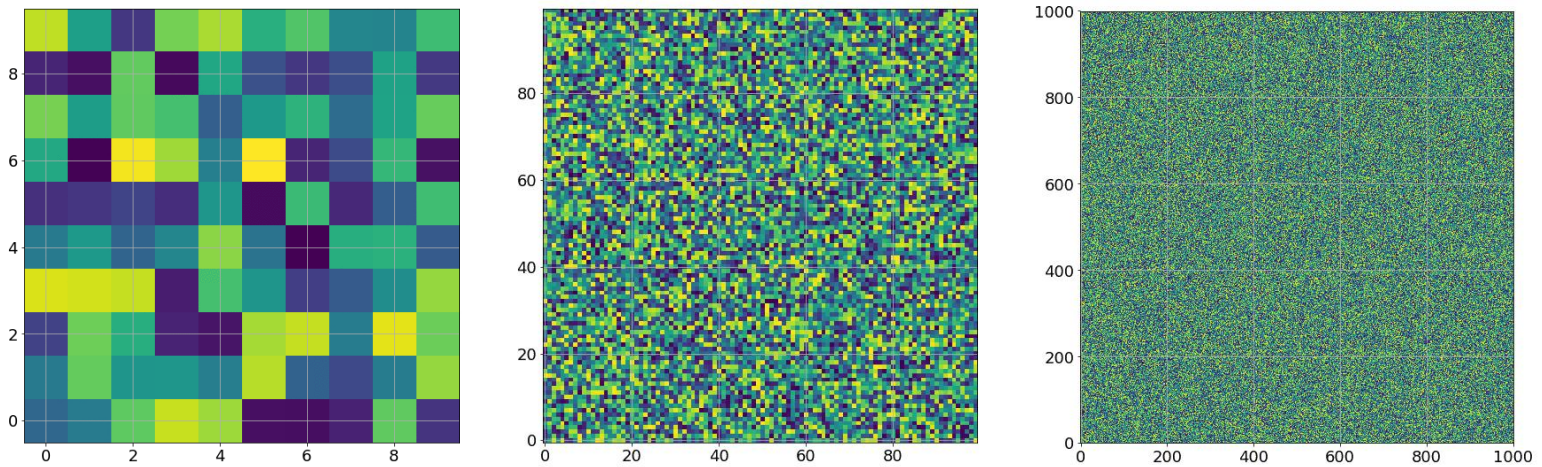
北京大學
PEKING UNIVERSITY

Task Arithmetic in the Tangent Space: Improved Editing of Pre-Trained Models

NeurIPS 2024

Neural Tangent Kernel



- 整个训练过程中网络存在一个不变量，不依赖于网络参数，就是NTK，神经正切核

$$K(\xi, \xi') = \langle \nabla_\theta f(\xi; \theta_0),\ \nabla_\theta f(\xi'; \theta_0) \rangle$$

$$K(x, x') = \langle \phi(x), \phi(x') \rangle$$

- 在无穷宽度条件下，宽神经网络由在初始参数处的一阶泰勒展开式线性模型主导

$$f(\xi; \theta_0 + \Delta\theta) \approx f(\xi; \theta_0) + \langle \nabla_\theta f(\xi; \theta_0), \Delta\theta \rangle$$

Task Arithmetic in the Tangent Space: Improved Editing of Pre-Trained Models

权重解耦

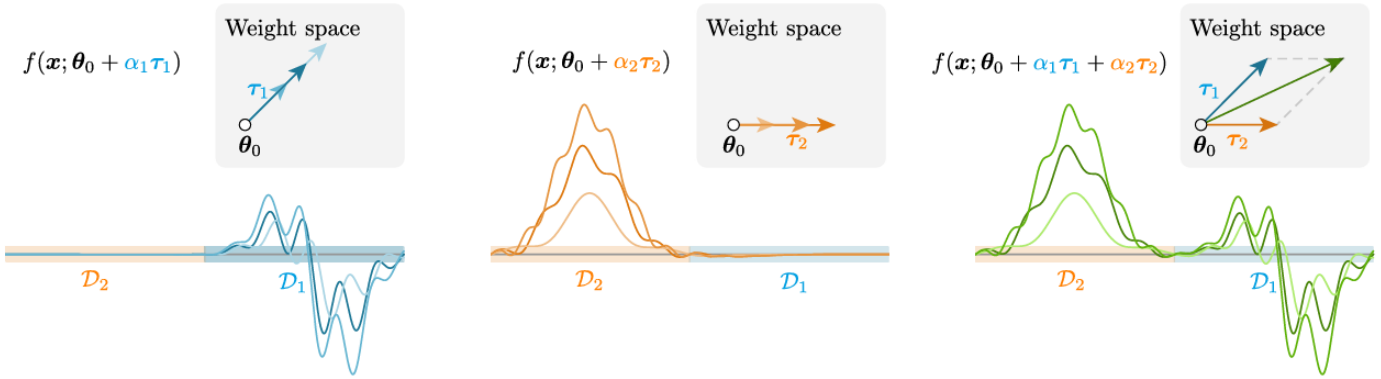


Figure 1: **Illustration of weight disentanglement**, where distinct directions in the weight space, τ_t, are associated with localized areas of the input space, $\mathcal{D}_t$. This allows a model, f, to manipulate these areas independently by adding linear combinations of τ_t's to a pre-trained checkpoint θ_0.

