Exemplar-Free Continual Transformer with Convolutions

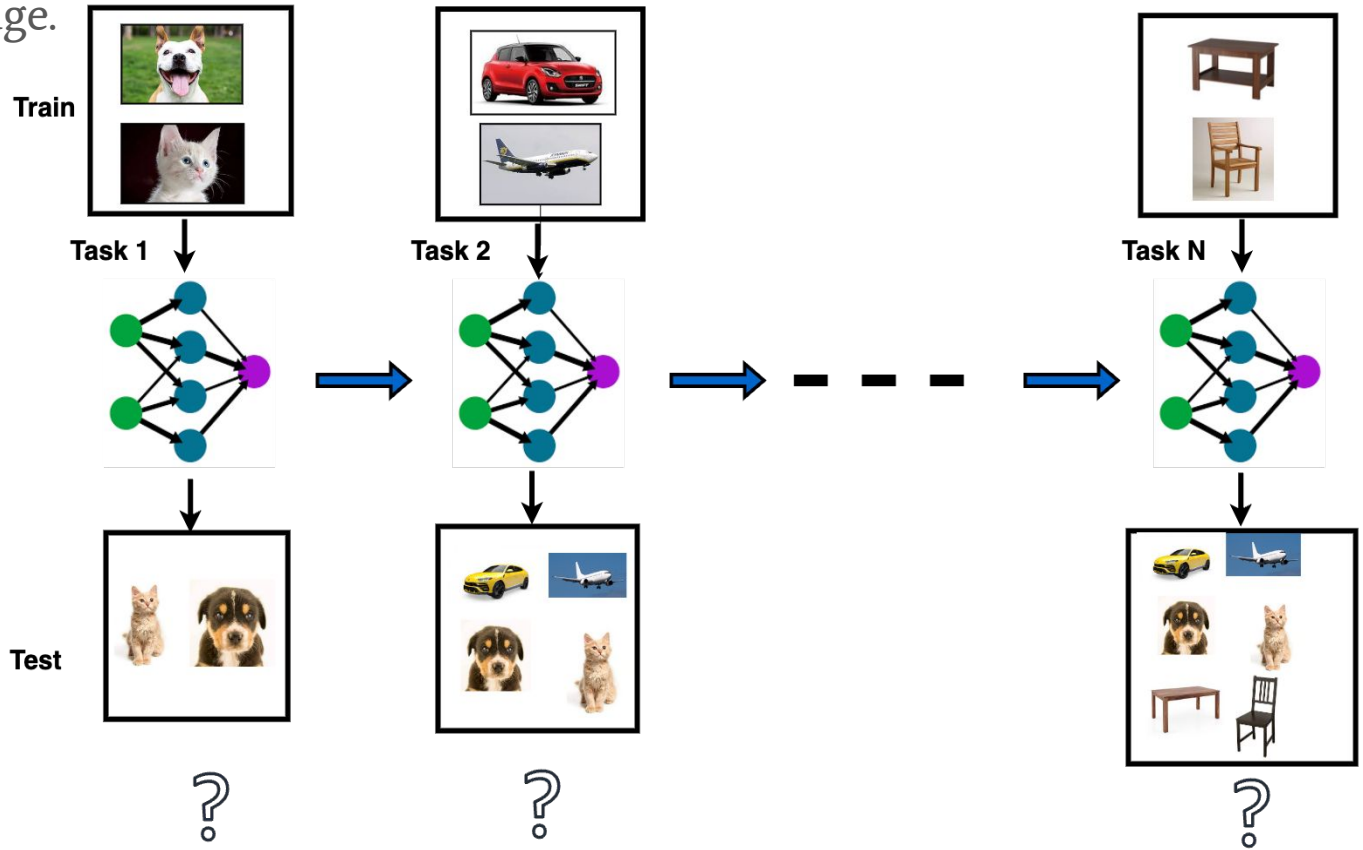
Anurag Roy[3], Vinay K. Verma[1], Sravan Voonna[3], Kripabandhu Ghosh[2], Saptarshi Ghosh[3], Abir Das[3]

[1]IML, Amazon India,
[2]Indian Institute of Science Education and Research (IISER) Kolkata, India,
[3]Indian Institute of Technology Kharagpur, India

ICCV23
PARIS

Continual Learning

- A learning paradigm where a model can learn a new task without forgetting the previous tasks' knowledge.



