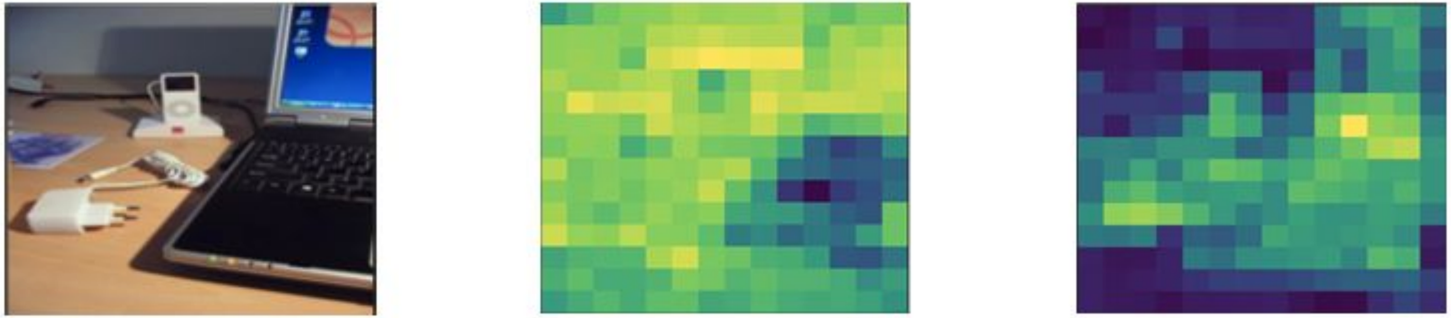
Mitigating the effect of Incidental correlations on part-based learning

Gaurav Bhatt*[1,3], Deepayan Das[2], Leonid Sigal[1,3], Vineeth N Balasubramanian [2]

[1]The University of British Columbia, [2]Indian Institute of Technology Hyderabad
[3] The Vector Institute, Canada

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

The problem of Incidental Correlations



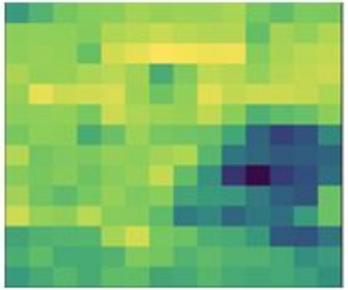
(a) Input image (b) ViT with parts (c) Proposed - DPViT

- Some specific configuration or background could dominate the training data:
  - This could lead towards bias towards those configurations.
- These configurations may not be spurious or anti-causal:
  - They provide relevant context for identifying parts.

Effect of Incidental correlations on Part-learners

- Reduces interpretability of learned parts.
- Reduces generalization of part representations.

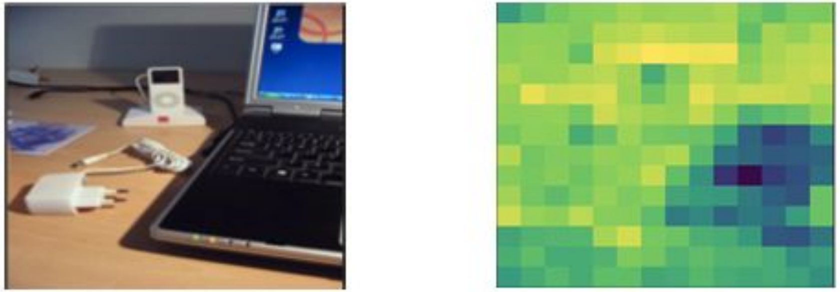
(a) Input image



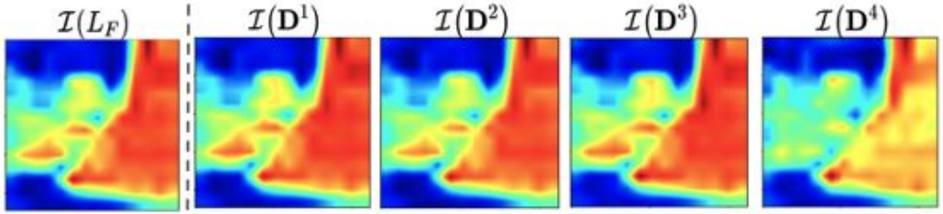
(b) ViT with parts

Effect of Incidental correlations on Part-learners

- Reduces interpretability of learned parts.
- Reduces generalization of part representations.