任务算术

Property 1 (Task arithmetic)**.** *Consider a set of task vectors* $\mathcal{T} = \{\tau_t\}_{t\in[T]}$ *with associated non-intersecting task supports* $\mathcal{D} = \{\mathcal{D}_t \subset \mathcal{X}\}_{t\in[T]}$*, i.e.,* $\forall t, t'$*, if* $t \neq t'$ *then* $\mathcal{D}_t \cap \mathcal{D}_{t'} = \varnothing$*. We say a network* f *satisfies the task arithmetic property around* θ_0 *with respect to* $\mathcal{T}$ *and* $\mathcal{D}$ *if*

$$f\left(x;\theta_0+\sum_{t=1}^{T}\alpha_t\,\tau_t\right)=\begin{cases}f(x;\theta_0+\alpha_t\,\tau_t) & x\in\mathcal{D}_t\\ f(x;\theta_0) & x\notin\bigcup_{t=1}^{T}\mathcal{D}_t\end{cases}\tag{1}$$

with $(\alpha_1,\ldots,\alpha_T)\in\mathcal{A}\subseteq\mathbb{R}^T$.

NTK

$$f(x;\theta)\approx f(x;\theta_0)+(\theta-\theta_0)^\top\nabla_\theta f(x;\theta_0).$$

$$k_{\mathrm{NTK}}(x,x')=\nabla_\theta f(x;\theta_0)^\top\nabla_\theta f(x';\theta_0)$$

检验假设：任务算术之所以可能，是因为模型内在地在一个线性区间（linear regime）内运行，其行为由NTK所支配

↓

预训练LM/VLM?

Task Arithmetic in the Tangent Space: Improved Editing of Pre-Trained Models

- 验证 CLIP 微调是否在线性空间内

Property 2 (Post-hoc linearization). *The change in the network output after training can be approximated by its first-order Taylor expansion, i.e.,* $f(\boldsymbol{x};\boldsymbol{\theta}^\star) - f(\boldsymbol{x};\boldsymbol{\theta}_0) \approx (\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0)^\top \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x};\boldsymbol{\theta}_0)$.

That is, we apply the fine-tuned task vectors $\boldsymbol{\tau} = \boldsymbol{\theta}^\star - \boldsymbol{\theta}_0$ to the linear approximation of f at $\boldsymbol{\theta}_0$, *i.e.*,

$$f_{\text{lin}}(\boldsymbol{x};\boldsymbol{\theta}_0 + \boldsymbol{\tau}) = f(\boldsymbol{x};\boldsymbol{\theta}_0) + \boldsymbol{\tau}^\top \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x};\boldsymbol{\theta}_0),$$

and we check whether $f_{\text{lin}}(\cdot\,;\boldsymbol{\theta}^\star)$ performs similarly to $f(\cdot\,;\boldsymbol{\theta}^\star)$[2].

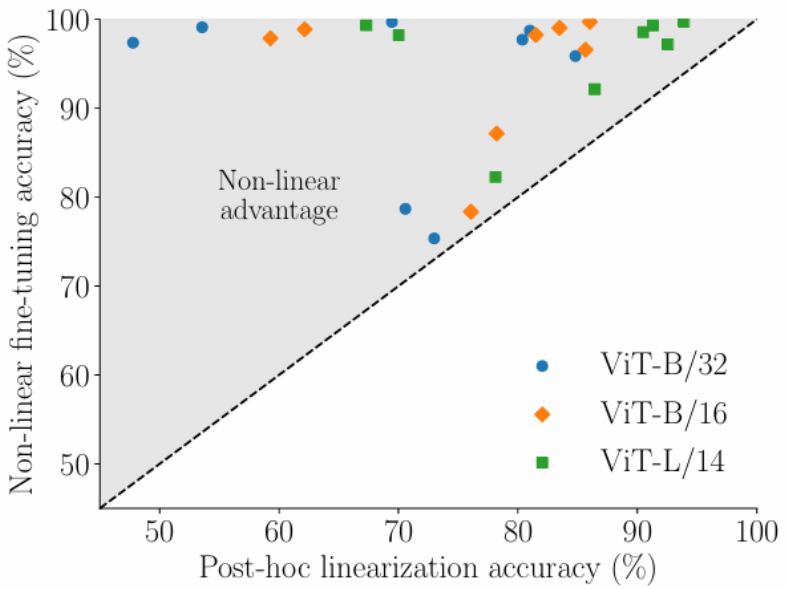


Figure 2: **Non-linear advantage.** Single-task accuracies of non-linearly fine-tuned models $f(\cdot\,;\boldsymbol{\theta}^\star)$ and their *post-hoc* linearization $f_{\text{lin}}(\cdot\,;\boldsymbol{\theta}^\star)$. Markers represent different ViTs.