Motivation

- Deep neural networks suffer from catastrophic forgetting.
- Retraining model from scratch or training a separate model for each task incurs a lot of resource.
- Most CL approaches in vision are based on CNN backbones

Prior Works on CL using Transformers

There have been a few prior-works in this field:

- LVT[1]: Uses an inter-task attention mechanism that absorbs the previous tasks' information and slows down the information drift between new and current tasks.

- Dytox[2]: learns new task through expansion of new tokens known as task tokens.

- MEAT[3]: Uses learnable masks to help isolate previous tasks' parameters that are required for current task.

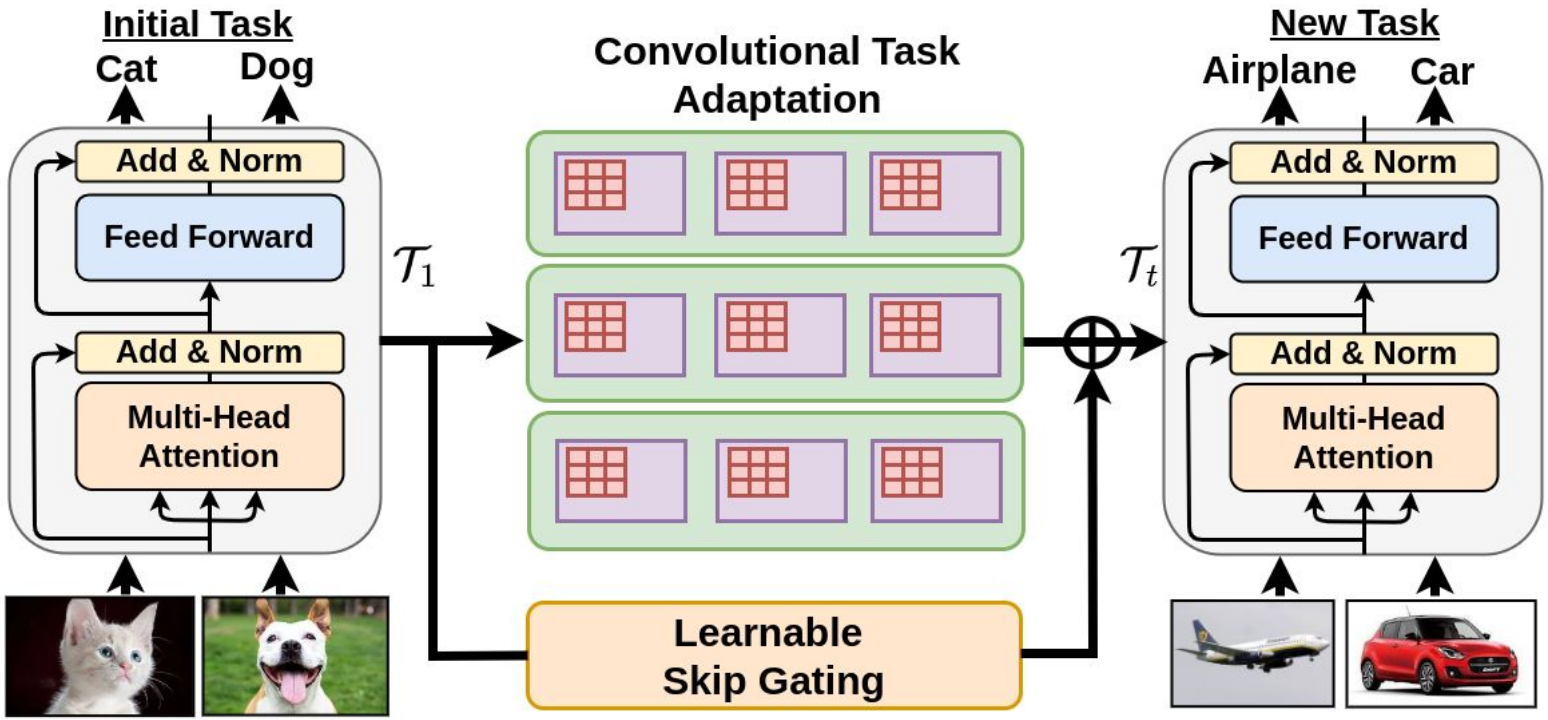
[1]Zhen Wang, et. al. Continual learning with lifelong vision transformer. CVPR 2022.
[2]Arthur Douillard, et. al. Dytox: Transformers for continual learning with dynamic token expansion. CVPR, 2022.
[3]Mengqi Xue, et. al. Meta-attention for vit-backed continual learning. CVPR 2022