(a) Input image (b) ViT with parts



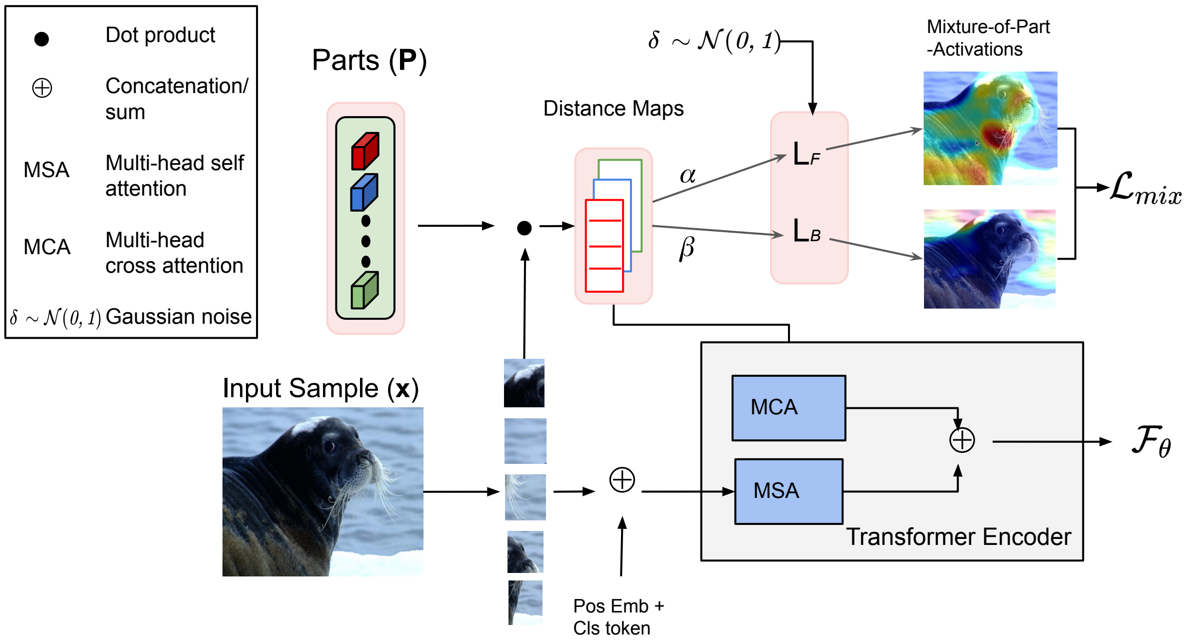
(a) Visualization of learned parts

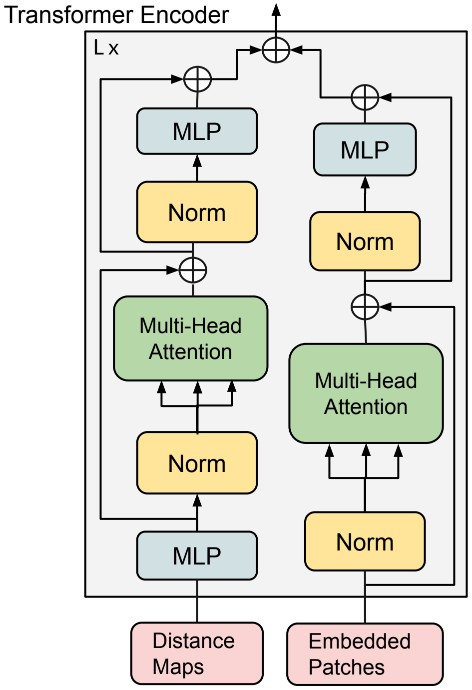
- Degeneracy of parts on a common solution.
- Less diversity among the learned part representations.

