Universität St.Gallen

ICML International Conference On Machine Learning

WeightCLIP: Aligning Datasets and Models for Weight Space Learning

Aron Asefaw, Konstantinos Tzevelekakis, Damian Falk, Léo Meynent, Damian Borth
University of St. Gallen (Switzerland)

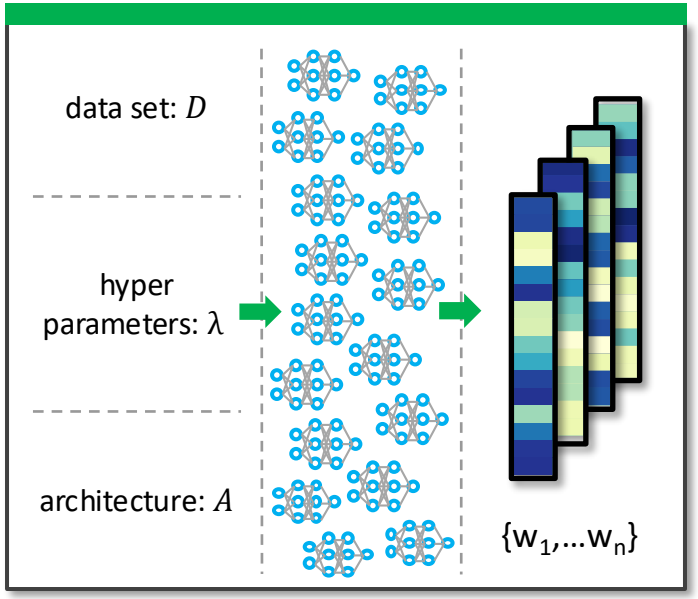
International Conference On Machine Learning, ICML 2026
Seoul, Korea

Weight Space Learning

Instead of learning from **images** or **text**, we learn from the **neural network parameters** themselves

Weight Space Learning

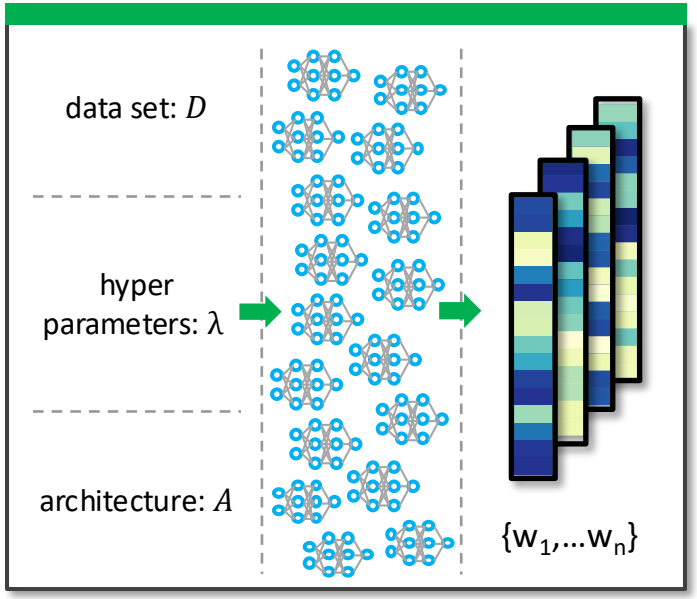
Instead of learning from **images** or **text**, we learn from the **neural network parameters** themselves