Limitations of Prior Works

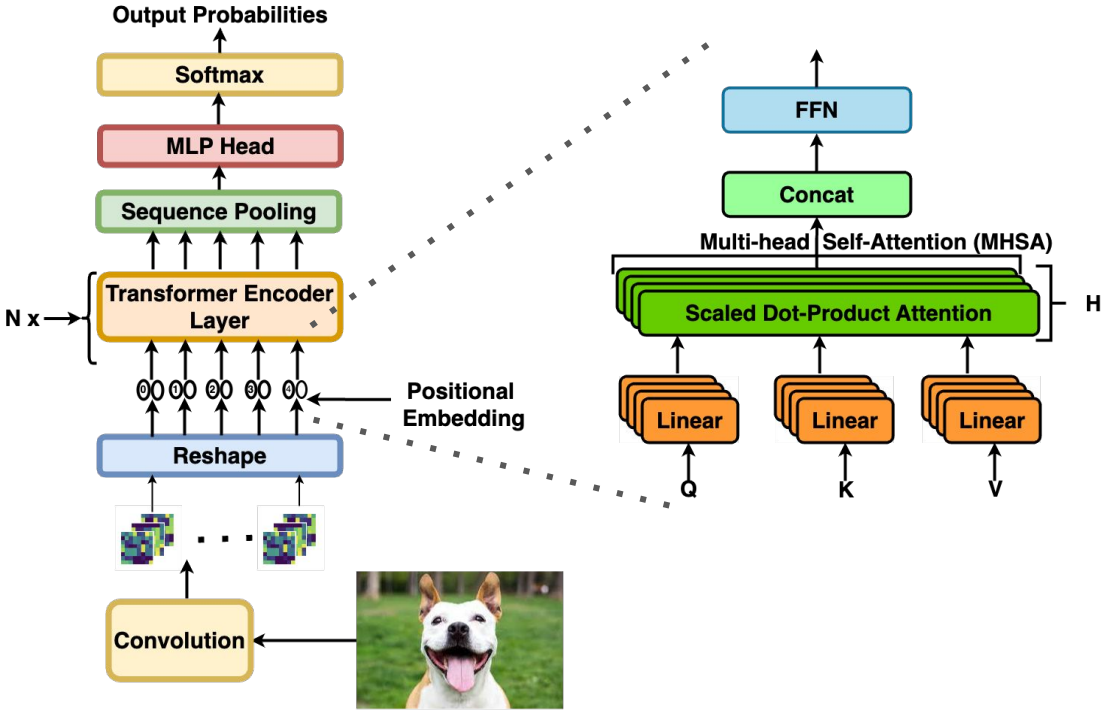
- LVT and Dytox: Requires to store few representative data samples from previous tasks also known as exemplars and use them when training for the new task. Cannot be applied for cases where data storage is not allowed.

- MEAT: Requires task-id to be present during inference for identification of task-specific masks. Not practical for scenarios where task-id is not present during inference.