Limitations of existing works

- Current SOTA part-learners suffers from the problem of incidental correlations:
  - [1] Concept Vision Transformers (CViT), ICLR 2022
  - [2] CORL, WACV 2023
  - [3] ConstellationNet, ICLR 2021
- Does not enforce strict regularization to enforce diversity among the parts:
  - [4] CompoNet, ICCV 2019
  - [5] TUSK, ICCV 2021

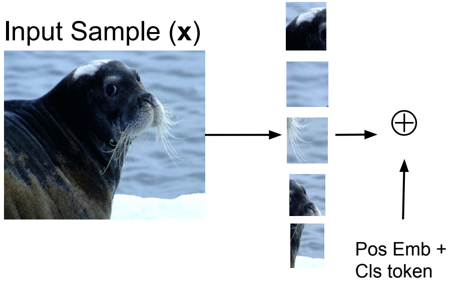
Our method: DPViT (Pretraining phase)



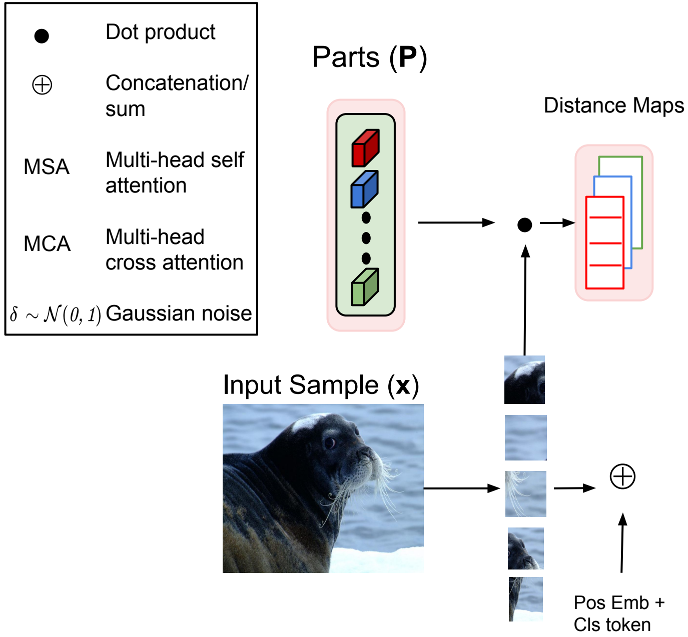


DPViT : Patch generation from the input image

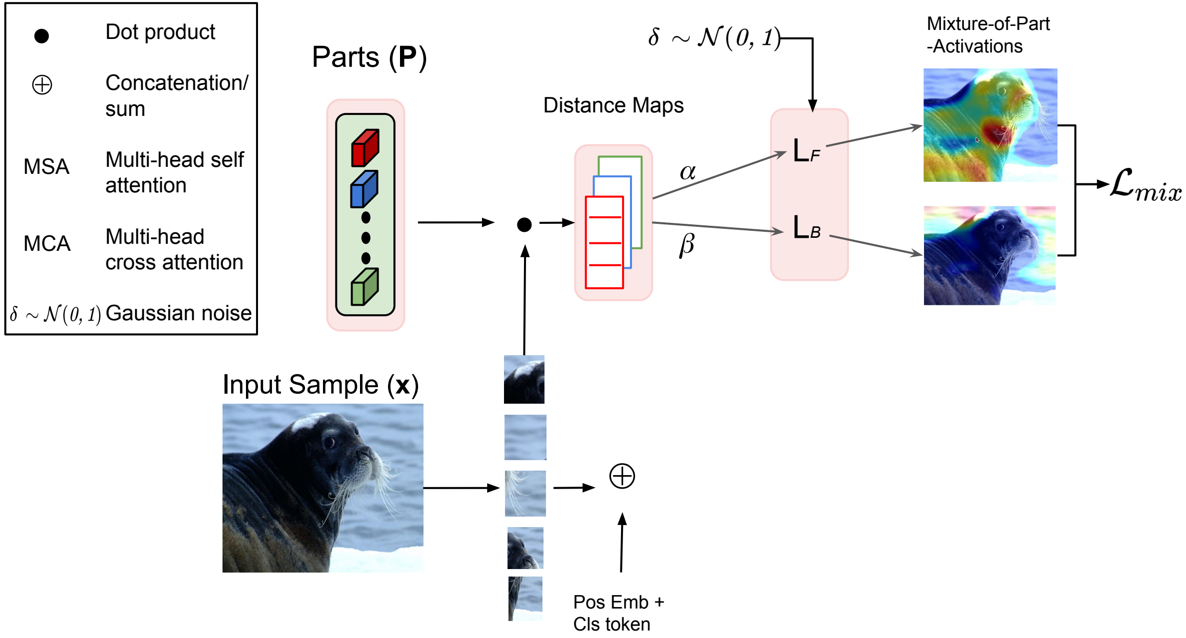
$\bullet$	Dot product
$\oplus$	Concatenation/ sum
MSA	Multi-head self attention
MCA	Multi-head cross attention
$\delta \sim \mathcal{N}(0, 1)$	Gaussian noise



