D³Former: Debiased Dual Distilled Transformer for Incremental Learning

Abdelrahman Mohamed [1*] Rushali Grandhe [1*] K J Joseph [2] Salman Khan [1,3] Fahad Khan [1,4]

[1] MBZUAI [2] Adobe Research [3] Australian National University [4] Linköping University

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

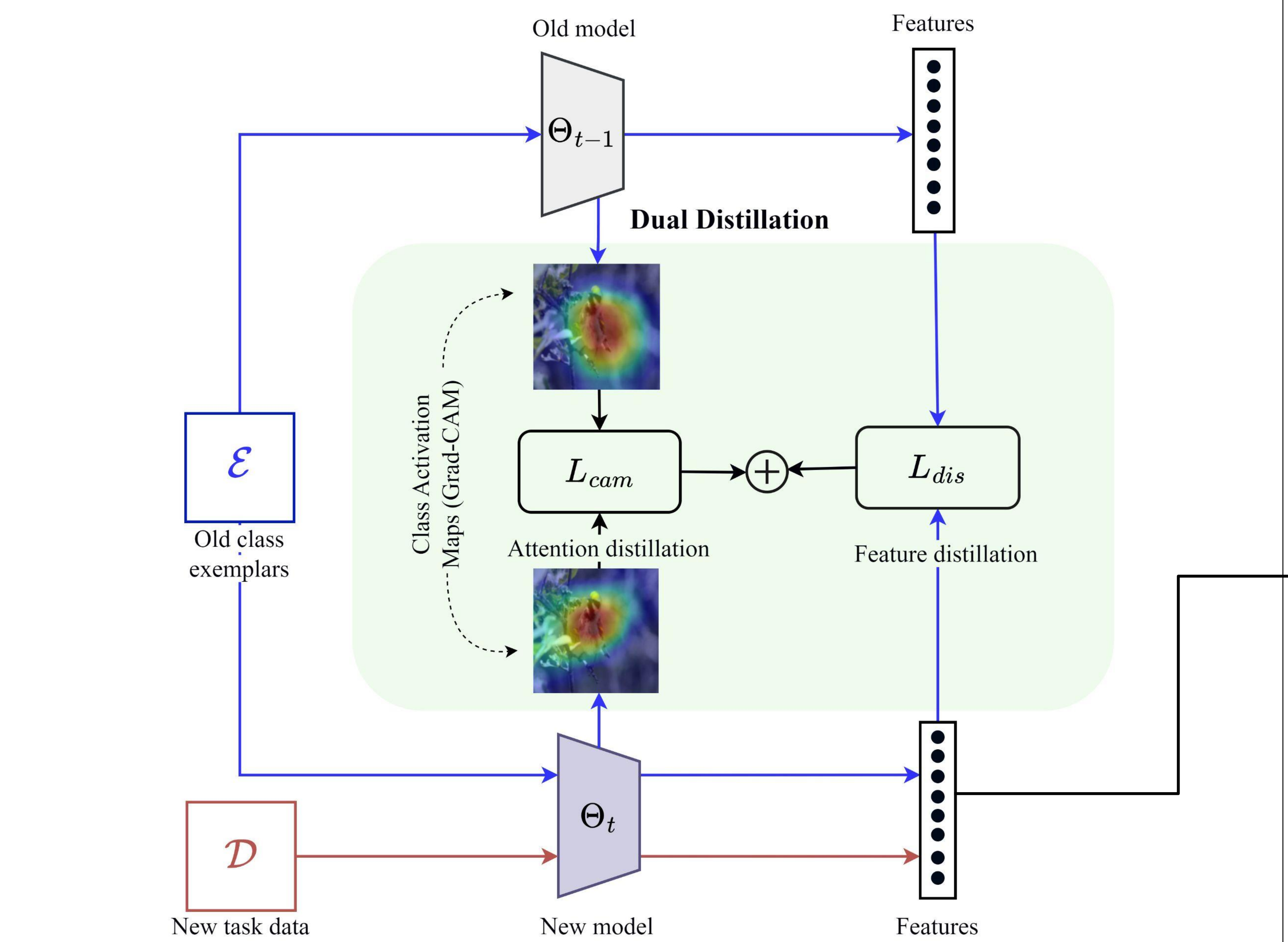
MOTIVATION

- Previous works extensively analyzed CNNs in class incremental learning (CIL), while limited works have studied the behaviour of ViTs.
- Catastrophic Forgetting Problem in CIL
  a. Class imbalance due to few exemplars vs new class samples.
  b. Distribution shift between incremental sets of classes.

MAIN CONTRIBUTIONS

- We study NesT[a] *a data-efficient and interpretable hybrid ViT* in CIL.
- Treat class imbalance in CIL as a long tail problem: reducing bias by simple logit adjustment strategy.
- Retain attention over important regions of exemplars using interpretability methods.
- Scalable to small and large datasets without expanding architecture.