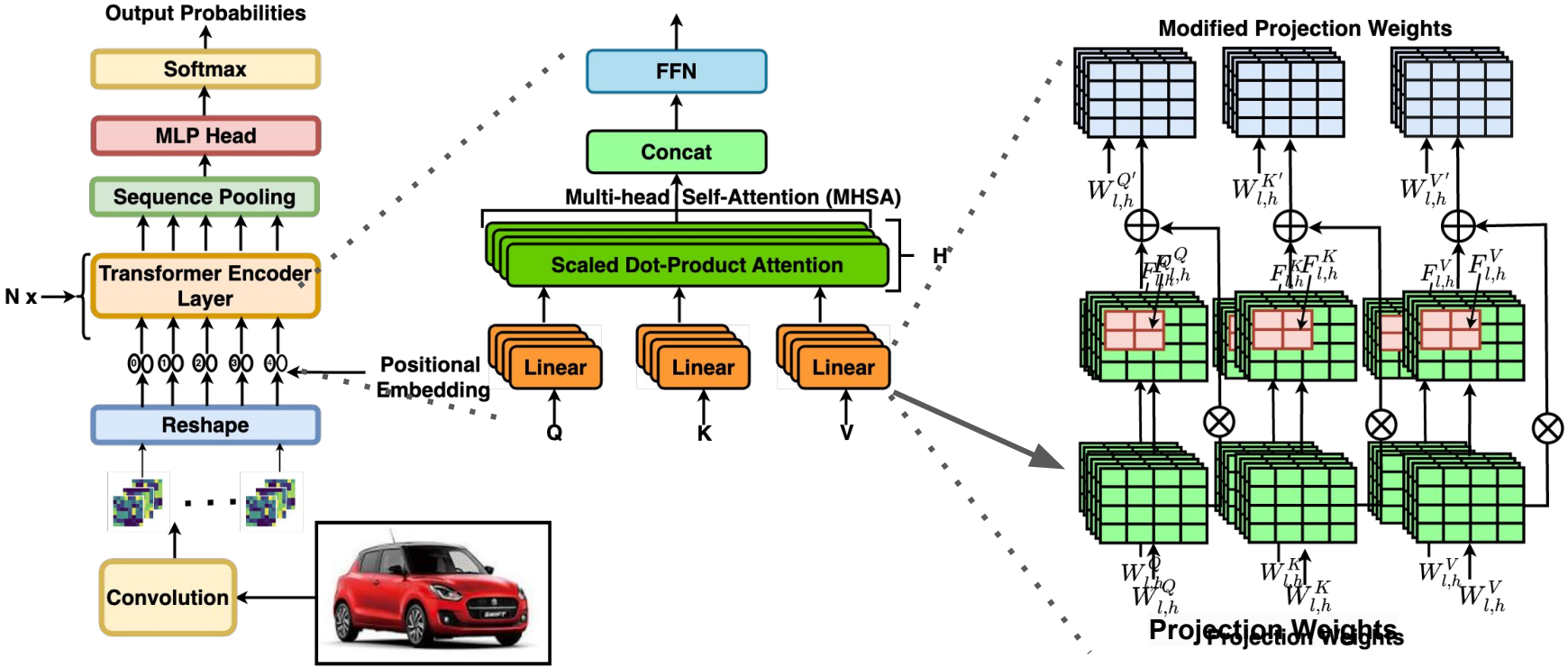
ConTraCon


Initial Task
Cat
Dog
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
$\mathcal{T}_1$
Convolutional Task Adaptation
Learnable Skip Gating
$\mathcal{T}_t$
New Task
Airplane
Car
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention