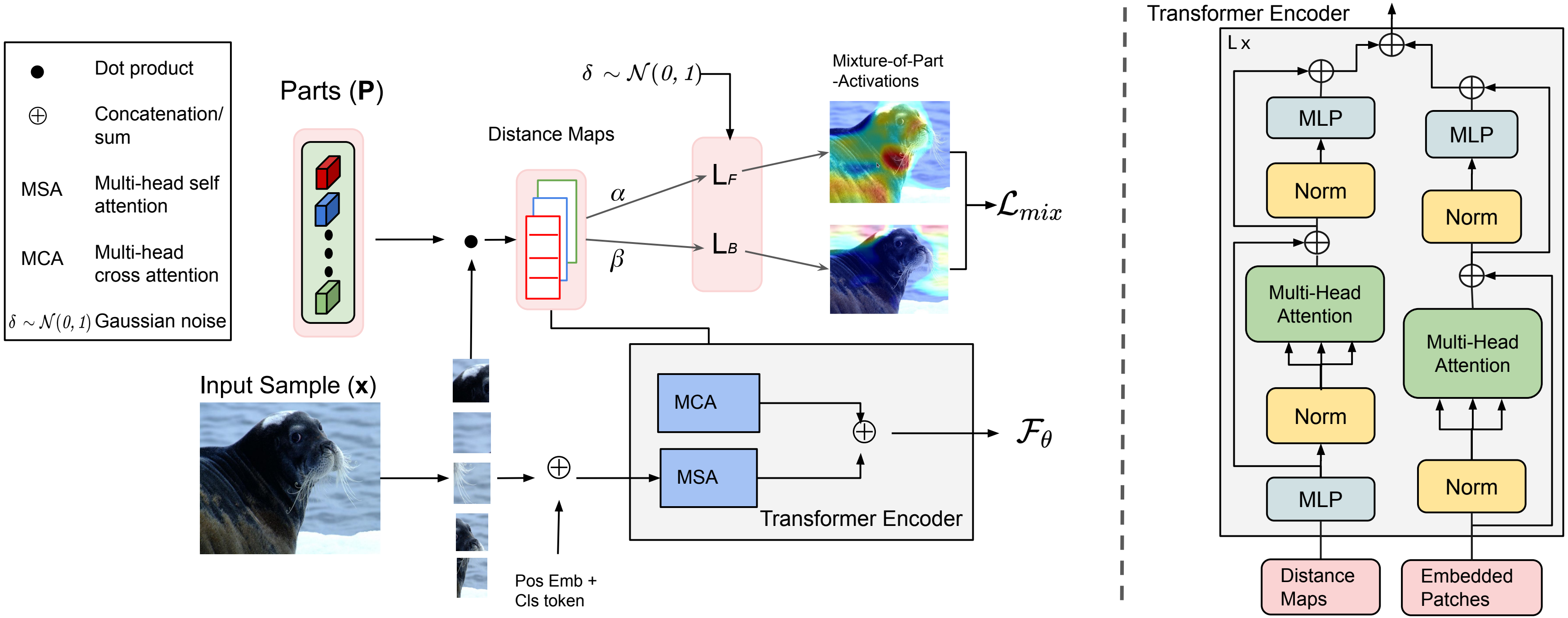
DPViT : Compute distance maps using randomly initialized part dictionary