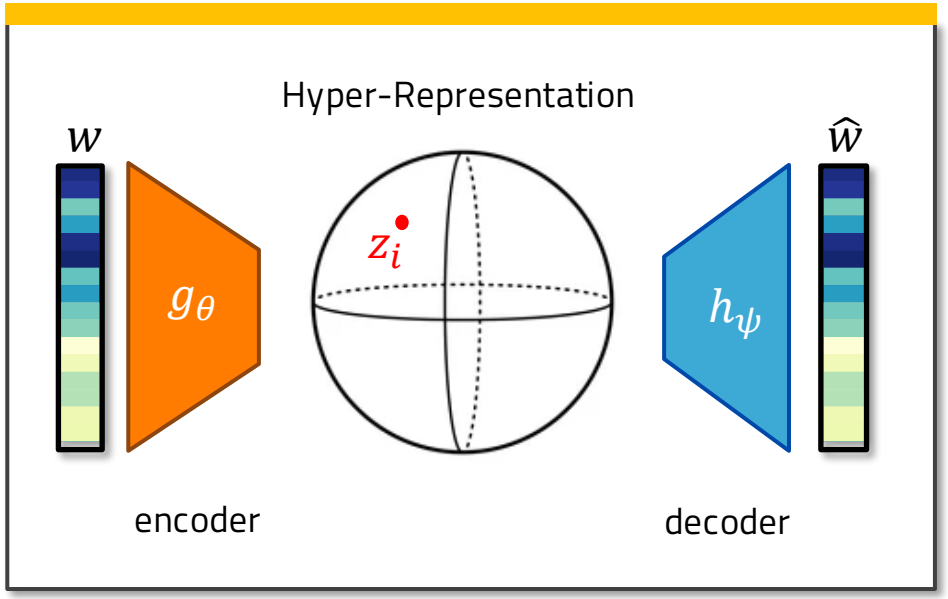
(I) Model Zoos

Weight Space Learning

Instead of learning from **images** or **text**, we learn from the **neural network parameters** themselves



