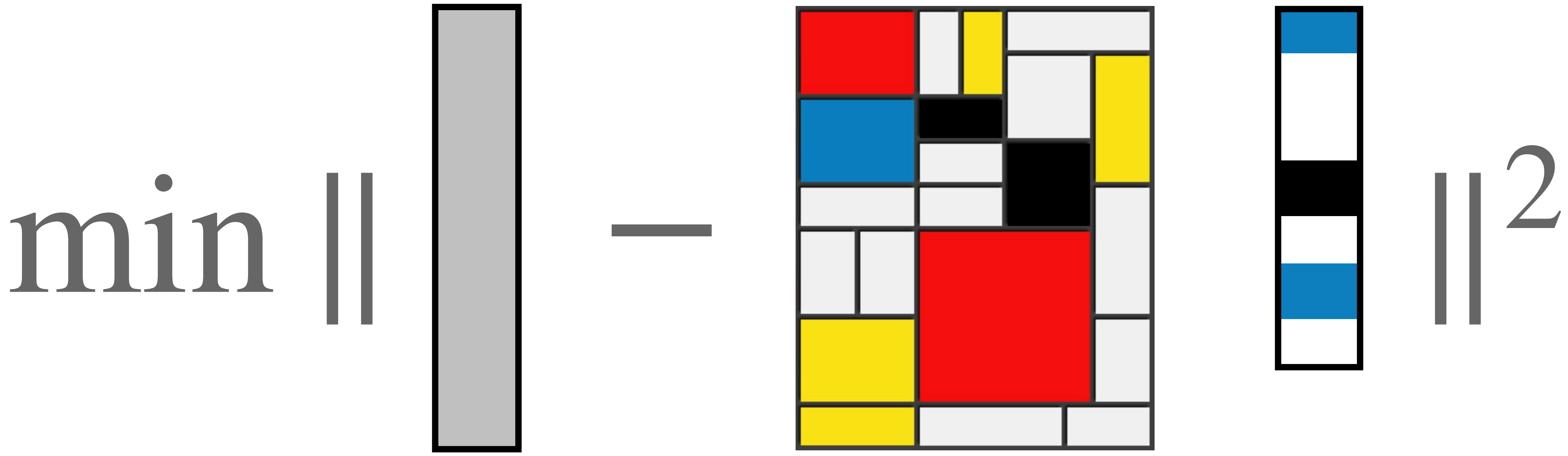
- y^{DeepNet} learns some **coarse structures.** (the flowers)
- y^{ResMem} memorizes the **fine-grained details.** (the roses)