DPViT : Compute $\mathcal{L}_{mix}$ *(mixture-of-parts)* using L_F & L_B



DPViT : Use MSA and MCA layers to form transformer encoder



DPViT pretraining : Quality assurance regularization

- Construct foreground and background latent variables to form mixture-of-parts

$$L_F = \sum_{k \in n_f} \alpha_k \mathbf{D}^k + \delta_f; L_B = \sum_{k \in n_b} \beta_k \mathbf{D}^k + \delta_b$$

DPViT pretraining : Quality assurance regularization

- Construct foreground and background latent variables to form mixture-of-parts

$$L_F = \sum_{k \in n_f} \alpha_k \mathbf{D}^k + \delta_f ; L_B = \sum_{k \in n_b} \beta_k \mathbf{D}^k + \delta_b$$

- Compute the mixture loss on weakly-supervised foreground-background masks

$$\mathcal{L}_{mix} = ||\mathcal{I}(L_F) - \mathcal{M}_f||_2 + ||\mathcal{I}(L_B) - \mathcal{M}_b||_2$$

DPViT pretraining : Quality assurance regularization

- Construct foreground and background latent variables to form mixture-of-parts

$$L_F = \sum_{k \in n_f} \alpha_k \mathbf{D}^k + \delta_f ; L_B = \sum_{k \in n_b} \beta_k \mathbf{D}^k + \delta_b$$

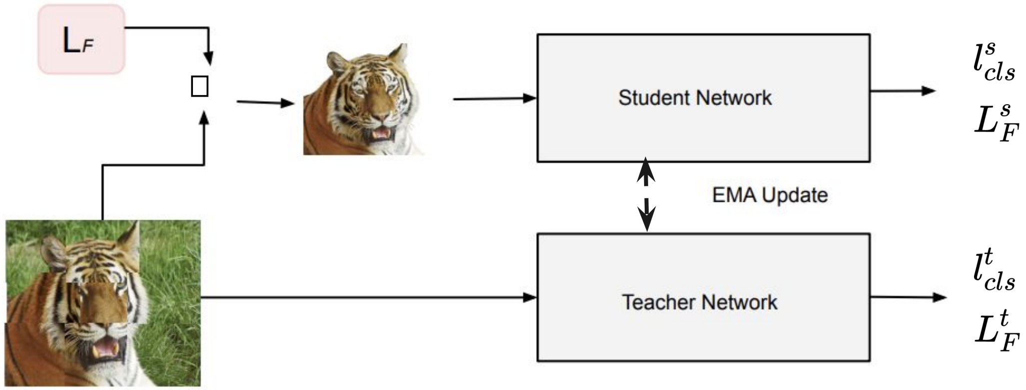
- Compute the mixture loss on weakly-supervised foreground-background masks

$$\mathcal{L}_{mix} = ||\mathcal{I}(L_F) - \mathcal{M}_f||_2 + ||\mathcal{I}(L_B) - \mathcal{M}_b||_2$$

- Enforce sparsity on parts $(\mathbf{P})$, while orthogonal spectral norm on $\mathbf{P_F}$ & $\mathbf{P_B}$

$$\mathcal{L}_Q(\lambda_s, \lambda_o) = \lambda_s ||\mathbf{P}||_1 + \lambda_o \Big[\sigma(\mathbf{P_F} \cdot \mathbf{P_F}^T - \mathbf{I}) + \sigma(\mathbf{P_B} \cdot \mathbf{P_B}^T - \mathbf{I}) \Big]$$

