Unifying Synergies between Self-supervised Learning and Dynamic Computation

Tarun Krishna, Ayush Rai, Alexandru Drimbarean, Eric Arazo, Paul Albert, Alan F. Smeaton, Kevin McGuinness and Noel O'Connor

Insight SFI RESEARCH CENTRE FOR DATA ANALYTICS

A World Leading SFI Research Centre

Science Foundation Ireland For what's next

Goal: *Can we unify the learning of a lightweight sub-network along with a dense network from scratch and in a completely self-supervised fashion?*

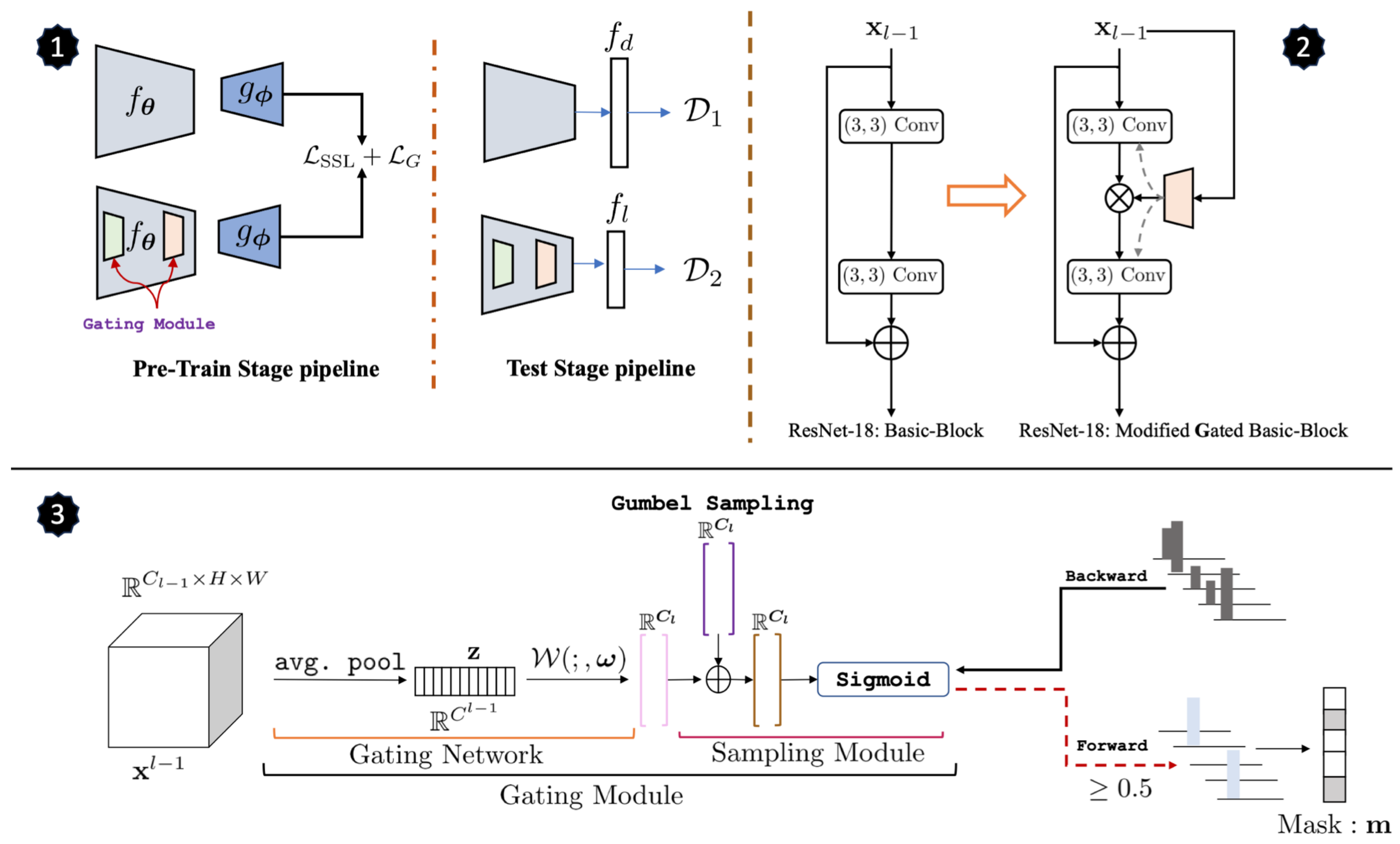
Motivation:

- Computationally expensive training strategies make self-supervised learning (SSL) impractical for resource constrained industrial settings.
- Knowledge distillation (KD), dynamic computation (DC) and pruning are often used for obtaining lightweight models, but this requires multiple fine-tuning steps of a large pre-trained model, posing computational challenges.
- Downstream tasks are diverse and vary widely any change in the task requires repeating the procedure multiple times, reducing efficiency and transferability.

