[Paper](#) [GitHub](#)

Label-Focused Inductive Bias over Latent Object Features in Visual Classification

Ilmin Kang[1], HyounYoung Bae[2], Kangil Kim[*]

Intelligence Representation &Reasoning Lab

ICLR

[Problem]

Preference in Neural Networks in Image Classification

- In most well-known neural networks in image classification
  - Models learn features based on visual similarity and differentiation depending on their classes
  - Preference for the input-domain(=visual information) to handle the similarity of features is a common property
  - Neural networks learn relation over the latent objects based on the preference for the input domain

Geirhos et al.2018, Park & Kim.2021, Naseer et al.2021

[Problem]

Undescribed World Knowledge (UWK)

- However, relation on the latent objects learned by neural networks may be different in human labeling based on the world knowledge

Visually similar, but semantically unrelated samples

- Regarded as ambiguous samples in visual classification
- But, we can simply differentiate them by using unobserved relations in the data

- Undescribed World Knowledge (UWK)
  - Undescribed relations over internal objects for determining output human labeling
  - Conflicts between input-domain information and UWK has not been discussed

[Problem]

Overview of Our Work

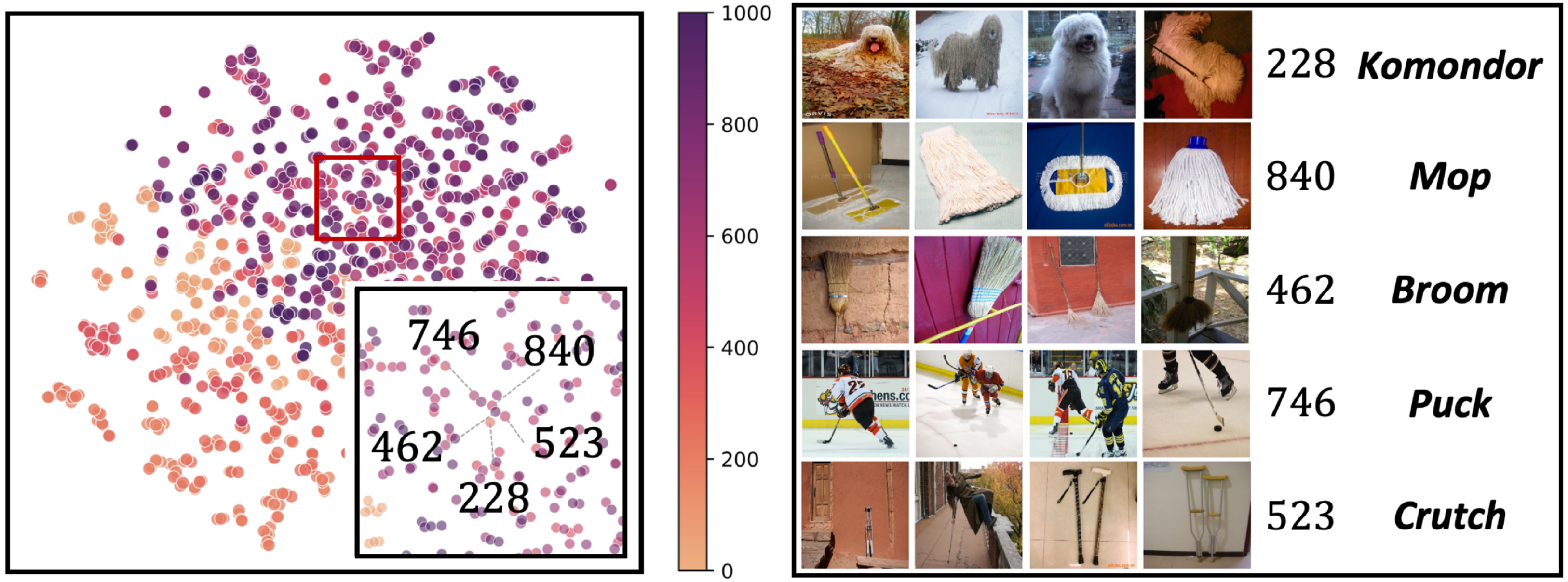
- What do we want to solve?
  - The dominance of input-domain focused inductive bias in visual neural networks
  - Conflicts between input-domain focused inductive bias and UWK limits the generalization