- 微调并没有发生在线性区间中

Task Arithmetic in the Tangent Space: Improved Editing of Pre-Trained Models

北京大学 PEKING UNIVERSITY

- 任务算术是否只依赖于网络的线性化部分

Table 1: **Task addition.** Average absolute (%) and normalized accuracies (%) of different CLIP ViTs edited by adding the sum of the task vectors of 8 tasks. We report results for the non-linear and linearized models of Sections 3 and 5 normalizing performance by their single-task accuracies.

Method		ViT-B/32		ViT-B/16		ViT-L/14	
		Abs. (↑)	Norm. (↑)	Abs. (↑)	Norm. (↑)	Abs. (↑)	Norm. (↑)
Pre-trained	$f(\cdot\,;\boldsymbol{\theta}_0)$	48.4	–	55.2	–	64.4	–
Non-lin. FT	$f(\cdot\,;\boldsymbol{\theta}_0+\boldsymbol{\tau})$	71.4	76.5	75.5	80.0	85.1	88.8
Post-hoc lin.	$f_{\text{lin}}(\cdot\,;\boldsymbol{\theta}_0+\boldsymbol{\tau})$	57.1	81.9	65.0	85.2	75.2	90.0
Linear. FT	$f_{\text{lin}}(\cdot\,;\boldsymbol{\theta}_0+\boldsymbol{\tau}_{\text{lin}})$	**76.5**	**85.4**	**81.3**	**86.0**	**88.5**	**93.5**

Table 2: **Task negation.** Minimum accuracy (%) of different CLIP ViTs edited by negating a task vector from a target task while retaining 95% of their performance on the control task. We report average performances over eight tasks on non-linear and linearized models as introduced in Sections 3 and 5.

Method		ViT-B/32		ViT-B/16		ViT-L/14	
		Targ. (↓)	Cont. (↑)	Targ. (↓)	Cont. (↑)	Targ. (↓)	Cont. (↑)
Pre-trained	$f(\cdot\,;\boldsymbol{\theta}_0)$	48.4	63.4	55.2	68.3	64.4	75.5
Non-lin. FT	$f(\cdot\,;\boldsymbol{\theta}_0-\boldsymbol{\tau})$	24.0	60.7	19.2	64.6	18.0	**72.5**
Post-hoc lin.	$f_{\text{lin}}(\cdot\,;\boldsymbol{\theta}_0-\boldsymbol{\tau})$	14.8	60.3	**10.8**	**64.8**	12.1	71.8
Linear. FT	$f_{\text{lin}}(\cdot\,;\boldsymbol{\theta}_0-\boldsymbol{\tau}_{\text{lin}})$	**10.9**	**60.8**	11.3	**64.8**	**7.9**	**72.5**

权重解耦

- 论点：用一个模型来执行任务算术的唯一必要条件，是该模型相对于微调任务集合而言是权重解耦的

Property 3 (Weight disentanglement). *A parametric function $f : \mathcal{X} \times \Theta \to \mathcal{Y}$ is weight disentangled with respect to a set of task vectors $\mathcal{T} = \{\boldsymbol{\tau}_t\}_{t\in[T]}$ and the corresponding supports $\mathcal{D} = \{\mathcal{D}_t\}_{t\in[T]}$ if*

$$f\left(\boldsymbol{x}; \boldsymbol{\theta}_0 + \sum_{t=1}^{T} \alpha_t \boldsymbol{\tau}_t\right) = \sum_{t=1}^{T} g_t(\boldsymbol{x}; \alpha_t \boldsymbol{\tau}_t) + g_0(\boldsymbol{x}), \tag{4}$$

where $g_t(\boldsymbol{x}; \alpha_t \boldsymbol{\tau}_t) = \mathbf{0}$ for $\boldsymbol{x} \notin \mathcal{D}_t$ and $t = 1, \ldots, T$, and $g_0(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in \bigcup_{t\in[T]} \mathcal{D}_t$.