DPViT: Background Invariant fine-tuning phase



Invariant Feature Learning

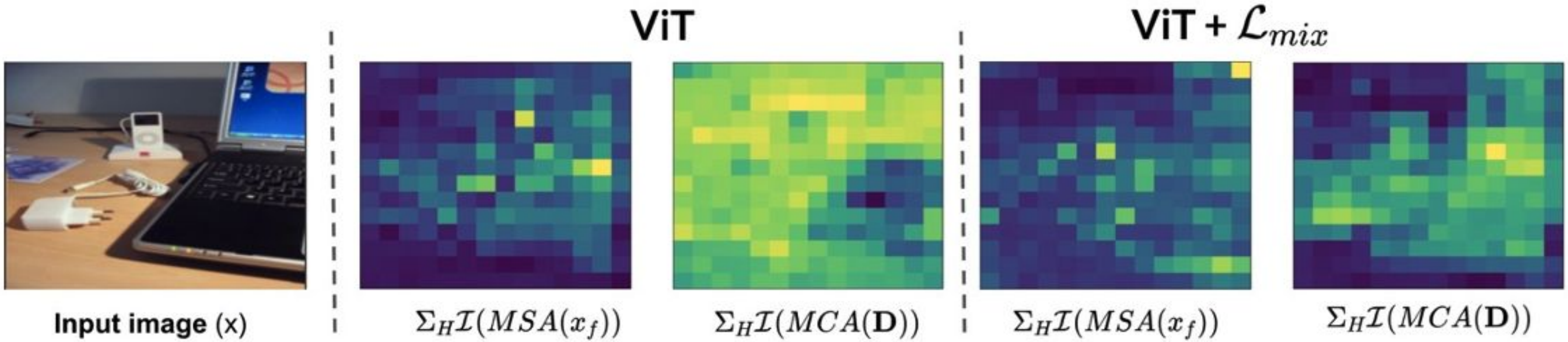
$$\mathcal{L}^{inv}_{cls} = \mathcal{L}_{ce}(\mathcal{F}^t_\phi(\mathcal{F}^t_\theta(x)), \mathcal{F}^s_\phi(\mathcal{F}^s_\theta(x_f)))$$

Invariant Parts Learning

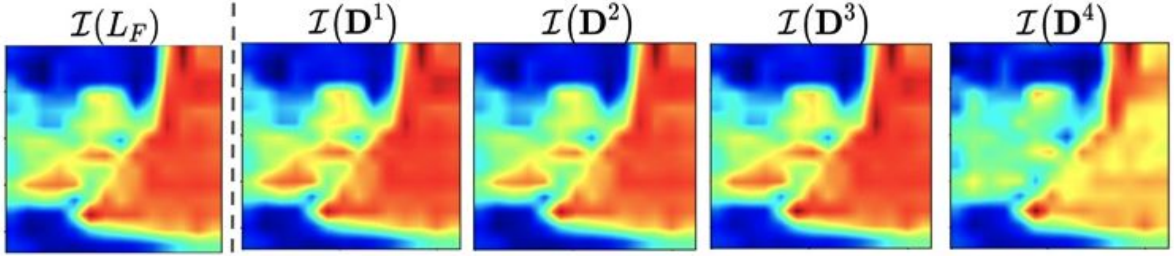
$$\mathcal{L}^{inv}_{p} = \mathcal{L}_{ce}(L^t_F(x), L^s_F(x_f))$$

Experiments, Results and Discussion

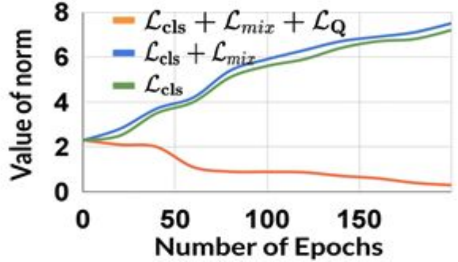
How do incidental correlations affect interpretability of part learners?



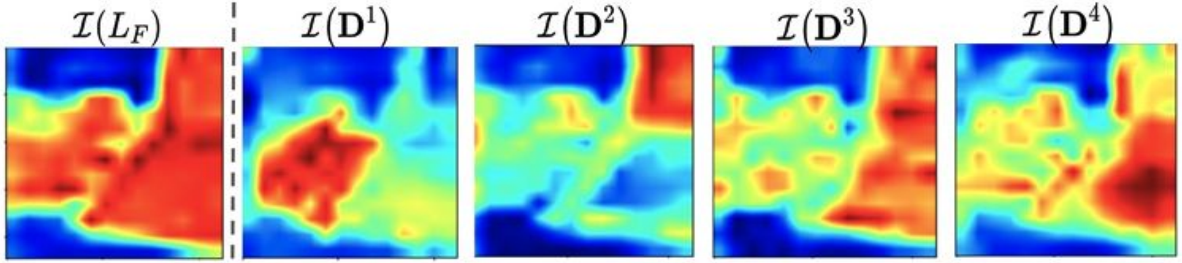
Studying the quality of learned part representations