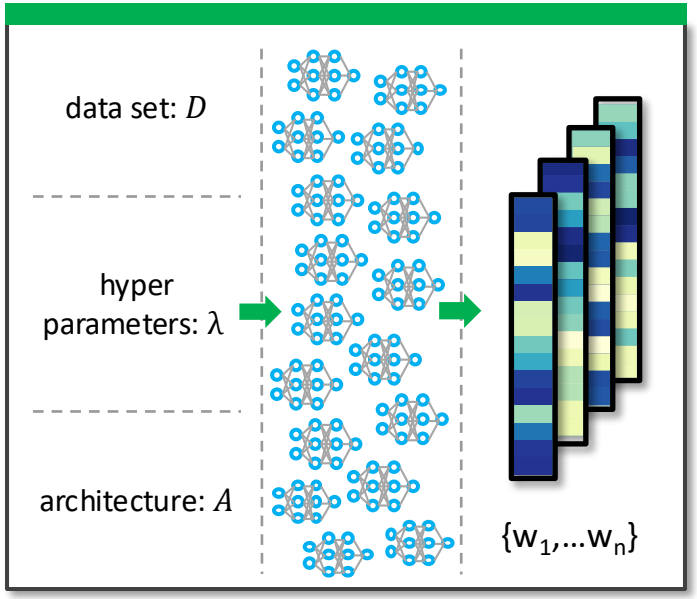
(I) Model Zoos



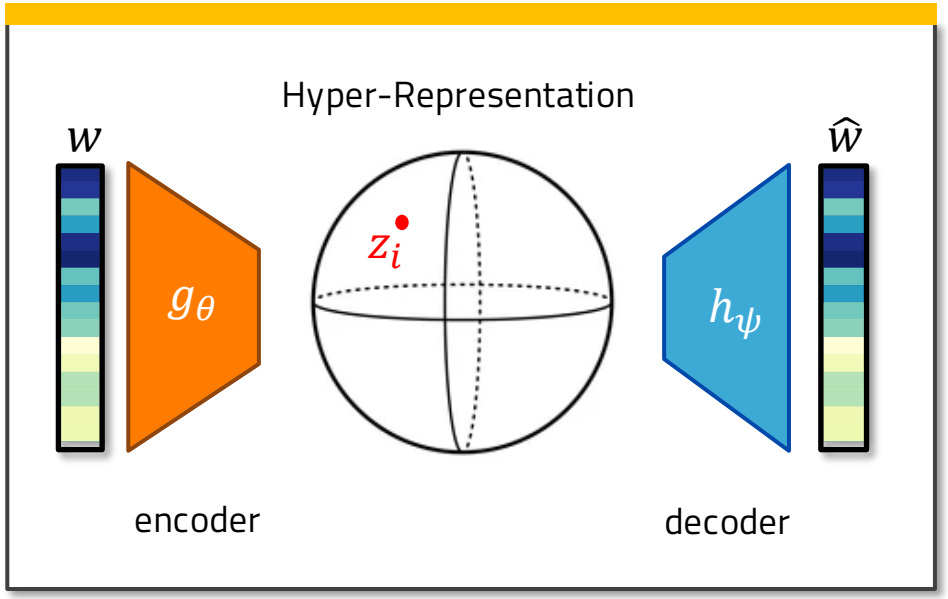
(II) Weight Space Learning

Weight Space Learning

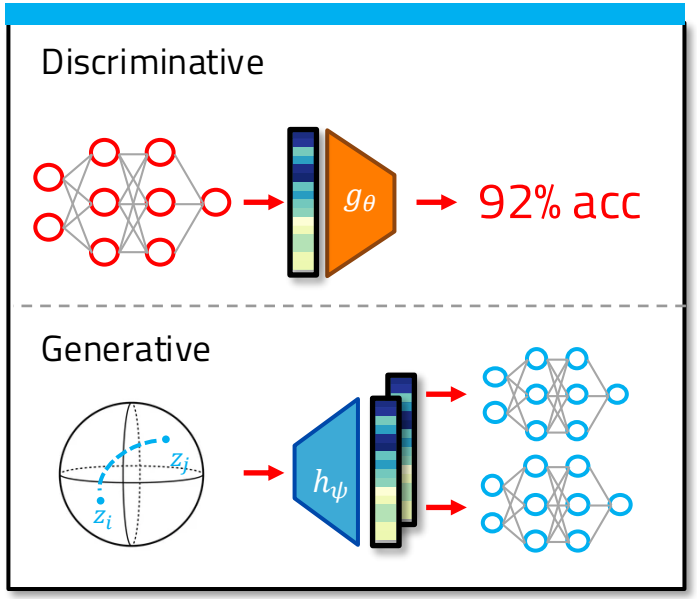
Instead of learning from **images** or **text**, we learn from the **neural network parameters** themselves