- How can we handle it?
  - Propose training strategy *Label-focused Latent-object Biasing (LLB)*
  - We disconnect visual dependencies
  - Learn *label- focused inductive bias* over latent object features determined solely by categorization of labels

- Strength of our LLB
  - Our LLB can be simply applied to any networks based on Transformer architectures
  - LLB shows general improvements compared to different ViT networks

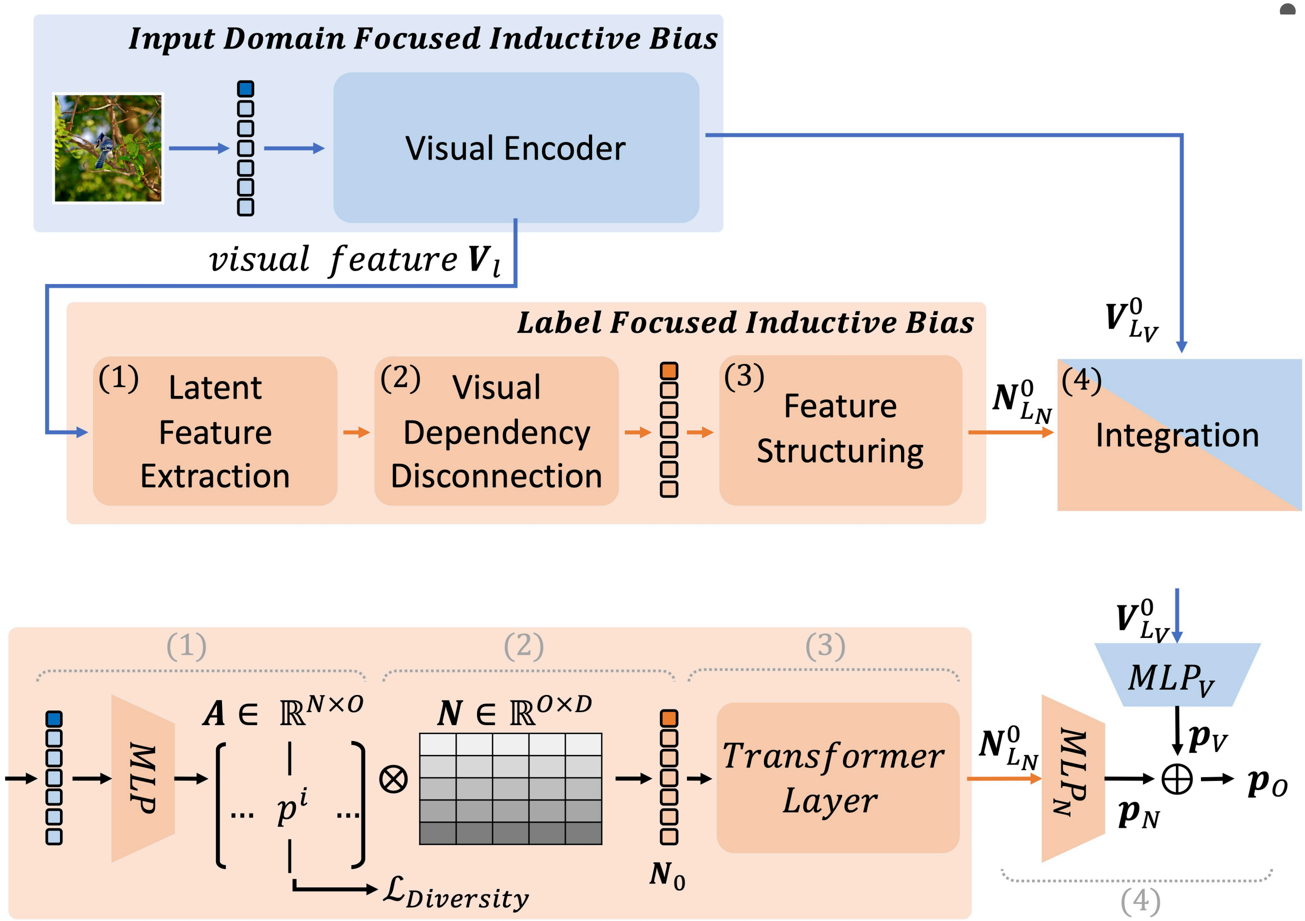
Preliminary Analysis for Problem Confirmation