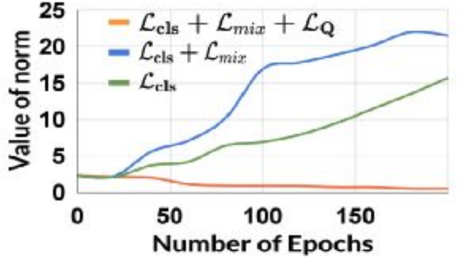
(a) Visualizing heatmaps of part projections using ViT+$\mathcal{L}_{\mathbf{mix}}$



(c) Computing $||\mathbf{P}||_1$



(b) Visualizing heatmaps of part projections using ViT+$\mathcal{L}_{\mathbf{mix}} + \mathcal{L}_Q$



(d) Computing $||\mathbf{P}\mathbf{P}^{\mathbf{T}} - \mathbf{I}||_1$

Generalization to limited data: Few-shot learning

Method		Backbone	MiniImageNet 1-shot	MiniImageNet 5-shot	TieredImageNet 1-shot	TieredImageNet 5-shot	FC100 1-shot	FC100 5-shot
ProtoNets (2017) [47]	Non-Parts	ResNet12	$60.39_{\pm0.16}$	$78.53_{\pm0.25}$	$65.65_{\pm0.92}$	$83.40_{\pm0.65}$	$37.50_{\pm0.60}$	$52.50_{\pm0.60}$
DeepEMD v2 (2020) [57]	Non-Parts	ResNet12	$68.77_{\pm0.29}$	$84.13_{\pm0.53}$	$71.16_{\pm0.87}$	$86.03_{\pm0.58}$	$46.47_{\pm0.26}$	$63.22_{\pm0.71}$
COSOC (2021) [32]	Non-Parts	ResNet12	$69.28_{\pm0.49}$	$85.16_{\pm0.42}$	$73.57_{\pm0.43}$	$87.57_{\pm0.10}$	-	-
MixtFSL (2021) [1]	Non-Parts	ResNet12	$63.98_{\pm0.79}$	$82.04_{\pm0.49}$	$70.97_{\pm1.03}$	$86.16_{\pm0.67}$	-	-
Match-feat (2022) [2]	Non-Parts	ResNet12	$68.32_{\pm0.62}$	$82.71_{\pm0.46}$	$71.22_{\pm0.86}$	$85.43_{\pm0.55}$	-	-
Label-Halluc (2022) [24]	Non-Parts	ResNet12	$67.04_{\pm0.70}$	$85.87_{\pm0.48}$	$71.97_{\pm0.89}$	$86.80_{\pm0.58}$	$47.37_{\pm0.70}$	$67.92_{\pm0.70}$
FeLMi (2022) [44]	Non-Parts	ResNet12	$67.47_{\pm0.78}$	$86.08_{\pm0.44}$	$71.63_{\pm0.89}$	$87.07_{\pm0.55}$	$49.02_{\pm0.70}$	$68.68_{\pm0.70}$
SUN (2022) [10]	Non-Parts	VIT	$67.80_{\pm0.45}$	$83.25_{\pm0.30}$	$72.99_{\pm0.50}$	$86.74_{\pm0.33}$	-	-
FewTure (2022) [23]	Non-Parts	Swin-Tiny	$72.40_{\pm0.78}$	$86.38_{\pm0.49}$	$76.32_{\pm0.87}$	$89.96_{\pm0.55}$	$47.68_{\pm0.78}$	$63.81_{\pm0.75}$
HCTransformer (2022) [22]	Non-Parts	3× VIT-S	$\mathbf{74.74}_{\pm0.17}$	$89.19_{\pm0.13}$	$\mathbf{79.67}_{\pm0.20}$	$91.72_{\pm0.11}$	$48.27_{\pm0.15}$	$66.42_{\pm0.16}$
SMKD (2023) [30]	Non-Parts	VIT-S	$74.28_{\pm0.18}$	$88.82_{\pm0.09}$	$78.83_{\pm0.20}$	$91.02_{\pm0.12}$	$50.38_{\pm0.16}$	$68.37_{\pm0.16}$
ConstNet (2021) [54]	Parts	ResNet12	$64.89_{\pm0.23}$	$79.95_{\pm0.17}$	$70.15_{\pm0.76}$	$86.10_{\pm0.70}$	$43.80_{\pm0.20}$	$59.70_{\pm0.20}$
TPMN (2021) [52]	Parts	ResNet12	$67.64_{\pm0.63}$	$83.44_{\pm0.43}$	$72.24_{\pm0.70}$	$86.55_{\pm0.63}$	$46.93_{\pm0.71}$	$63.26_{\pm0.74}$
CORL (2023) [21]	Parts	ResNet12	$65.74_{\pm0.53}$	$83.03_{\pm0.33}$	$73.82_{\pm0.58}$	$86.76_{\pm0.52}$	$44.82_{\pm0.73}$	$61.31_{\pm0.54}$
VIT-with-parts (L_{cls})	Parts	VIT-S	$72.15_{\pm0.20}$	$87.61_{\pm0.15}$	$78.03_{\pm0.19}$	$89.08_{\pm0.19}$	$48.92_{\pm0.13}$	$67.75_{\pm0.15}$
Ours - DPViT	Parts	VIT-S	$73.81_{\pm0.45}$	$\mathbf{89.85}_{\pm0.35}$	$79.32_{\pm0.48}$	$\mathbf{91.92}_{\pm0.40}$	$\mathbf{50.75}_{\pm0.23}$	$\mathbf{68.80}_{\pm0.45}$