ConTraCon: Training



ConTraCon: Training

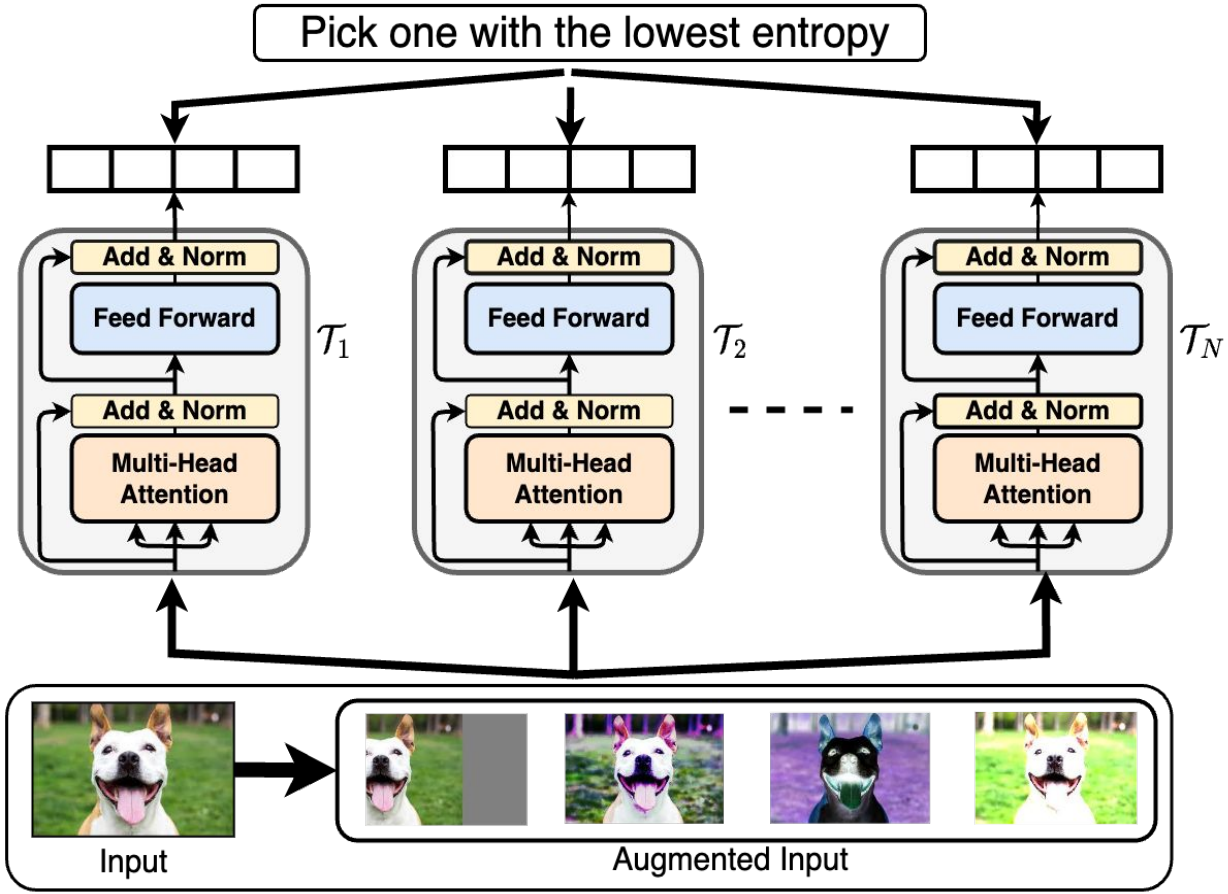


ConTraCon: Training

- We train the entire transformer on the first task.
- For every new incoming task, the weights of the MHSA layers of the pre-trained transformer are re-weighted using learnable task-specific convolutions.

ConTraCon: Inference

- **Task-id Prediction:**
- **Class Prediction.**



ConTraCon: Inference

- **Class Prediction:** Pass input through the parameters of the predicted task-id to get the prediction.

Experiments: Datasets used.

Dataset		Image size	#Train	#Test	#Classes
CIFAR-100		32 x 32	50K	10K	100
TinyImageNet-200		64 x 64	100K	10K	200
ImageNet-100		224 x 224	130K	5K	100
5-Datasets	CIFAR-10	32 x 32	50K	10K	10
	MNIST	32 x 32	60K	10K	10
	SVNH	32 x 32	73K	26K	10
	FashionMNIST	32 x 32	60K	10K	10
	notMNIST	32 x 32	60K	10K	10

Experiments

- We create T-tasks by dividing the classes equally among all the tasks.