(I) Model Zoos



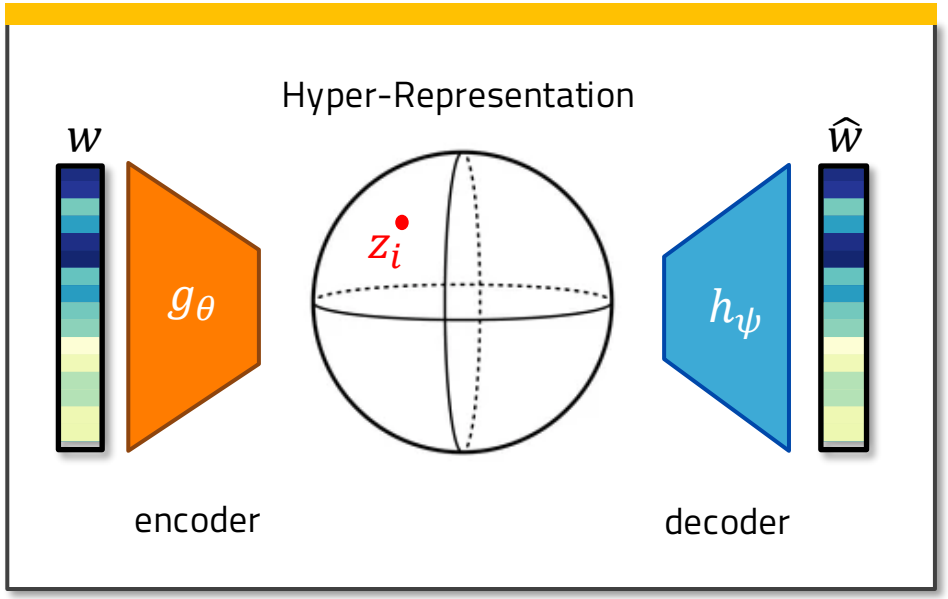
(II) Weight Space Learning



(III) Down-stream Tasks

Weight Space Learning

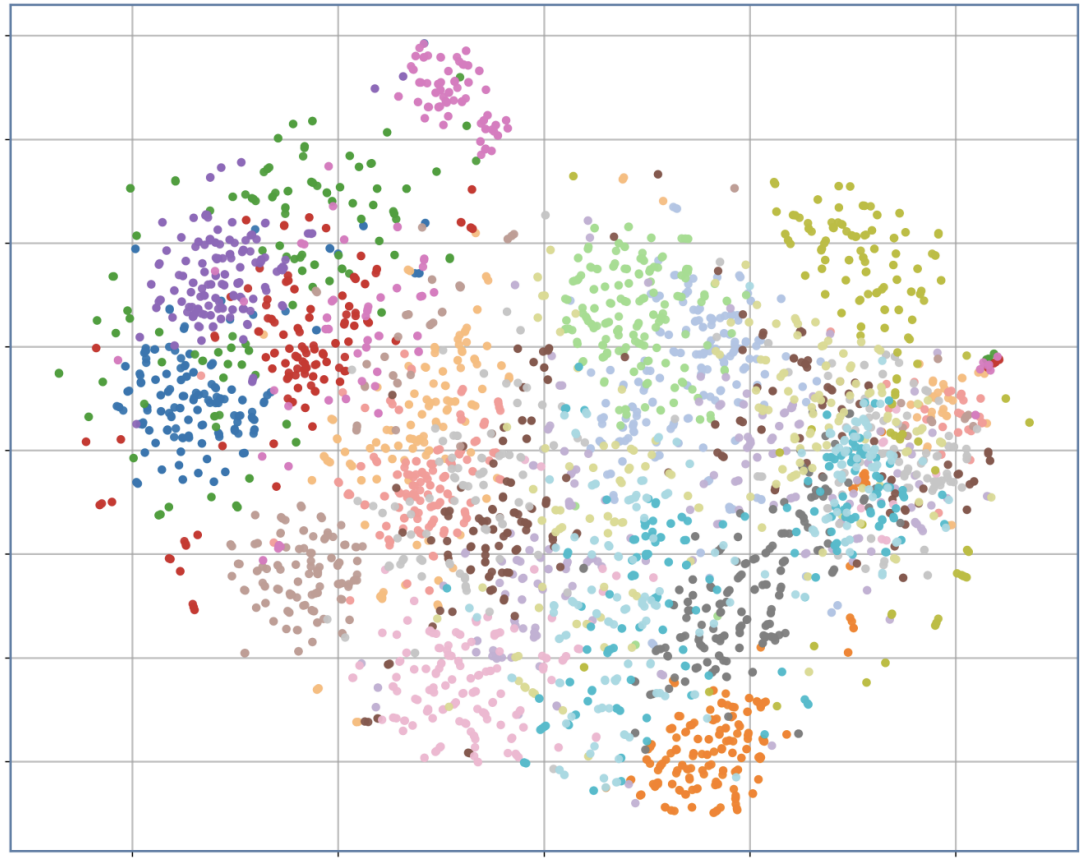
Instead of learning from **images** or **text**, we learn from the **neural network parameters** themselves