Theoretical analysis: linear regression

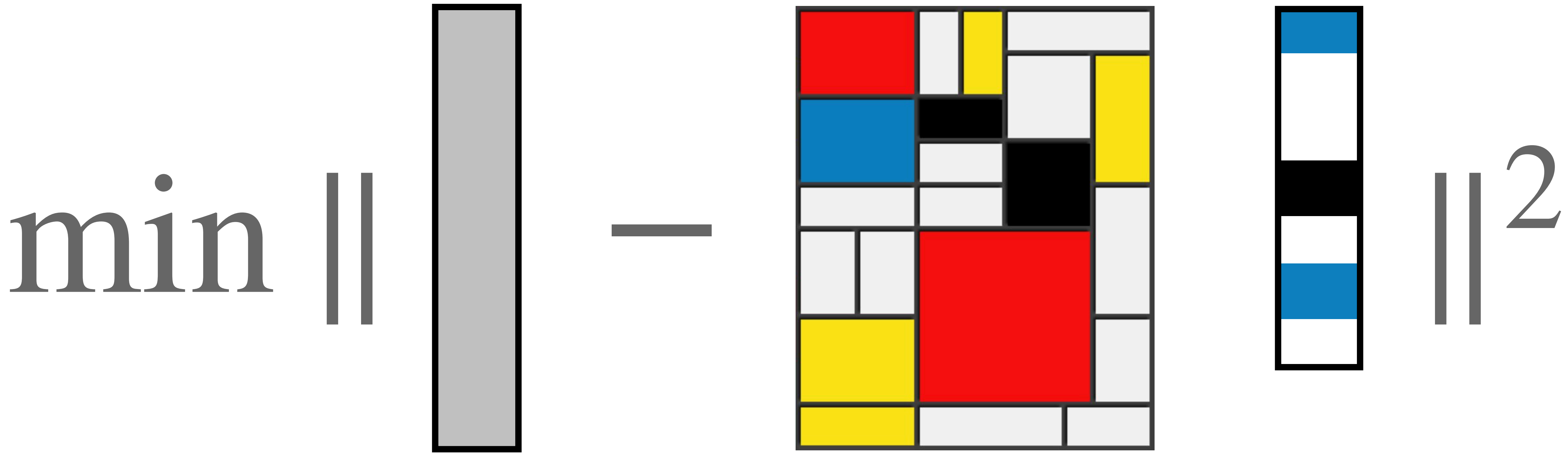
$\min \|$  $\|^2$

- Why don’t we memorize the **labels y directly?**

Theoretical analysis: assumptions



Theoretical analysis: assumptions

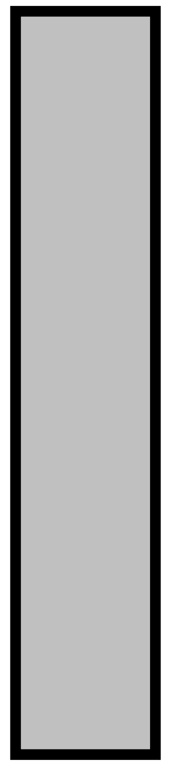
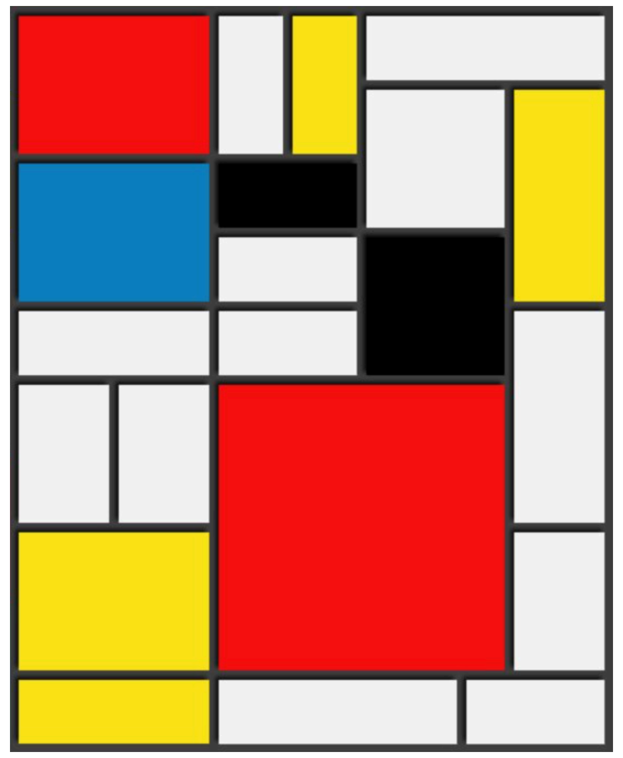
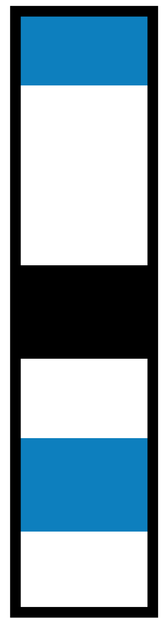


$X \in \mathbb{R}^{n \times d}$

- $X \in \mathbb{R}^{n \times d}$ is the usual design matrix with row $x_i \overset{i.i.d.}{\sim}$ some distribution.

Theoretical analysis: assumptions

y

$\min ||$  $-$   $||^2$

$X \in \mathbb{R}^{n \times d}$

- $X \in \mathbb{R}^{n \times d}$ is the usual design matrix with row $x_i \overset{i.i.d.}{\sim}$ some distribution.
- $y = X\theta_\star$ is generated by some true $\theta_\star$ with $||\theta_\star||$ = 1

Theoretical analysis: assumptions

$$\theta_n = \min_{\|\theta\| < L} \| \underbrace{y}_{} - \underbrace{X}_{X \in \mathbb{R}^{n \times d}} \underbrace{\theta}_{} \|^2$$

- $X \in \mathbb{R}^{n \times d}$ is the usual design matrix with row $x_i \overset{i.i.d.}{\sim}$ some distribution.
- $y = X\theta_\star$ is generated by some true $\theta_\star$ with $\|\theta_\star\| = 1$
- Empirical risk minimization performed over the functions class

$$\mathscr{F} = \{x \mapsto \langle x, \theta \rangle, \|\theta\| < L\}, L < 1$$

Theoretical analysis: results

Main theoretical result [Theorem 5.3, Yang et al., '23]

$$\text{Test Risk of ResMem} \lesssim \underbrace{d^2L^2n^{-2/3}}_{\text{parametric rate from linear regression}} + \overbrace{d^2(1-L)^2\left[\frac{\log\left(n^{1/d}\right)}{n}\right]^{1/d}}^{\text{stays as a irreducible error without ResMem}} \longrightarrow 0 \text{ as } n \to \infty$$

Theoretical analysis: results

Main theoretical result [Theorem 5.3, Yang et al., '23]

$$\text{Test Risk of ResMem} \lesssim \underbrace{d^2L^2n^{-2/3}}_{\text{parametric rate from linear regression}} + \overbrace{d^2(1-L)^2\left[\frac{\log\left(n^{1/d}\right)}{n}\right]^{1/d}}^{\text{stays as a irreducible error without ResMem}} \longrightarrow 0 \text{ as } n \to \infty$$

- The nearest neighbors component **expands the capacity** of the linear regressor by adding a non parametric component.
- If we memorize the labels y directly, we end up with a **slow rate of $n^{-1/d}$**.

Learn the flowers, remember the roses.