$$f\left(\boldsymbol{x}; \boldsymbol{\theta}_0 + \sum_{t=1}^{T} \alpha_t \boldsymbol{\tau}_t\right) = \begin{cases} f(\boldsymbol{x}; \boldsymbol{\theta}_0 + \alpha_t \boldsymbol{\tau}_t) & \boldsymbol{x} \in \mathcal{D}_t \\ f(\boldsymbol{x}; \boldsymbol{\theta}_0) & \boldsymbol{x} \notin \bigcup_{t=1}^{T} \mathcal{D}_t \end{cases} \tag{1}$$

$$f\left(\boldsymbol{x}; \boldsymbol{\theta}_0 + \sum_{t=1}^{T} \alpha_t \boldsymbol{\tau}_t\right) = \sum_{t=1}^{T} f(\boldsymbol{x}; \boldsymbol{\theta}_0 + \alpha_t \boldsymbol{\tau}_t) \mathbb{1}(\boldsymbol{x} \in \mathcal{D}_t) + f(\boldsymbol{x}; \boldsymbol{\theta}_0) \mathbb{1}\left(\boldsymbol{x} \notin \bigcup_{t\in[T]} \mathcal{D}_t\right), \tag{5}$$

and identify $g_t(\boldsymbol{x}; \alpha_t \boldsymbol{\tau}_t) = f(\boldsymbol{x}; \boldsymbol{\theta}_0 + \alpha_t \boldsymbol{\tau}_t) \mathbb{1}(\boldsymbol{x} \in \mathcal{D}_t)$ and $g_0(\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{\theta}_0) \mathbb{1}(\boldsymbol{x} \notin \mathcal{D}_t)$. It is

权重解耦

- disentanglement error

$$\xi(\alpha_1, \alpha_2) = \sum_{t=1}^{2} \mathbb{E}_{\boldsymbol{x} \sim \mu_t} \left[\text{dist} \left(f(\boldsymbol{x}; \boldsymbol{\theta}_0 + \alpha_t \boldsymbol{\tau}_t), f(\boldsymbol{x}; \boldsymbol{\theta}_0 + \alpha_1 \boldsymbol{\tau}_1 + \alpha_2 \boldsymbol{\tau}_2) \right) \right],$$

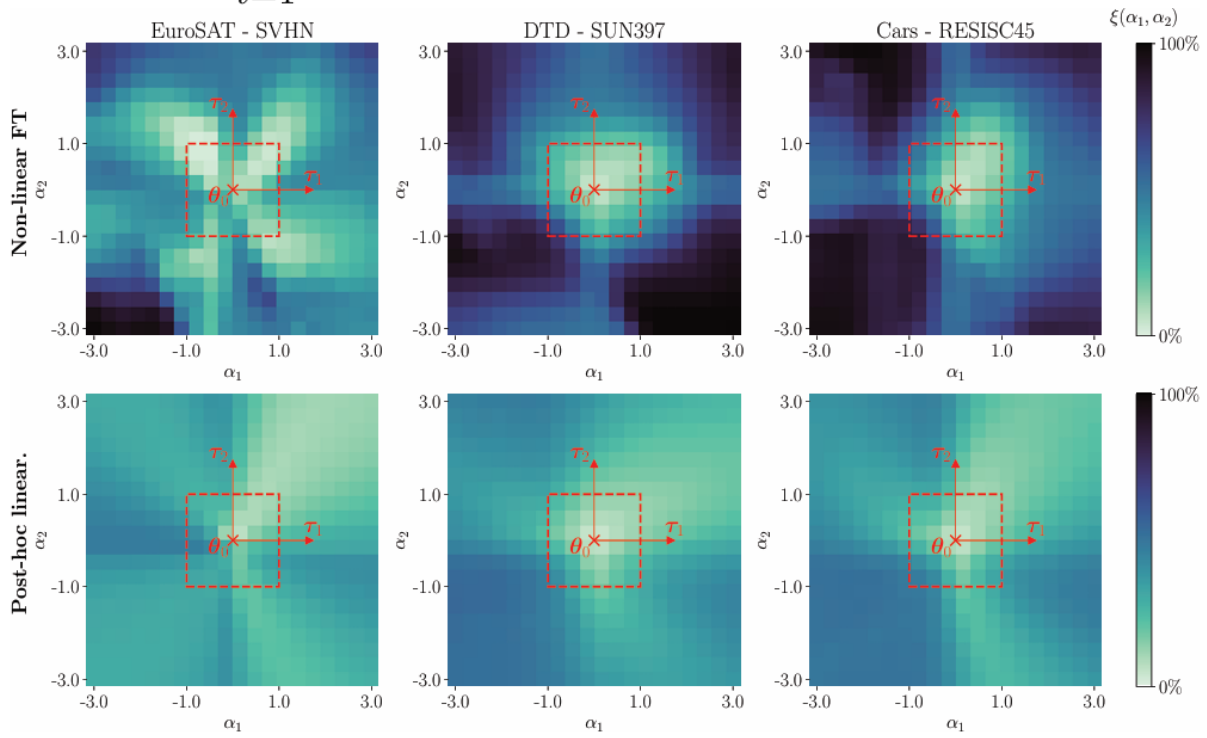


Figure 3: **Visualization of weight disentanglement.** The heatmaps show the disentanglement error $\xi(\alpha_1, \alpha_2)$ of a non-linear CLIP ViT-B/32 (top) and its post-hoc linearization (bottom) on different example task pairs. The light regions denote areas of the weight space where weight disentanglement is stronger. The red box delimits the search space used to compute the best α in all our experiments.