(II) Weight Space Learning

The Missing Link

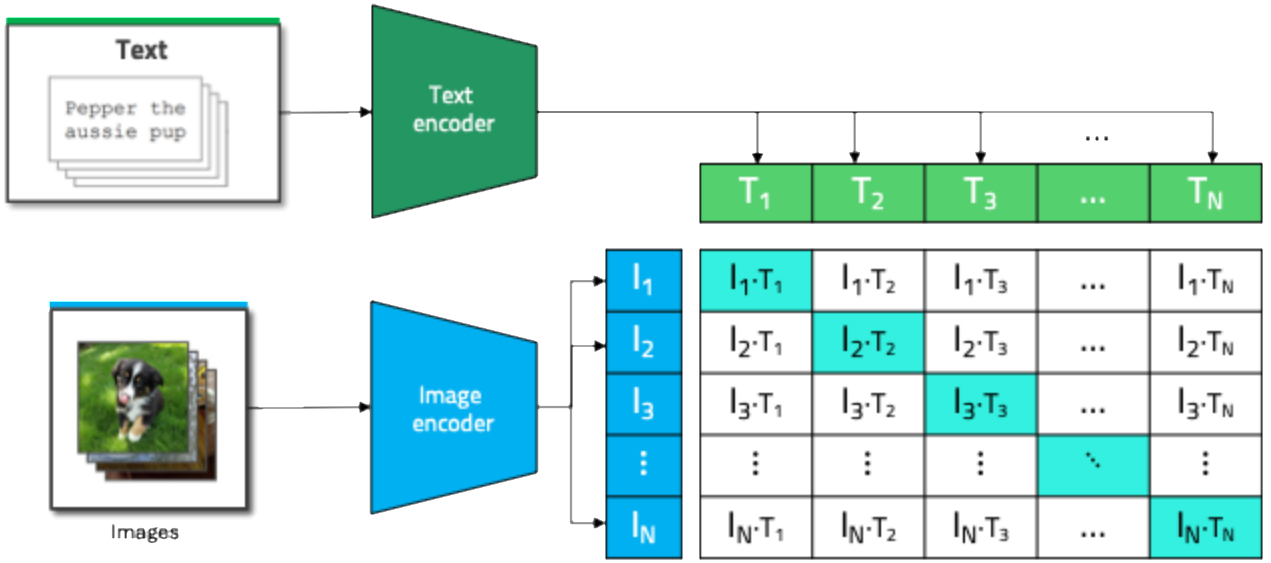
- Reconstruction alone does not impose **dataset semantics.**
- Models trained on similar datasets are not guaranteed to be close.
- This makes the space hard to **prompt** or **navigate** in the **learned latent space**.
- **Goal:** Organize the latent space around **datasets**



t-SNE plot of the model latent space without any dataset alignment

Proposed solution: WeightCLIP

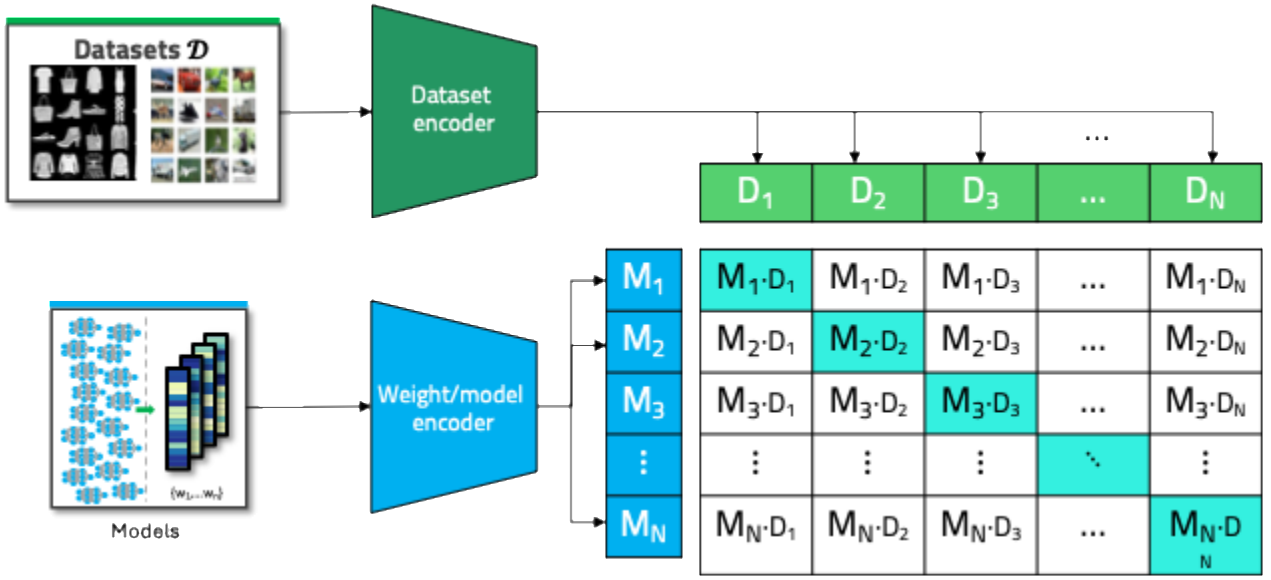
- CLIP takes **two different modalities** and aligns them in a **shared latent space.**
- The shared latent space becomes aligned for **prompting.**



