BIRD: Behavior Induction via Representation Structure Distillation

Galen Pogoncheff and Michael Beyeler

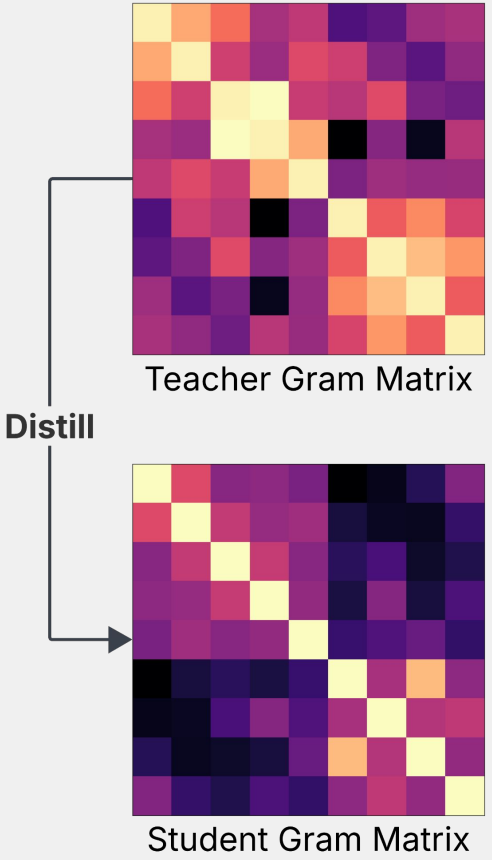
Introduction

- Aligning AI models with human values (safety, robustness, fairness, …) becomes increasingly important as models become more capable
  - Behavioral alignment supervision is traditionally expensive
  - Aligned behavior is easily forgotten during fine-tuning (Qi et al., 2023)

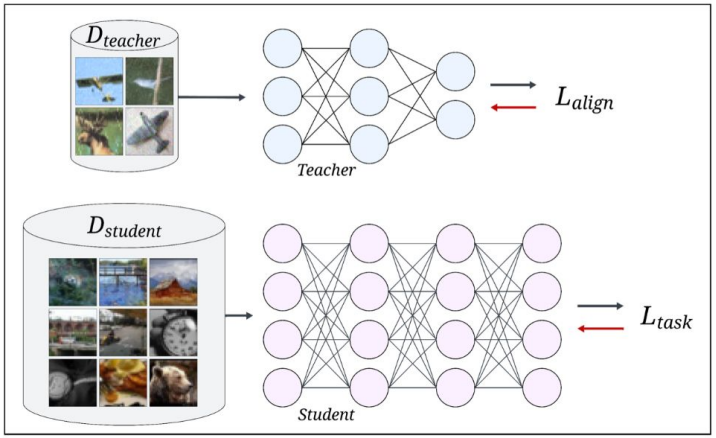
Can aligned behavior be transferred from one model to another, despite differences in tasks and data distributions?

Hypothesis

- **Hypothesis:** behavioral properties (e.g., robustness, safety) are encoded not only in model outputs and directions in feature space, but also in representation structure
  - Representation space geometry reflects how information is organized
- BIRD mechanizes this in the context of knowledge distillation
  - Instead of distilling predictions or matching activations, a student model is trained to **match the representation structure** of a behaviorally-aligned teacher