- 初始化权重附近的disentanglement error 很小
- 线性化模型表现出比其非线性更强的解耦

Task Arithmetic in the Tangent Space: Improved Editing of Pre-Trained Models

Linearization FT

- 在切空间中对每个任务进行显式微调

$$f_{\text{lin}}(x) = f(x;\theta_0) + J_\theta f(x;\theta_0) \cdot \Delta\theta$$

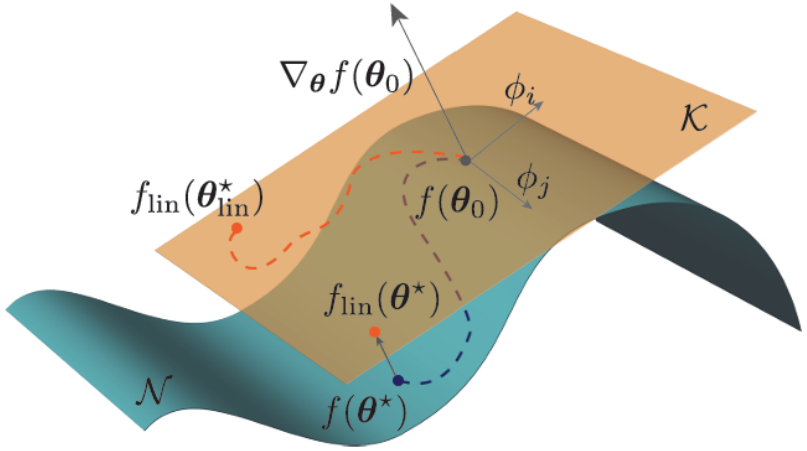


Figure 4: Conceptual illustration of the different approaches we use to edit a pretrained model $f(\cdot\,;\theta_0)$. Here $\mathcal{N}$ represents the space of neural network functions f, non-linearly parameterized by $\theta \in \Theta$; and $\mathcal{K}$ its tangent space, given by the space of linearized functions f_{lin}.

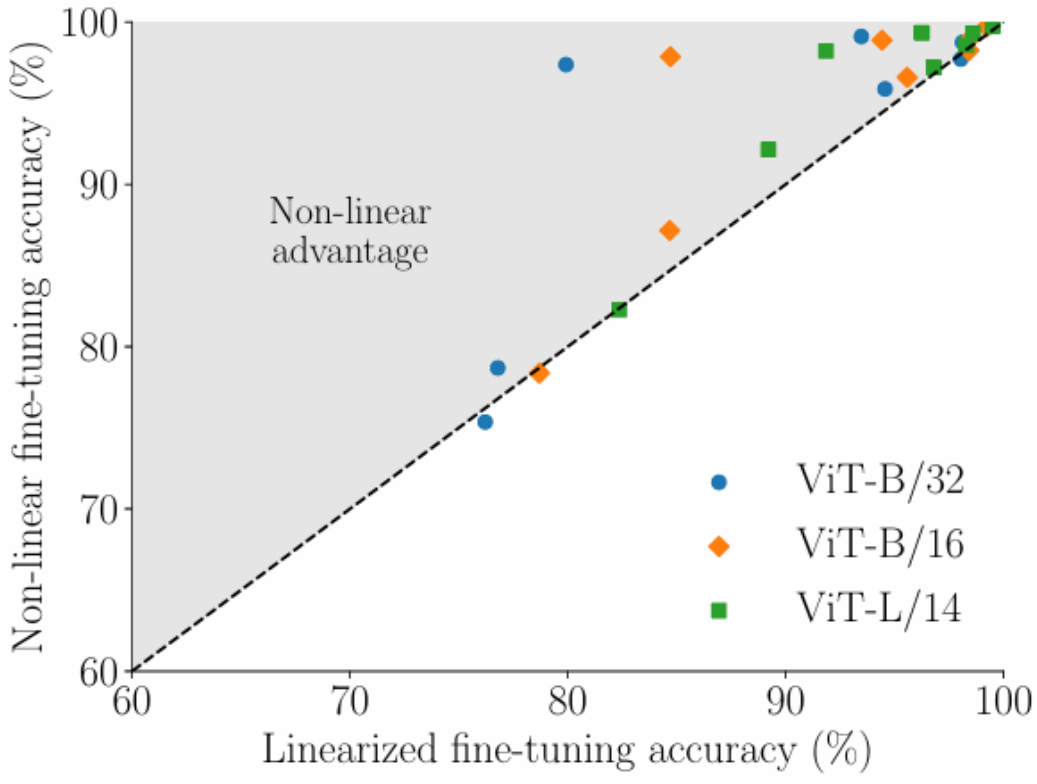


Figure 5: Single-task accuracies of non-linearly FT, $f(\cdot\,;\theta^\star)$ and linearly FT, $f_{\text{lin}}(\cdot\,;\theta^\star_{\text{lin}})$, models.