- For each of approach we report:
    - Accuracy averaged over all the tasks after the model has been trained on the final task with both task-id provided (TIL) and task-id not provided (CIL) during inference.

    - No of Parameters required for the backbone architecture and the no of parameters required per task (in brackets)

Results

Memory Buffer	Model	Approach	Backbone	# Params	5 Tasks		10 Tasks		20 Tasks	
					TIL	CIL	TIL	CIL	TIL	CIL
200	iCARL [40]	Rehearsal	ResNet 18	11.2 M	55.70	30.12	60.81	22.38	62.17	12.62
	FDR [4]				63.75	22.84	65.88	14.84	59.13	6.70
	DER++ [6]				62.55	27.46	59.54	21.76	61.98	15.16
	ERT [7]				54.75	21.61	58.49	12.91	62.90	10.14
	RM [2]				62.05	32.23	66.28	22.71	68.21	15.15
	LVT [52]		Transformer	8.9 M	66.92	39.68	72.80	35.41	73.41	20.63
	Dytox [16]			10.7 M	75.17	40.97	84.84	32.08	85.24	15.96
500	iCARL [40]	Rehearsal	ResNet 18	11.2 M	64.4	35.95	71.02	30.25	72.26	20.05
	FDR [4]				69.11	29.99	74.22	22.81	73.22	13.10
	DER++ [6]				70.74	38.39	73.31	36.15	70.55	21.65
	ERT [7]				62.85	28.82	68.26	23.00	73.50	18.42
	RM [2]				69.27	39.47	73.51	32.52	75.06	23.09
	LVT [52]		Transformer	8.9 M	71.54	44.73	76.78	43.51	78.15	26.75
	Dytox [16]			10.7 M	76.1	57.66	**88.72**	**47.34**	87.23	29.89
–	EFT [51]	Dynamic Arch	ResNet 18	4.9 M (32k)	79.04	**49.68**	83.14	40.42	76.75	19.15
	PASS [60]	Regulaization	ResNet 18	11.2 M	70.11	47.31	71.28	35.24	71.14	23.15
	GPM [43]	Regularization	AlexNet	6.7 M	65.90	–	72.54	–	77.59	–
	ConTraCon	Dynamic Arch	Transformer	**3.1 M (26k)**	**79.37**	48.46	85.69	41.26	**88.94**	**30.07**

Classification accuracy on CIFAR-100 dataset.

Results

Memory Buffer	Model	Approach	Backbone	# Params	ImageNet-100/10		TinyImageNet-200/10	
					TIL	CIL	TIL	CIL
200	iCARL [40]	Rehearsal	ResNet 18	11.2 M	33.75	12.59	28.41	8.64
	FDR [4]				37.80	10.08	40.15	8.77
	DER++ [6]				31.96	11.92	40.97	11.16
	ERT [7]				36.94	13.51	39.54	10.85
	RM [2]				35.18	16.76	41.96	13.58
	LVT [52]		Transformer	9.0 M	41.78	19.46	46.15	17.34
	Dytox [16]			10.7 M	70.12	41.76	61.71	19.14
500	iCARL [40]	Rehearsal	ResNet 18	11.2 M	36.89	16.44	35.89	10.69
	FDR [4]				42.60	11.78	49.91	10.58
	DER++ [6]				35.46	14.52	51.90	19.33
	ERT [7]				41.56	20.42	50.87	12.13
	RM [2]				38.66	14.56	52.08	18.96
	LVT [52]		Transformer	9.0 M	47.84	26.32	57.93	23.97
	Dytox [16]			10.7 M	73.64	40.94	**64.29**	26.39
–	EFT [51]	Dynamic Arch	ResNet 18	4.9 M (32k)	72.18	32.98	60.00	24.08
	PASS [60]	Regularization	ResNet 18	11.2 M	39.9	34.52	43.9	22.76
	GPM [43]	Regularization	AlexNet	6.7 M	40.65	–	45.48	–
	ConTraCon	Dynamic Arch	Transformer	**3.6 M (28k)**	**76.78**	**42.2**	62.76	**27.46**