Proposed solution: WeightCLIP

- We move this analogy to Weight Space Learning. Instead of **image–text** pairs, we use **model–dataset** pairs as the alignment signal.

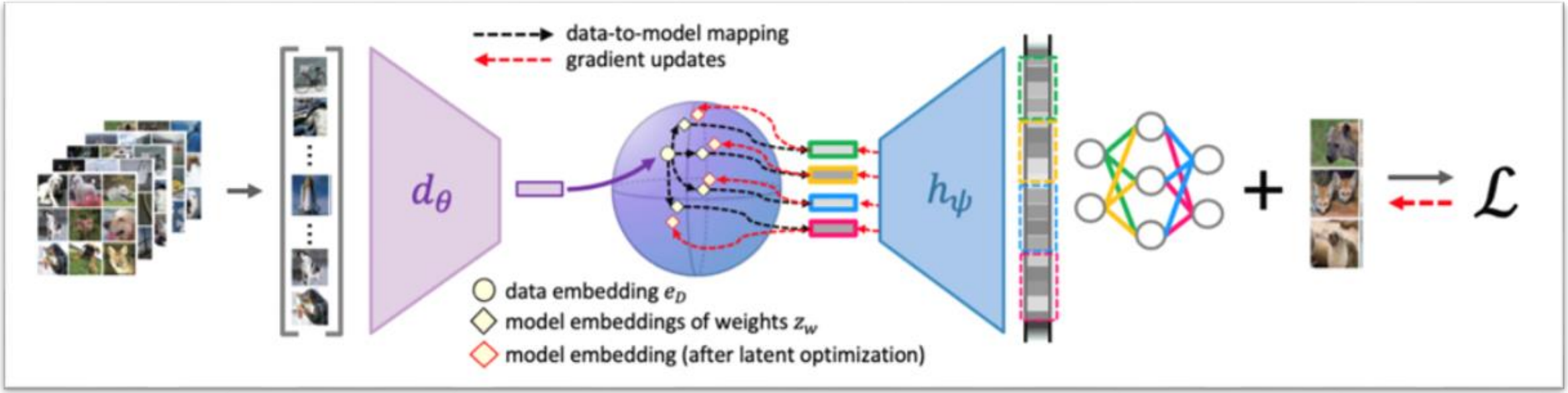
- Using a **dataset encoder** and a **weight encoder** we **contrastively align** dataset and model embeddings.



Proposed Solution - Generation

Given a **new target dataset**:

1. Encode a **few examples** as a **dataset prompt**
2. **Map** the **dataset embedding** to **model latent vectors**
3. Decode the latent vectors into **neural network weights**