Key Contributions:

- We present a novel perspective of unifying the learning of dense and lightweight networks by exploiting a symmetric joint embedding architecture of the SSL paradigm.
- We demonstrate that a single encoder can be exploited as a dense as well as a lightweight network. This not only reduces computational overhead during training but also gives enough flexibility to use a single network and exploit it accordingly.
- This unification preserves feature quality across different experimental settings.

End-to-End Unification Pipeline for SSL and DC



- Our approach comprises of training a dense branch and sparse branch derived from dense branch via gating mechanism during pre-training only.
- Both the branches share different batch-normalization layers, because each branch have different batch statistics.
- We exploit **VICReg** [1] as our SSL-objective as it regularizes each branch independently making it suitable for the task at our end.

Quantitative Results

Baselines: To exhaustively compare the performance of the dense and gated models we consider VICReg [1] as an SSL ***dense*** baseline while VICReg augmented with sparsity loss L_G (following Krishna et al. [2]) serves as a ***gated*** baseline.

Dataset	VICReg Baseline-1 Bardes *et al.* [4] Dense	FLOPs	t_d(%)	VICReg-Gating Baseline-2 Krishna *et al.* [44] Gated	FLOPs R.	VICReg-Dual-Gating *this work* Dense ↑	Gated ↑	FLOPs R. ↑
CIFAR-10	91.11 ±0.03	7.03E8	10%	87.75 ±0.03	85.92%	88.99 ±0.04 (↓2.12)	88.94 ±0.06 (↓2.17) (↑1.19)	81.49% (↓4.43)
			30%	89.49 ±0.04	69.27%	90.38 ±0.04 (↓0.73)	90.27 ±0.03 (↓0.84) (↑0.78)	66.43% (↓2.84)
			50%	90.70 ±0.04	51.62%	90.20 ±0.02 (↓0.91)	90.40 ±0.06 (↓0.71) (↓0.30)	49.02% (↓2.60)
STL-10	86.15 ±0.10	3.33E8	10%	82.48 ±0.15	82.85%	84.29 ±0.21 (↓1.86)	83.29 ±0.05 (↓2.86) (↑0.81)	78.34% (↓4.51)
			30%	84.16 ±0.11	68.38%	84.90 ±0.05 (↓1.25)	84.85 ±0.04 (↓1.30) (↑0.69)	65.24% (↓3.14)
			50%	85.40 ±0.20	49.93%	85.75 ±0.02 (↓0.40)	85.72 ±0.02 (↓0.43) (↑0.32)	48.41% (↓1.52)
CIFAR-100	65.86 ±0.10	7.03E8	10%	63.12 ±0.09	84.82%	65.21 ±0.06 (↓0.65)	64.31 ±0.08 (↓1.55) (↑1.19)	81.71% (↓3.11)
			30%	65.41 ±0.09	68.68%	65.90 ±0.10 (↑0.04)	65.64 ±0.00 (↓0.22) (↑0.23)	66.83% (↓1.85)
			50%	65.75 ±0.12	50.04%	66.41 ±0.05 (↑0.55)	66.40 ±0.14 (↑0.54) (↑0.65)	49.06% (↓0.98)
ImageNet-100	77.74 ±0.12	1.81E9	30%	74.04 ±0.09	67.95%	75.12 ±0.07 (↓2.62)	75.04 ±0.10 (↓2.17) (↑1.00)	64.98% (↓2.97)
			50%	75.83 ±0.07	50.11%	76.42 ±0.26 (↓1.32)	76.24 ±0.12 (↓1.51) (↑0.41)	47.69% (↓2.42)

- The lightweight gated network achieves improved performance across all datasets and target budgets (t_d) as compared to Baseline-2 [2], with a negligible drop at t_d = 50% for CIFAR-10 only.
- The performance gain is compensated by a slightly smaller reduction in FLOPs as compared to Baseline-2 [2].
- Another important aspect of our learning method is the performance of the ***dense*** (f_θ) model. Aim is to achieve fewer fluctuations with varying t_d with a performance equivalent to Baseline-1 [1]. However, we find that the performance of the dense network (this work) is slightly below the performance of the dense Baseline-1 [1].
- The learned structure is similar to dense (VICReg [1]) at a very low budget.