- Dominance of the input-domain focused inductive bias over the UWK
  - Visually similar, but semantically unrelated class centroids are distributed close



ViT representation centroids of of Class-wise Features (IN1K classes)

Label-focused Latent-object Biasing (LLB)



- Four sequential steps

1) Latent Object Extraction from Visual Features

2) Visual Dependency Disconnection

3) Non-visual Feature Structuring

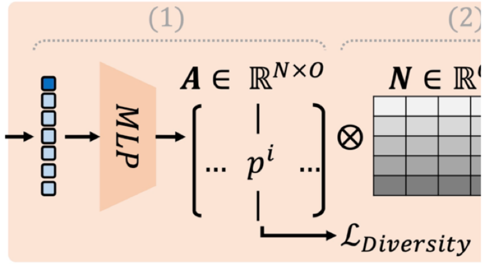
4) Integration of Non-visual and Visual Feature

[Method]

1. Latent Object Extraction from Visual Features

- Aims to quantize visual features into a set of latent objects
  - We denote visual features from ViT as $\boldsymbol{V}_l = [\boldsymbol{v}_l^i]_{i=0}^N$
  - $\boldsymbol{p}^i$ determines the probability of selecting the latent object of index i of O objects

$$\boldsymbol{p}^i = \texttt{Softmax}(\texttt{MLP}(\boldsymbol{v}_l^i)) \quad , \quad \boldsymbol{p}^i \in \mathbb{R}^O$$



  - Additional regularization term that promotes diversity of probabilities

$$\mathcal{L}_{Diversity} = -\mathcal{H}(\tilde{p}), \quad \text{where } \tilde{p} = \frac{1}{N}\sum_{i=0}^{N} \boldsymbol{p}^i \quad \text{and} \quad \mathcal{H}(p) = -\sum_{i=1}^{O} p^i \log(p^i)$$

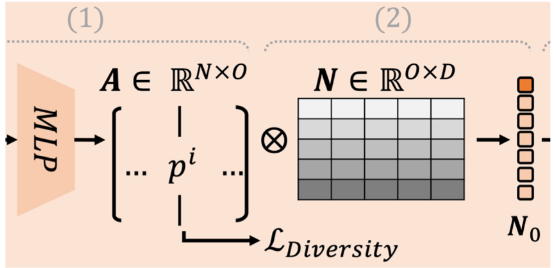
2. Visual Dependency Disconnection

- Main idea of visual disconnection is to assign separate embedding parameters to visually determined latent objects
  - Latent object is mapped to separate learnable embedding
  - Learnable embedding *Non–visual features* $N = \left[n^i\right]_{i=0}^{O} \in \mathbb{R}^{O \times D}$
  - Operates matrix multiplication of assign matrix $A = \left[p^i\right]_{i=0}^{N}$

$$\boldsymbol{N}_0 = \mathtt{Disconnect}(\boldsymbol{A}, \boldsymbol{N}) = \boldsymbol{A} \times \boldsymbol{N}$$

$\boldsymbol{A}$: A matrix of patch-wise probability vectors to select latent objects

$\boldsymbol{N}$: A matrix of non-visual features in disconnected parameters from input

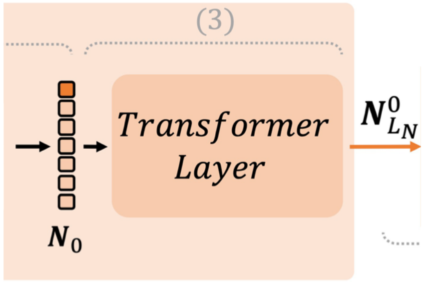


3. Non-visual Feature Structuring

- Non-visual Feature Structuring aims to redefine the similarity of features built over latent objects via solely the categorization of labels