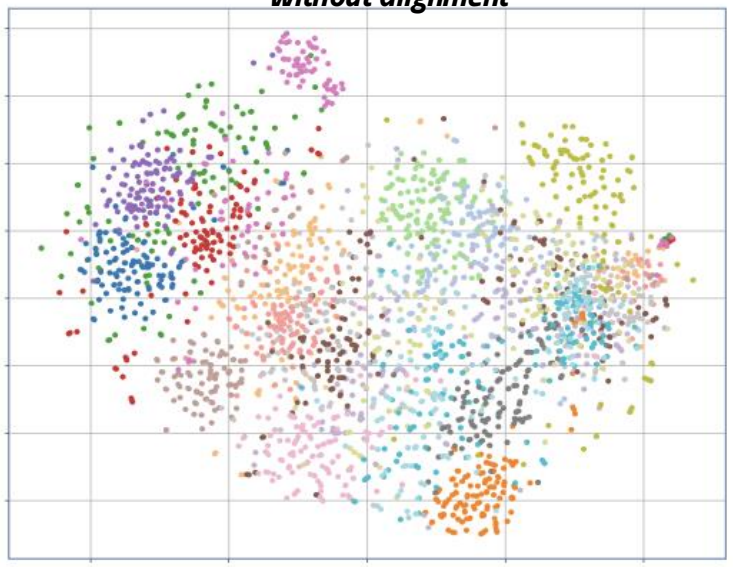
The dataset becomes a prompt for generating a model

Universität St.Gallen

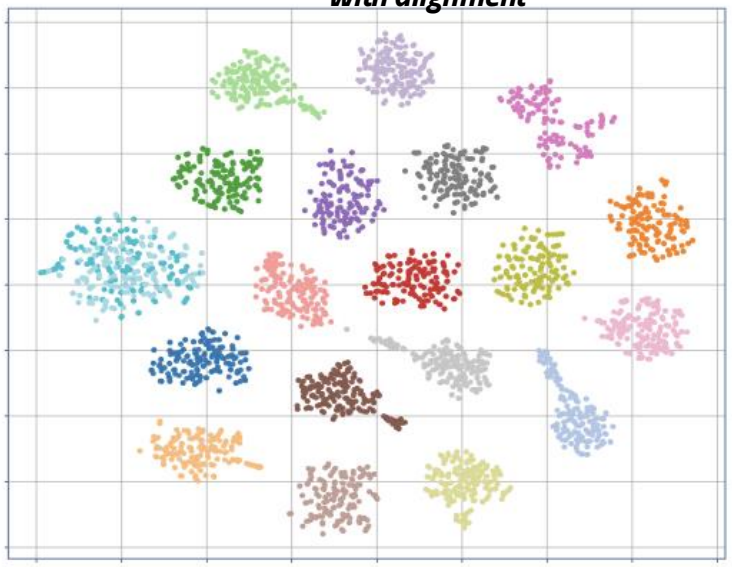
Main Results

Qualitatively there is a clear improvement in the **clustering** of the models based on training data.