METHOD



Dual Distillation:

- Knowledge distillation over exemplars' attention-maps improves spatial awareness.
- Feature distillation[b] provides additional stability during CIL.

$$L_{cam}(x) = \| CAM(\boldsymbol{\Theta_t}, \boldsymbol{x}) - CAM(\boldsymbol{\Theta_{t-1}}, \boldsymbol{x}) \|_1$$

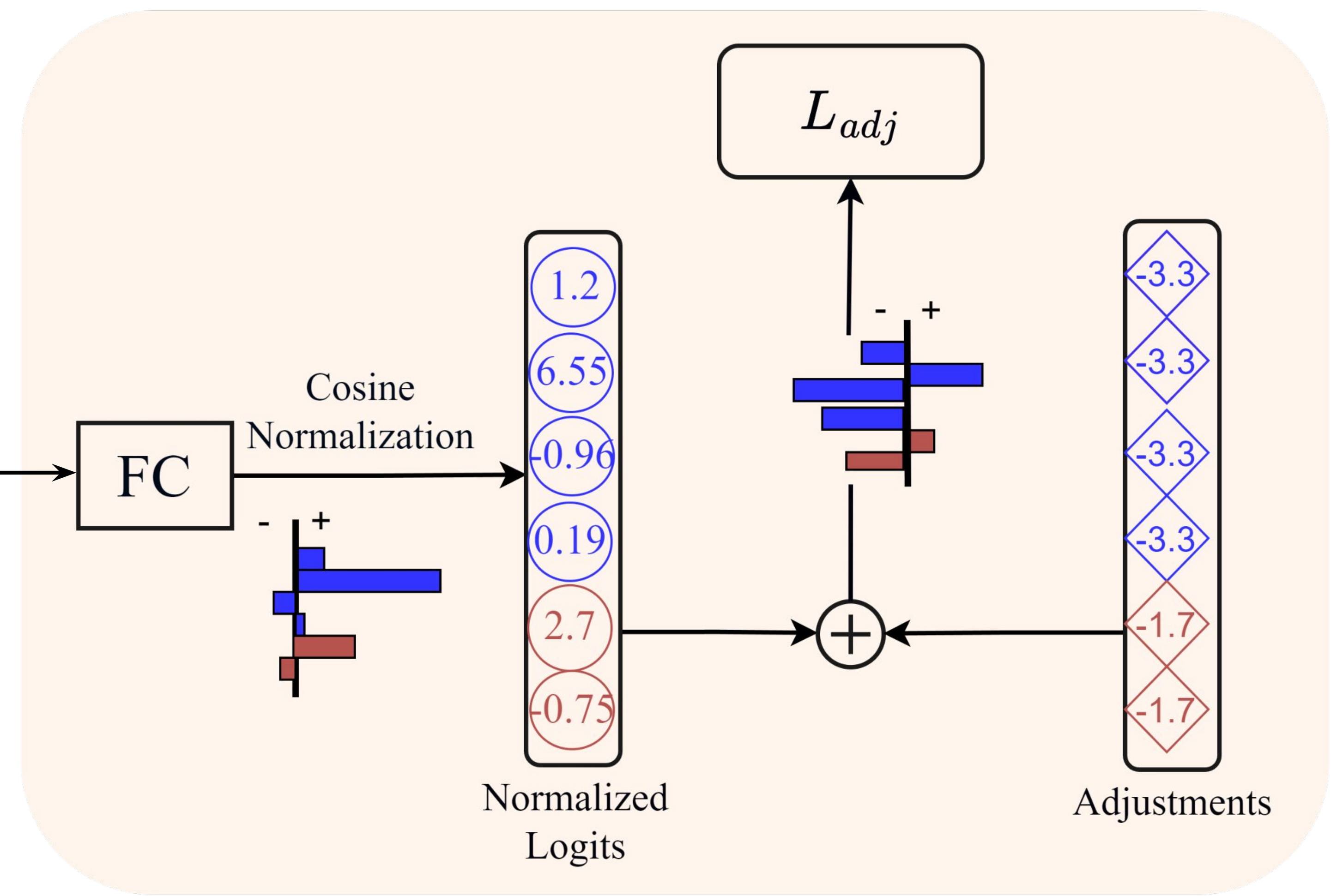
$$L_{dis} = 1 - \langle \bar{\theta}_{t-1}(\boldsymbol{x}), \bar{\theta}_t(\boldsymbol{x}) \rangle$$

Debiasing via Logit Adjustment:

- Adjusting the logits[c] of old and new class samples by an offset dampens bias caused by inherent class imbalance in CIL.
- Offset equals $\tau \log \pi_y$, where τ is a hyperparameter controlling adjustment strength, π_y is the estimated prior for class y.

$$L_{adj}(\boldsymbol{x}) = -\log \frac{e^{f_y(\boldsymbol{x}) + \tau \log \pi_y}}{\sum_{y' \in \mathcal{T}} e^{f_{y'}(\boldsymbol{x}) + \tau \log \pi_{y'}}}$$