Linearization FT

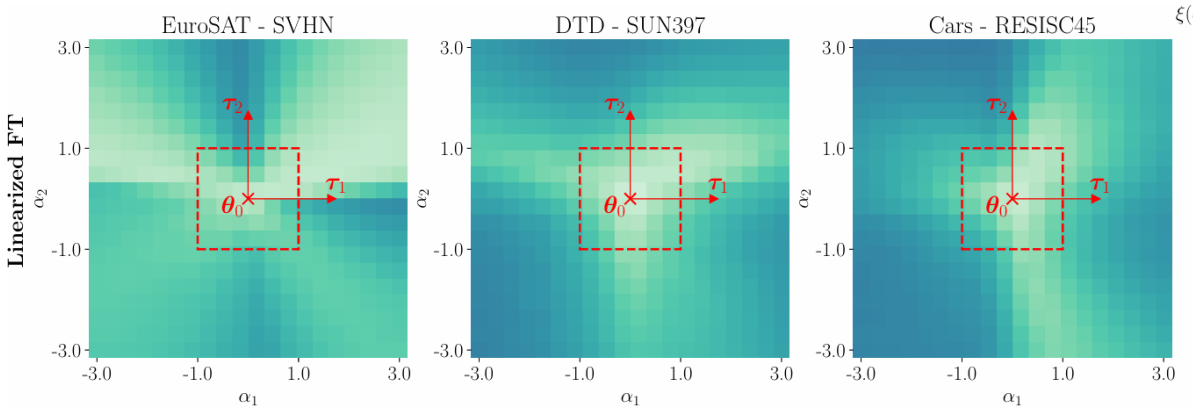


Figure 10: **Visualization of weight disentanglement from linearized models.** The heatmaps the disentanglement error $\xi(\alpha_1, \alpha_2)$ of a ViT-B/32 linearly fine-tuned on different example task The light regions denote areas of the weight space where weight disentanglement is stronger. T box delimits the search space used to compute α in our experiments.