Qualitative Results

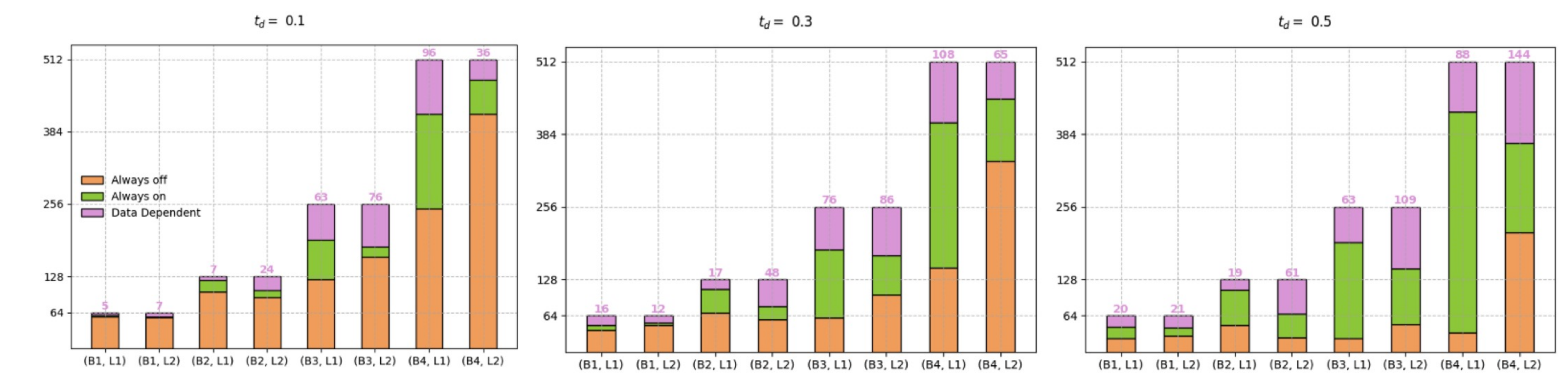


Figure 1: Learned channel distribution for CIFAR-100 with varying t_d.

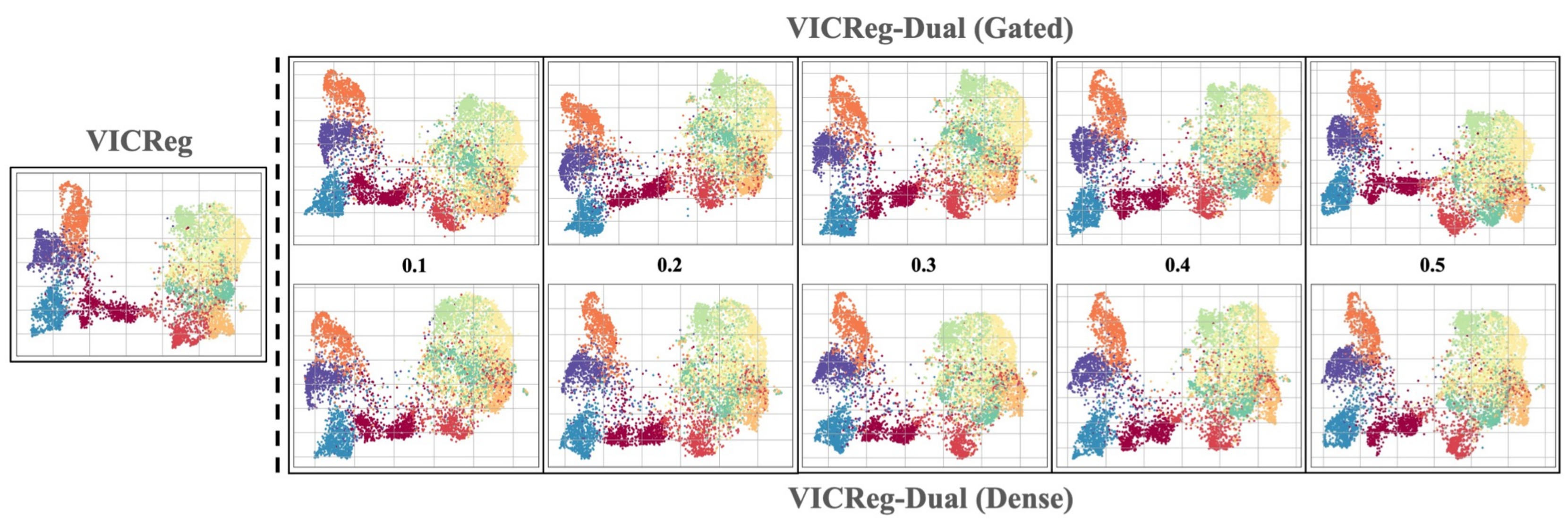


Figure 2: **Qualitative analysis:** UMAP embeddings of the learned representations: *lightweight* gated network (*top* row), while dense network (*bottom*) row over different target budgets t_d. This is compared with embeddings of VICReg (dense) trained without any sort of sparsity. Best viewed in color.

Limitations

- Dense model performance degrades and fluctuates with varying (t_d).
- No constraints to enforce more conditional computation during inference.

References

1. Bardes, Adrien, Jean Ponce, and Yann LeCun. "Vicreg: Variance-invariance-covariance regularization for self-supervised learning." *arXiv preprint arXiv:2105.04906* (2021).
2. Krishna, Tarun, et al. "Dynamic Channel Selection in Self-Supervised Learning." 24th Irish Machine Vision and Image Processing Conference. 2022.

Acknowledgements: This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183, co-funded by the European Regional Development Fund and Xperi.

BMVC 2023

Insight SFI RESEARCH CENTRE FOR DATA ANALYTICS

DCU Ollscoil Chathair Bhaile Átha Cliath Dublin City University

XPERI

SCAN ME