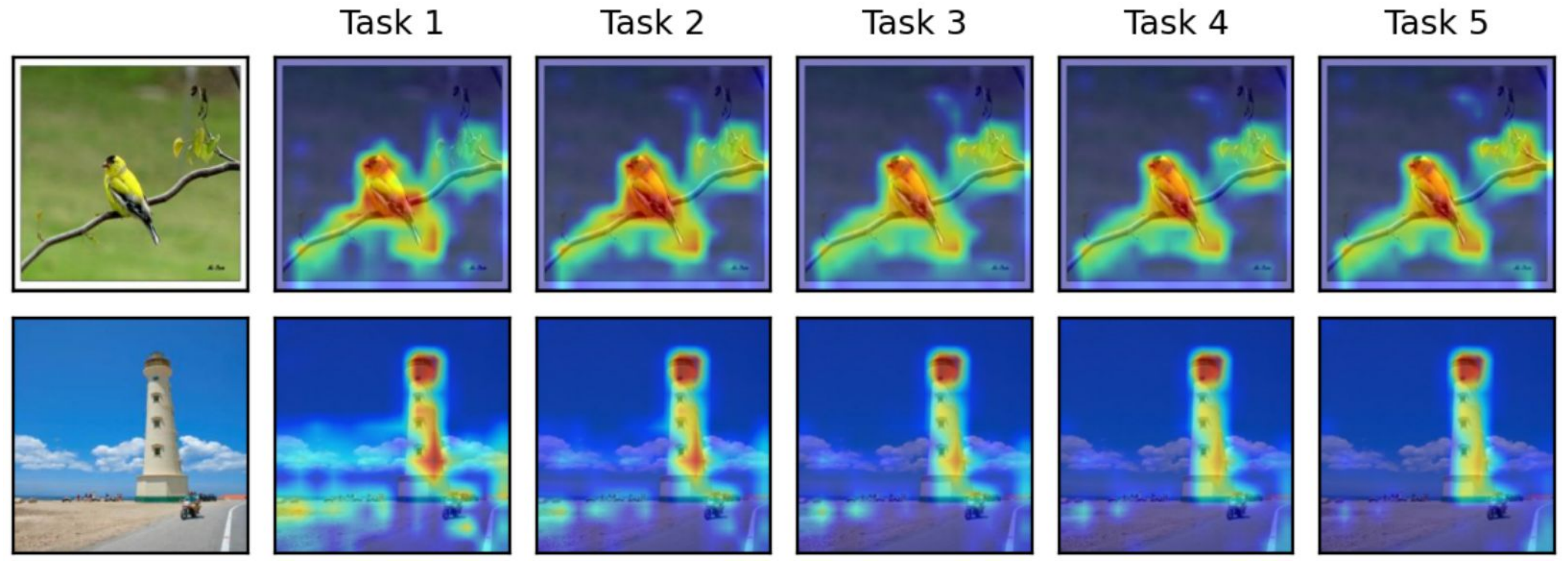
Debiasing Block



RESULTS

Method	25 tasks CIFAR-100 Avg ↑	25 tasks CIFAR-100 Last ↑	25 tasks CIFAR-100 $\mathcal{F}$ ↓	5 tasks ImageNet-100 Avg ↑	5 tasks ImageNet-100 Last ↑	5 tasks ImageNet-100 $\mathcal{F}$ ↓
LwF	45.51	38.25	41.66	53.62	40.10	55.32
BiC	50.00	-	34.60	70.07	-	27.04
iCaRL	48.22 $\pm$ 0.76	39.39	36.48	65.44 $\pm$ 0.35	53.60	43.40
LUCIR	57.54 $\pm$ 0.43	48.35	26.46	70.84 $\pm$ 0.69	60.00	31.88
Mnemonics	60.96 $\pm$ 0.72	50.78	**19.80**	75.54 $\pm$ 0.85	61.36	**17.40**
PODNet-CNN	60.72 $\pm$ 1.54	51.40	-	76.96 $\pm$ 0.29	67.60	-
DyTox	62.83	53.95	33.72	77.08	**70.24**	21.21
D³Former (ours)	**68.68** $\pm$ 0.4	**59.79** $\pm$ 0.44	<u>21.23</u>	**77.31** $\pm$ 0.41	67.82 $\pm$ 0.36	25.92
D³Former-NCM (ours)	<u>67.03</u> $\pm$ 0.59	<u>58.12</u> $\pm$ 0.80	22.84	<u>77.21</u> $\pm$ 0.22	<u>69.89</u> $\pm$ 0.18	<u>17.98</u>

Results with Average accuracy (%), last phase accuracy (%) and forgetting rate F (%) for small and large datasets in multiple task settings. D³Former achieves performance gains over several methods.



Grad-CAMs visualized for 5 incremental tasks of ImageNet-100. The model exhibits minimal forgetting and leverages discriminatory regions to generate predictions.

CONCLUSION

- Hybrid Vision transformers have favorable performance over CNNs.
- Incremental learning can be treated as a long-tail distribution problem.
- Interpretability methods can help reduce catastrophic forgetting.
- Architectures with better interpretability have stronger continual learning capabilities.

REFERENCES

a. Zhang et al., Nested hierarchical transformer (NesT), AAAI'22
b. Hou et al., Learning a unified classifier incrementally via rebalancing, CVPR'19
c. Menon et al., Long-tail learning via logit adjustment, ICLR'21