  - We use transformer layers to extensively discern semantic relations

$$\boldsymbol{N}_{L_N} = g([\boldsymbol{c}_n || \boldsymbol{N}_0], \boldsymbol{W})$$

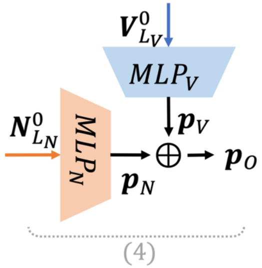


  - Outputs *structured non-visual features* N_{L_N}

4. Integration of Non-visual and Visual Feature

- Aims to leverage the strengths of both visual and non-visual features
  - We employ a separate classifier for each predictions and aggregate the output probability vectors with balancing parameter α

$$
\begin{aligned}
\boldsymbol{p}_V &= \mathtt{SoftMax}(\mathtt{MLP}_V(\boldsymbol{c}_v)) \\
\boldsymbol{p}_N &= \mathtt{SoftMax}(\mathtt{MLP}_N(\boldsymbol{c}_n)) \\
\boldsymbol{p}_O &= \alpha \times \boldsymbol{p}_V + (1-\alpha) \times \boldsymbol{p}_N
\end{aligned}
$$



[Experiment]

Quantitative Analysis Results

Model	Pre.	Resolution		Image Classification (Top1 acc.)			
		Pre.	Fine.	IN1K	IN-Real	Places365	iNat18
ViT B/16*	IN1K	224	224	$79.00_{0.00}$ (77.91)	$83.76_{0.00}$ (83.57)	-	-
+ LLB (Ours)	-	224	-	$79.43_{0.03}$	$84.25_{0.02}$	-	-
ViT B/16*	IN21K	224	224	$84.40_{0.00}$ (83.97)	$88.55_{0.00}$ (88.35)	-	-
+ LLB (Ours)	-	224	-	$84.80_{0.01}$	$88.90_{0.02}$	-	-
ViT L/16*	IN21K	224	224	$85.68_{0.00}$ (85.15)	$89.05_{0.00}$ (88.40)	-	-
+ LLB (Ours)	-	224	-	$85.92_{0.02}$	$89.26_{0.01}$	-	-
MAE B/16†	IN1K	224	224	$83.63_{0.00}$ (83.60)	$88.29_{0.00}$ (-)	$57.84_{0.07}$(57.90)	$74.20_{0.05}$ (75.40)
+ LLB (Ours)	-	224	-	$83.78_{0.02}$	$88.40_{0.02}$	$57.90_{0.06}$	$74.32_{0.06}$
MAE L/16†	IN1K	224	224	$86.08_{0.00}$ (85.90)	$89.63_{0.00}$ (-)	$59.60_{0.06}$ (59.40)	$80.06_{0.06}$ (80.10)
+ LLB (Ours)	-	224	-	$86.12_{0.01}$	$89.65_{0.02}$	$59.70_{0.05}$	$80.00_{0.06}$
SWAG B/16‡	IB3.6B	224	384	$85.28_{0.00}$ (85.30)	$89.00_{0.00}$ (89.10)	$58.90_{0.13}$ (59.10)	79.78 ($79.90_{0.06}$)
+ LLB (Ours)	-	224	-	$85.35_{0.04}$	$89.10_{0.09}$	$59.17_{0.02}$	$79.86_{0.03}$

(red: positive, blue: negative)

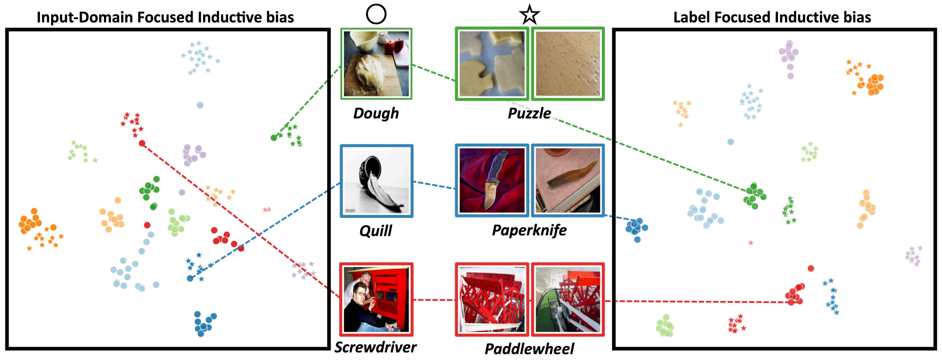
- LLB shows general improvements compared to different ViT networks across diverse data in image classification task