t-SNE plot of the model latent space ***without alignment***

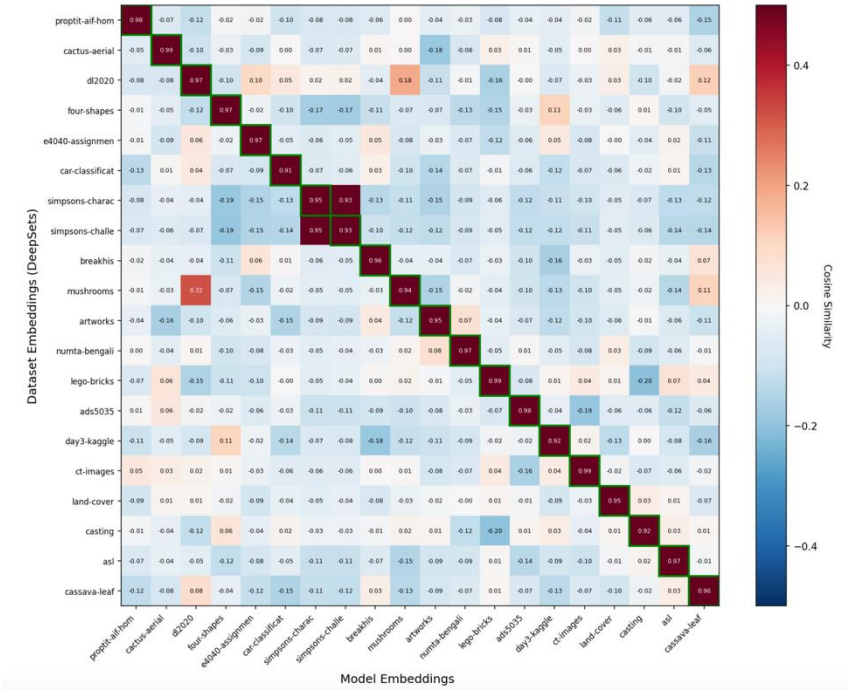


t-SNE plot of the model latent space ***with alignment***



The **dataset-model similarity matrix** also tells the same story.

Dataset-model similarity matric



Universität St.Gallen

Main Results

- Is this space now useful for generating new models based on OOD data prompts?
- **WeightCLIP** outperforms scratch, retrieval, and direct generation baselines.
- Strongest gains after short adaptation.
- Works beyond retrieval: generated models outperform nearest neighbors.

OOD model generation from dataset prompts. Accuracy after 0, 1, and 10 fine-tuning epochs.

Ep.	Method	Colo.H	COVID-19	CIFAR10	Speed.	Honey P.	Real/Draw.
0	tr. fr. scratch	∼12.5	∼33.3	∼10.0	∼25.0	∼50.0	∼10.0
	TANS	8.5±6.5	44.9±0.9	8.6±1.8	30.6±2.8	51.9±1.6	10.3±1.3
	Hypernetwork	12.4±0.0	35.6±0.0	9.5±0.0	13.9±0.0	**59.7±0.0**	12.6±0.0
	WeightCLIP-LM	13.7±0.3	**47.4±0.4**	10.8±0.1	36.7±3.2	56.9±1.2	16.2±0.9
	WeightCLIP-MBM	**18.3±1.8**	45.9±0.3	**11.0±0.2**	**40.0±1.0**	52.6±1.0	**18.1±0.2**
1	tr. fr. scratch	54.1±1.0	60.7±4.1	38.9±0.6	26.1±6.2	56.4±1.9	27.0±1.2
	TANS	29.2±9.3	73.5±1.9	41.5±0.3	32.4±1.6	60.7±7.7	21.9±2.8
	Hypernetwork	59.1±0.5	82.9±0.5	42.7±0.4	21.1±1.0	82.3±3.1	34.2±0.3
	WeightCLIP-LM	**68.4±4.6**	84.7±1.4	46.2±3.1	41.7±6.8	**82.8±6.4**	**35.9±2.7**
	WeightCLIP-MBM	68.4±3.8	**90.8±0.6**	**47.7±3.5**	**48.0±2.6**	72.9±3.7	35.4±1.7
10	tr. fr. scratch	74.8±5.7	81.7±7.1	68.1±0.7	48.3±8.4	84.7±1.2	42.6±1.2
	TANS	71.9±2.3	89.0±0.5	68.4±0.1	51.9±10.5	92.6±6.9	44.3±2.2
	Hypernetwork	75.1±0.8	90.0±0.3	**71.5±0.5**	58.2±1.2	92.8±2.3	48.7±0.8
	WeightCLIP-LM	**84.6±1.3**	92.6±0.8	70.5±4.3	70.6±5.4	**95.3±2.3**	**55.2±1.2**
	WeightCLIP-MBM	82.2±1.4	**95.0±0.3**	70.3±5.4	**74.3±2.3**	94.6±2.8	53.1±1.1

Takeaways

- Dataset semantics provide a reference frame for weight space.
- A subset of a dataset can serve as a prompt for model weights.
- Aligned latent spaces support retrieval, generation, and latent refinement.

WeightCLIP: Prompt neural network weights with data

Universität St.Gallen

ICML International Conference On Machine Learning

Thank You!

Email: aron.asefaw@unisg.ch

Code: github.com/HSG-AIML/WeightCLIP

Poster Session 6: HALL A