Studying impact of incidental correlations on IN9 benchmark

(a) Original (b) MIXED-SAME (c) MIXED-RAND

Figure 8: Visualizing the test splits from ImageNet-9 dataset.

Method	IN-9L ↑	Original ↑	M-SAME ↑	M-RAND ↑	BG-GAP ↓
ResNet-50 [53]	94.6	96.3	89.9	75.6	14.3
WRN-50×2 [53]	95.2	97.2	90.6	78.0	12.6
ConstNet	90.6	92.7	86.1	69.2	17.1
ViT-S pre [11]	82.5	84.9	72.2	50.3	21.9
CT [41]	84.7	85.5	73.1	51.5	21.6
VIT-with-parts	95.1	97.2	91.5	81.7	9.8
Ours - DPViT	**96.9**	**98.5**	**93.4**	**87.5**	**5.9**

Table 2: Performance evaluation on domain shift of varying background and common data corruptions on ImageNet-9. Evaluation metric is Accuracy %.

Conclusion and future work

- Dependent on weakly supervised off-the-shelf foreground extractor to guide the training.
    - Could be challenging to train in problem-specific datasets sometimes found in medical disease domain.

- DPViT does not consider the relationship among the parts.
    - Relationship among the parts could results in interesting properties useful for tasks such as scene graph generation.

Acknowledgements:

VECTOR INSTITUTE

THE UNIVERSITY OF BRITISH COLUMBIA

compute canada

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

References

Open Review https://openreview.net/forum?id=8Xn3D9OtqI

GitHub https://github.com/GauravBh1010tt/DPViT.git

[1] Rigotti, Mattia, et al. "Attention-based interpretability with concept transformers." *ICLR*. 2022.

[2] He, Ju, Adam Kortylewski, and Alan Yuille. "CORL: Compositional representation learning for few-shot classification." *WACV*. 2023.

[3] Xu, Weijian, Huaijin Wang, and Zhuowen Tu. "Attentional constellation nets for few-shot learning." *ICLR* 2021.

[4] Tokmakov, Pavel, Yu-Xiong Wang, and Martial Hebert. "Learning compositional representations for few-shot recognition." *ICCV.* 2019.

[5] Wu, Jiamin, et al. "Task-aware part mining network for few-shot learning." *ICCV.* 2021.