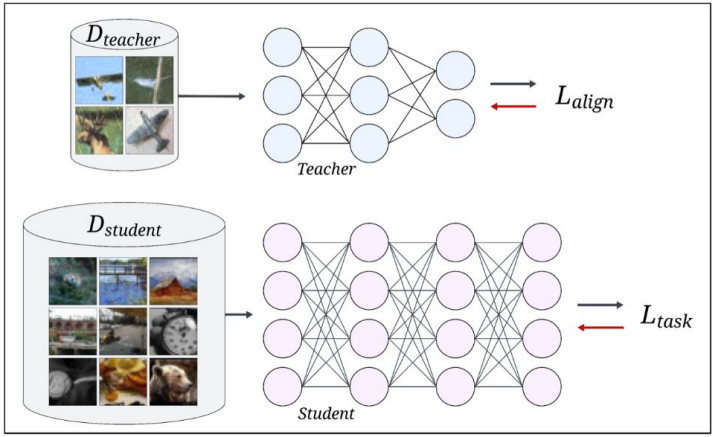


BIRD

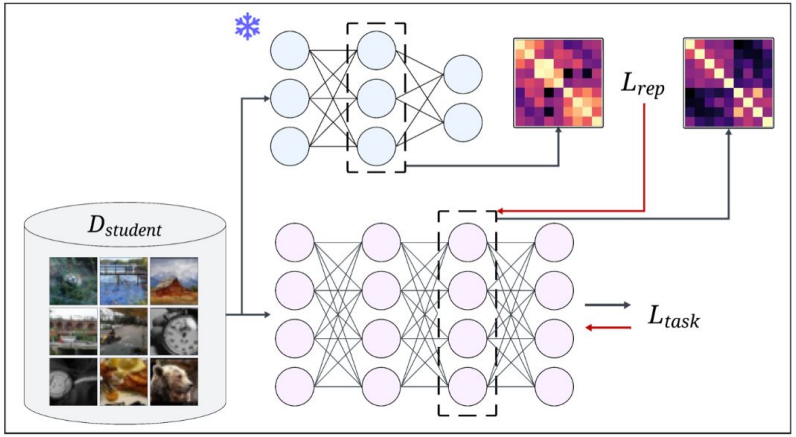


(a) Independent model pre-training

BIRD

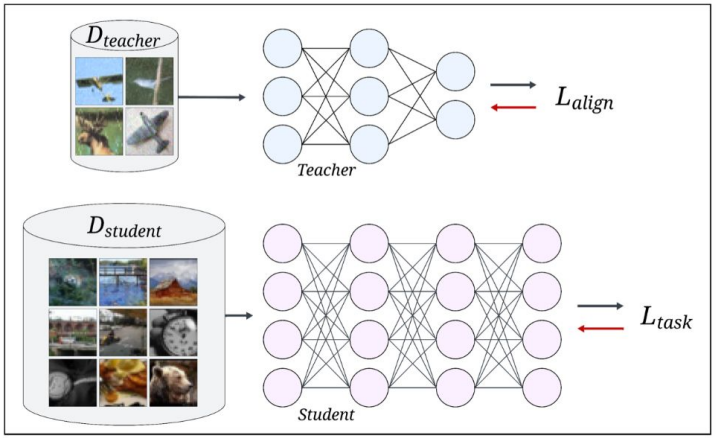


(a) Independent model pre-training

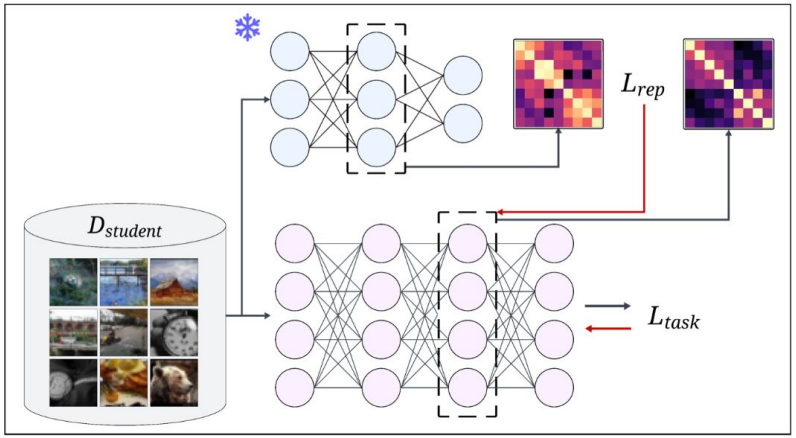


(b) Layer selection + representation-structure distillation

BIRD



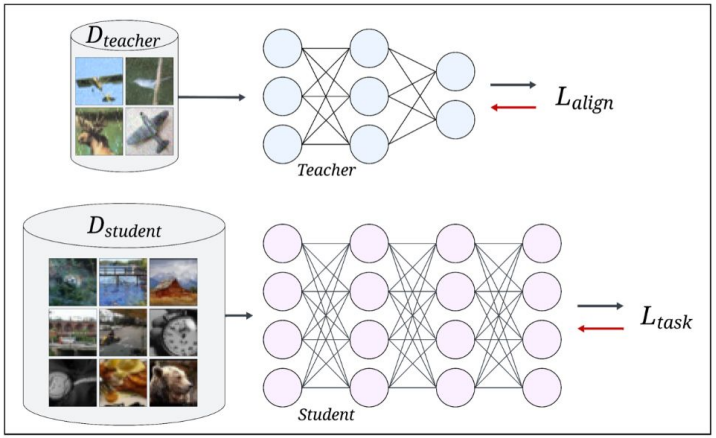
(a) Independent model pre-training



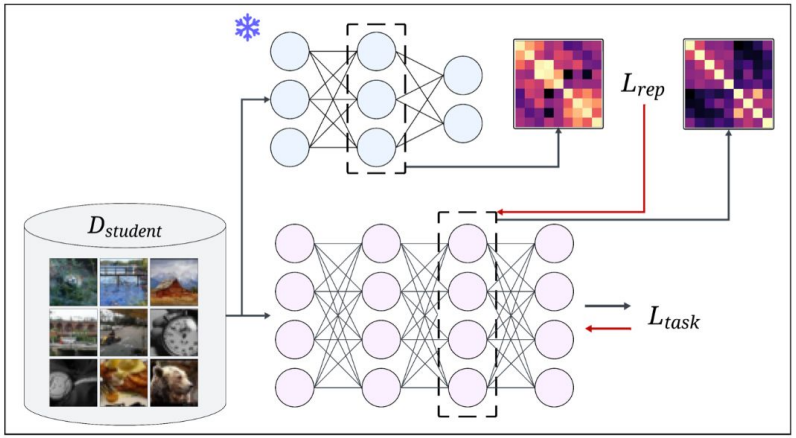
(b) Layer selection + representation-structure distillation