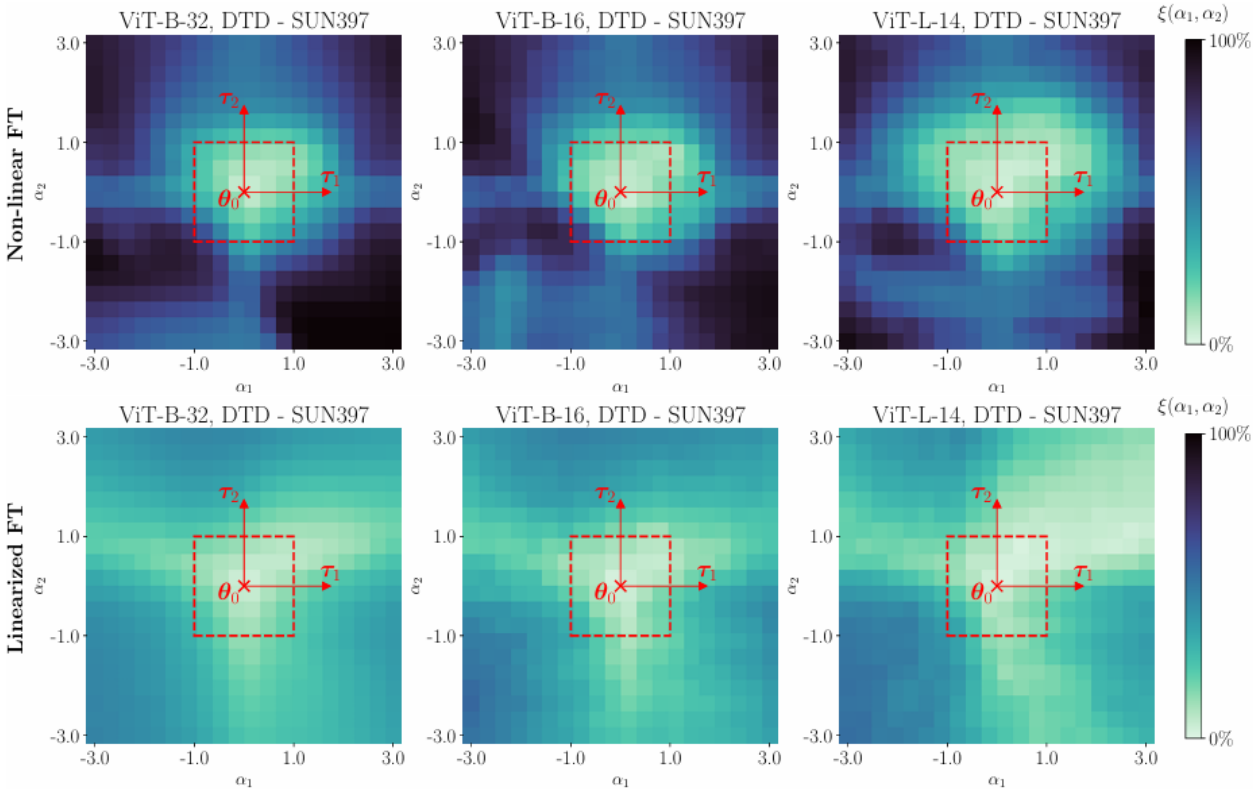


Figure 9: **Weight disentanglement and model scale.** The heatmaps show the disentanglement error $\xi(\alpha_1, \alpha_2)$ of different non-linear CLIP ViTs (top) and their post-hoc linearizations (bottom) on DTD and SUN397. The light regions denote areas of the weight space where weight disentanglement is stronger. The red box delimits the search space used to compute the best α in all our experiments.

Eigenfunction localization

- 将核视作特征函数-特征值对簇

Proposition 1 (Simplified). *Suppose that* $\{f_t^\star\}_{t\in[T]}$ *can be represented by the kernel* k*. The kernel* k *is capable of performing task arithmetic with respect to* $\{f_t^\star\}_{t\in[T]}$ *and* $\{\mathcal{D}_t\}_{t\in[T]}$ *if, for each task* t*, there exists a subset of localized eigenfunctions such that i)* $\mathrm{supp}(\phi) \subseteq \mathcal{D}_t$ *for each* ϕ *in the subset, and ii) the representation of* $f_t^\star$ *only involves these basis functions.*

- 模型训练的核表示为 $f_{\mathrm{lin}}(\boldsymbol{x}) = f(\boldsymbol{x};\boldsymbol{\theta}_0) + \sum_{\nu\in[n_t]} \beta_\nu\, k_{\mathrm{NTK}}(\boldsymbol{x}_\nu, \boldsymbol{x})$

- 对矩阵进行对角化 $(K_{\mathrm{NTK}})_{ij} = k_{\mathrm{NTK}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ with $\boldsymbol{x}_i \in \mathcal{D}_t$,

  $\boldsymbol{x}_j \in \mathcal{D}_t \cup \mathcal{D}_{t'}$, where $\mathcal{D}_{t'}$ is the support of a control task

- 局部能量计算 $\mathcal{E}_{\mathrm{loc}}(\boldsymbol{x}) = \sum_\rho \phi_\rho^2(\boldsymbol{x})$

Eigenfunction localization

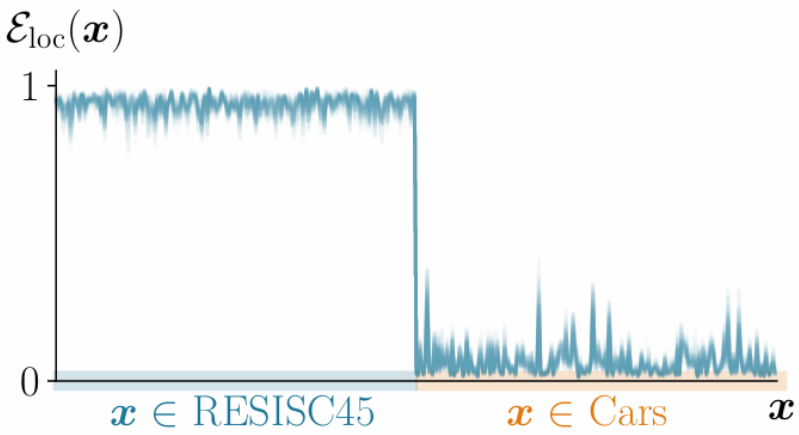


Figure 6: **Eigenfunction localization.** Estimated support of the eigenfunctions of the NTK of a ViT-B/32 CLIP model trained on RESISC45. The plot shows the sum of the local energy of the eigenfunctions over a random subset of the training and control supports (RESISC45 and Cars, respectively).