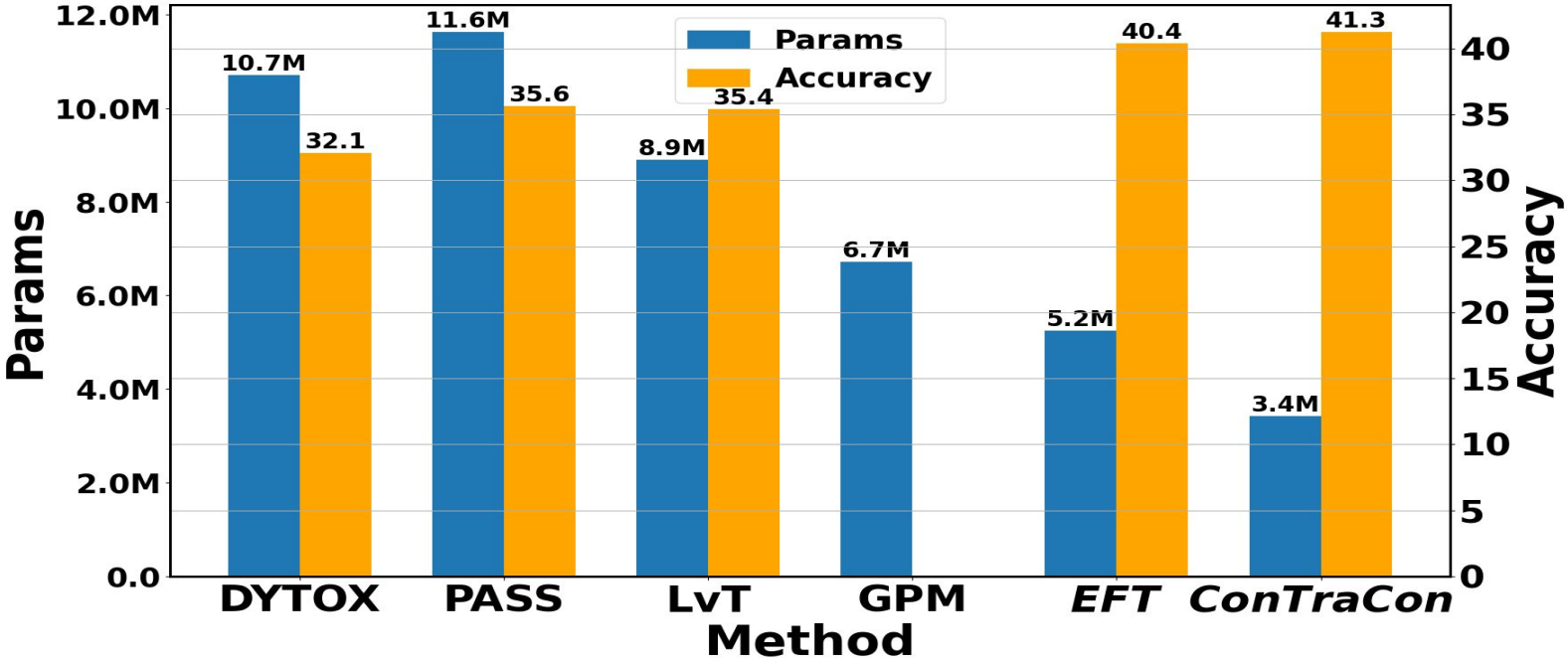
Classification accuracy on 10-task setup ofImagenet-100 and TinyImagenet-200 dataset.