$$\mathbb{E}_{B\sim\mathcal{D}_{\text{student}}}\Big[\alpha\mathcal{L}_{\text{task}}\big(f_\theta(B),\cdot\big)+\beta\mathcal{L}_{\text{rep}}\big(u(B),v(B)\big)\Big]$$

BIRD



(a) Independent model pre-training

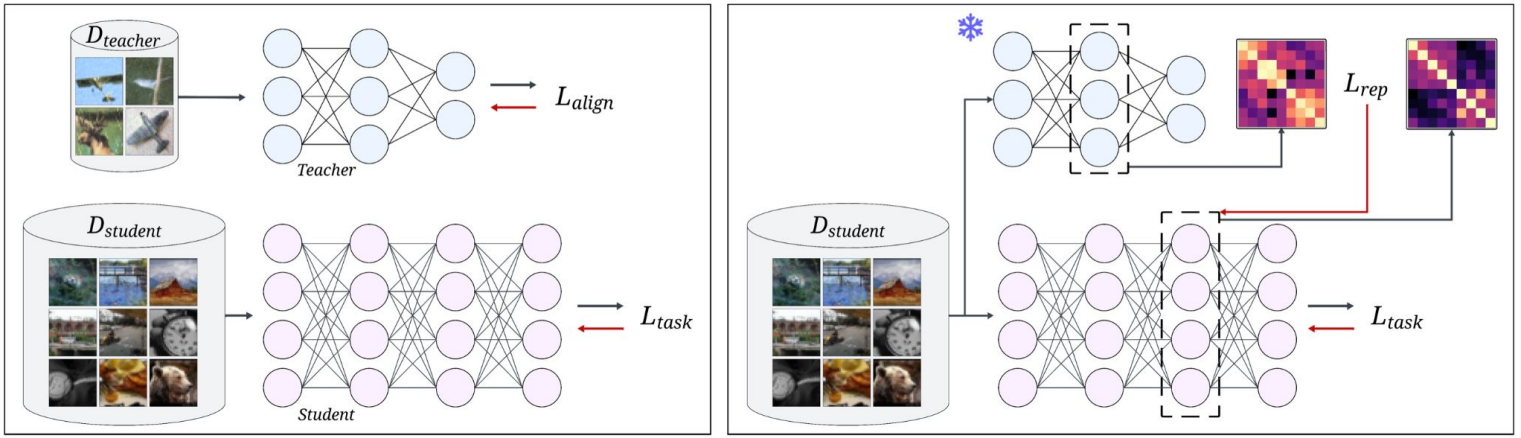


(b) Layer selection + representation-structure distillation

$$\mathbb{E}_{B\sim\mathcal{D}_{\text{student}}}\left[\alpha\mathcal{L}_{\text{task}}\left(f_\theta(B),\cdot\right)+\beta\mathcal{L}_{\text{rep}}\left(u(B),v(B)\right)\right]$$