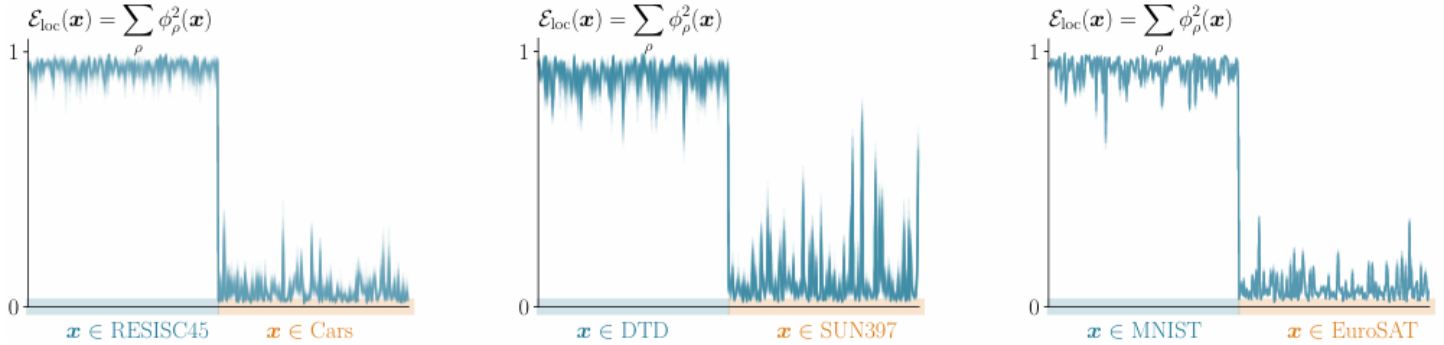


Figure 14: **Eigenfunction localization.** Estimated support of the eigenfunctions of the NTK of a ViT-B/32 CLIP model trained on different datasets. The plot shows the sum of the local energy of the eigenfunctions over a random subset of the training and control supports

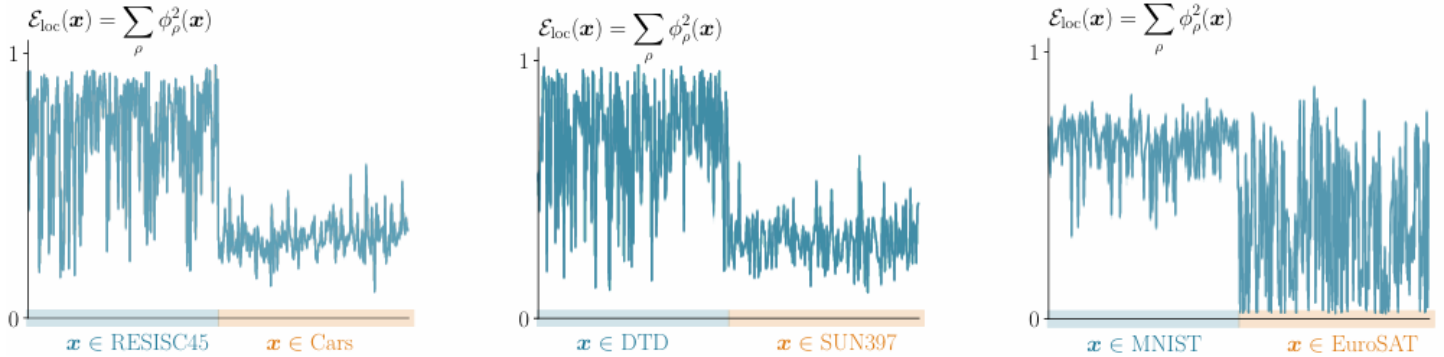


Figure 15: **Eigenfunction localization.** Estimated support of the eigenfunctions of the NTK of a randomly initialized ViT-B/32 model trained on different datasets. The plot shows the sum of the local energy of the eigenfunctions over a random subset of the training and control supports

Weight disentanglement emerges during pre-training

Table 3: **Task addition from random initialization.** We use the same setup as for the experiments in Table 1 but with task vectors obtained from fine-tuning randomly initialized ViTs. Results compare the average single-task accuracy (%) after fine-tuning and the multi-task accuracy (%) via task addition.

Method		ViT-B/32		ViT-B/16		ViT-L/14	
		Sing. (↑)	Multi (↑)	Sing. (↑)	Multi (↑)	Sing. (↑)	Multi (↑)
Random init	$f(\cdot\,;\boldsymbol{\theta}_0^{\text{rd}})$	5.3	–	4.8	–	5.2	–
Non-lin. FT	$f(\cdot\,;\boldsymbol{\theta}_0^{\text{rd}}+\boldsymbol{\tau}^{\text{rd}})$	48.5	5.5	40.6	4.5	18.0	4.8
Linear. FT	$f_{\text{lin}}(\cdot\,;\boldsymbol{\theta}_0^{\text{rd}}+\boldsymbol{\tau}_{\text{lin}}^{\text{rd}})$	27.8	3.8	24.7	4.0	24.8	6.1