Results

Model	Approach	Backbone	# Params	5-Datasets	
				TIL	CIL
Dytox [16](500)	Rehearsal	Transformer	10.7 M	77.12	**67.13**
Dytox [16](200)	Rehearsal	Transformer	10.7 M	75.81	65.04
EFT [51]	Dynamic Arch	ResNet 18	4.9 M (32k)	94.75	52.04
GPM [43]	Regularization	ResNet18	**1.2 M**	90.60	–
ConTraCon	Dynamic Arch	Transformer	3.9 M (28k)	**95.10**	65.21

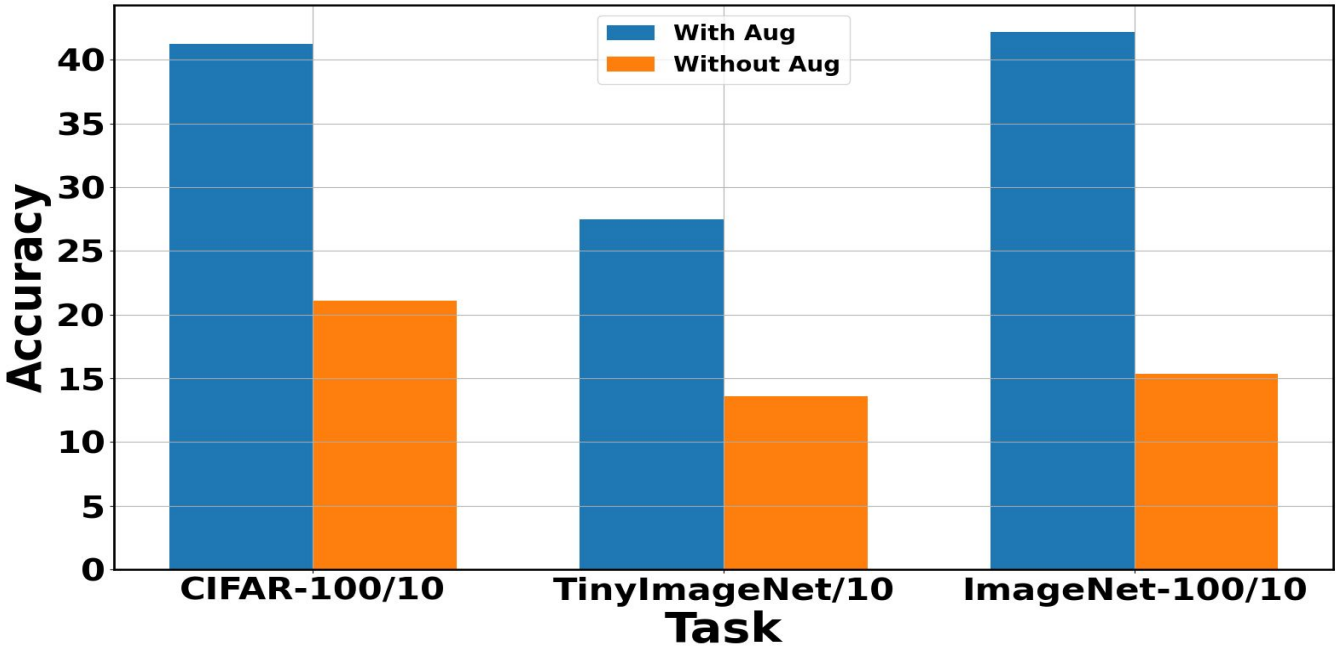
Classification accuracy on 5-datasets.

Parameters and Accuracy



Parameter and Accuracy bars of various approaches, in the 10-task setup of CIFAR-100 dataset.

Importance of Augmentation



Classification accuracy in CIL setup with and without Augmentation on 10 task setup.

Conclusion

- Proposed a novel method of adaptation to new tasks using convolution on the MHSA weights of the transformer – ConTraCon.

- Adopted an image augmentation and entropy based task-id prediction method thereby removing the need for task ids during inference.

- Performed extensive experimentation and ablation of our proposed approach.

THANK YOU!!