Maintain performance on original training task

BIRD



(a) Independent model pre-training

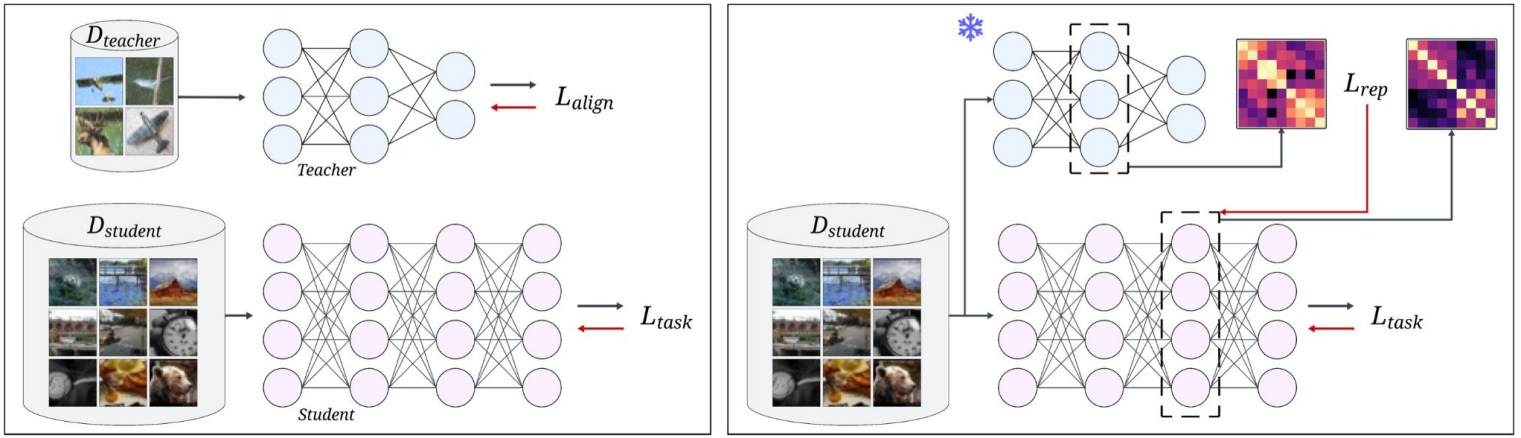
(b) Layer selection + representation-structure distillation

$$\mathbb{E}_{B\sim\mathcal{D}_{\text{student}}}\left[\alpha\mathcal{L}_{\text{task}}\left(f_\theta(B),\cdot\right)+\beta\mathcal{L}_{\text{rep}}\left(u(B),v(B)\right)\right]$$

Maintain performance on original training task

Match representation geometry of teacher

BIRD



(a) Independent model pre-training

(b) Layer selection + representation-structure distillation

$$\mathbb{E}_{B\sim\mathcal{D}_{\text{student}}}\left[\alpha\mathcal{L}_{\text{task}}\left(f_\theta(B),\cdot\right)+\beta\mathcal{L}_{\text{rep}}\left(u(B),v(B)\right)\right]$$

$$\mathcal{L}_{rep}(u(B),v(B))=1-\text{CKA}_{\text{linear}}(u(B),v(B))$$

Maintain performance on original training task

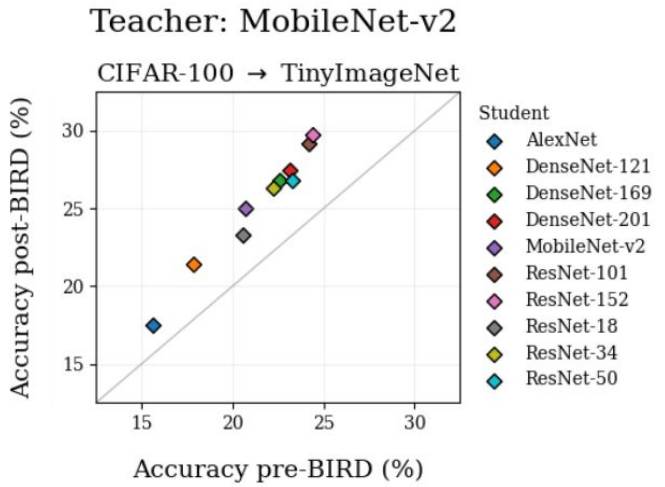
Match representation geometry of teacher