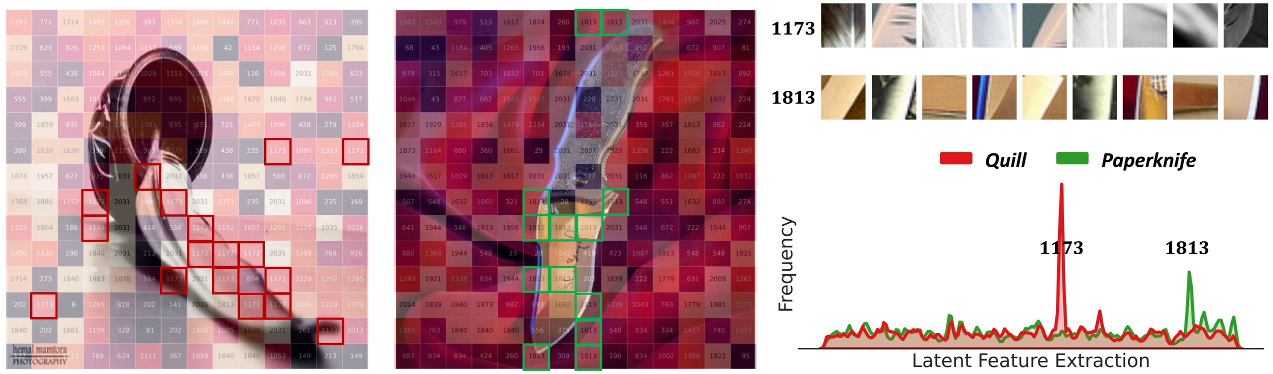
Ablation Study

Model	Visual Disc.	Diversity	w/o Pos.	Integration	IN1K (Top1-Acc_{std}%)
LLB (Ours)	✓	✓	✓	✓	$\mathbf{84.80_{0.01}}$
		✓	✓	✓	$84.25_{0.04}$
	✓		✓	✓	$84.74_{0.01}$
	✓	✓		✓	$84.77_{0.02}$
	✓	✓	✓		$82.58_{0.12}$
Baseline					$84.40_{0.00}$

- Our ablation study show that each configuration is required to obtain the best performance

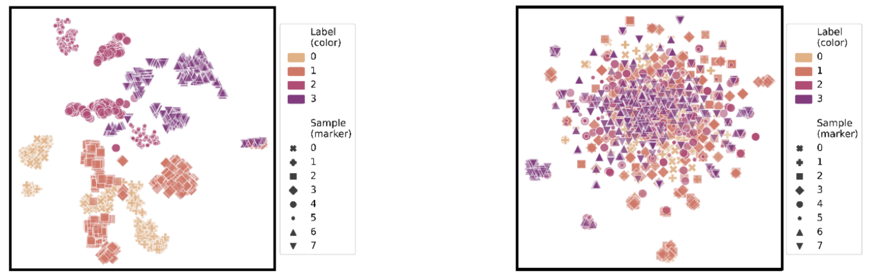
Qualitative Analysis Results





- Our LLB works as an effective identifiers for correct classification

- Our implementation
  - Effetely disconnects input-domain focused inductive bias
  - Learn *label- focused inductive bias* determined solely by categorization of labels



(a) visual feature (internal)
(b) non-visual feature (internal)
(c) visual feature (final)
(d) non-visual feature (final)

Contribution

- We raise the dominance of input-domain focused inductive bias of neural networks that conflicts with UWK
- We proposed training strategy *Label-focused Latent-object Biasing (LLB)*, to obtain UWK and utilize it as label-focused inductive bias
- We verified our method in various image classification benchmarks with quantitative and qualitative analysis

Thank You

For details of our proposal, check out our Paper & GitHub

[Paper](#) [GitHub](#)