Transferring robustness in vision models

Scenario

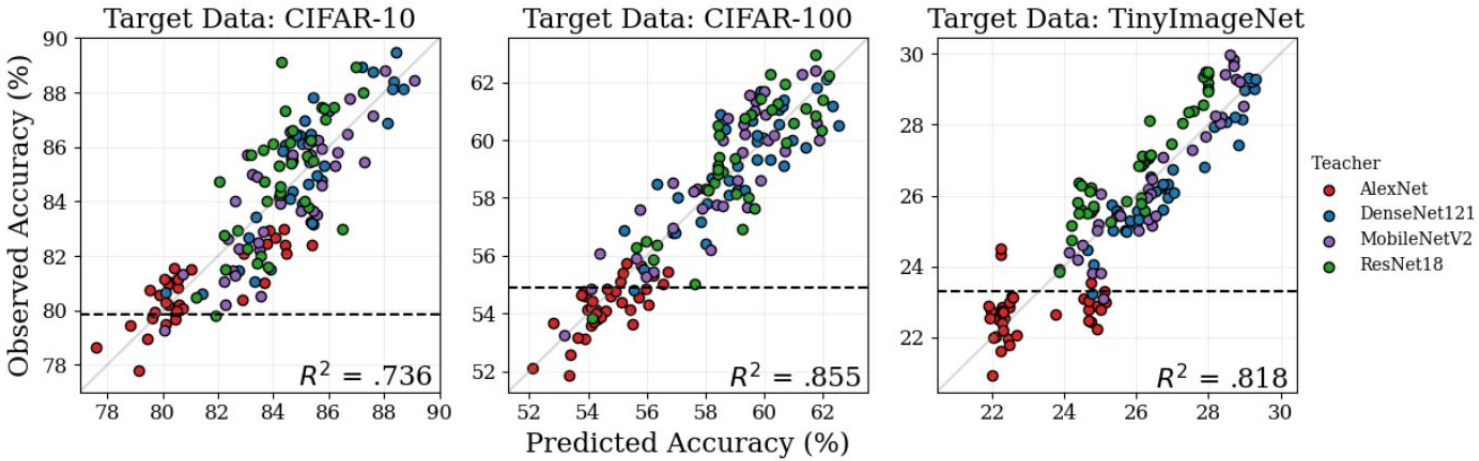
- We have
  - **Teacher model:** robust to out-of-distribution (OOD) image corruptions on dataset $D_{teacher}$
  - **Student model:** non-robust model, but performant on in-distribution data from $D_{student}$
- Want
  - Student with **improved robustness on $D_{student}$, without requiring additional data**

Model	Source Data	Target Data	Accuracy of Behavior Transfer Method (↑)						
			None	LP	FT	LP-FT	Hints	LwF	BIRD
MN2	C10	C100	51.31	10.95	51.12	47.56	51.53	52.21	**54.77**
	C10	TIN	20.74	5.24	20.00	18.12	21.27	20.52	**24.11**
	C100	TIN	20.74	18.84	20.66	23.52	21.26	23.18	**25.03**
	C10	IN	22.59	1.29	22.50	22.36	22.25	23.23	**23.38**
	C100	IN	22.59	6.18	22.36	22.34	22.49	23.27	**23.68**



Practical guidance on teacher selection

3 interpretable and computable properties of the teacher's representations, related to task and behavioral relevance, explain up to 85% of the variance in robustness transfer success



Generalization beyond vision

BIRD can be used alongside popular approaches in safety alignment (e.g., DPO) to **enhance safe behavior in language models**

Table 2: Safety alignment performance of generative models on the PKU-SafeRLHF test set. % *Safe*: percentage of query responses from that model evaluated as safe according to an LLM judge.

Student	% Safe (↑)		
	None	DPO	DPO+BIRD
SmolLM2-135M-Instruct	43.88	65.48	**71.28**
SmolLM2-360M-Instruct	47.63	86.57	**88.37**

Conclusion

- BIRD: a framework for transferring behavior between heterogeneous models
  - Built on the hypothesis that well-generalizing behavioral properties are encoded in the structure of a model's representation space
  - **Small, well-aligned models models can be used as scalable alignment seeds**, mitigating challenges from key bottlenecks in deploying safe AI systems
- We demonstrate BIRD's efficacy on improving robustness in image classification models and safety-alignment in small language models
  - Behavioral transfer success is predicted by task- and behavioral-relevance of teachers representations