ICLR

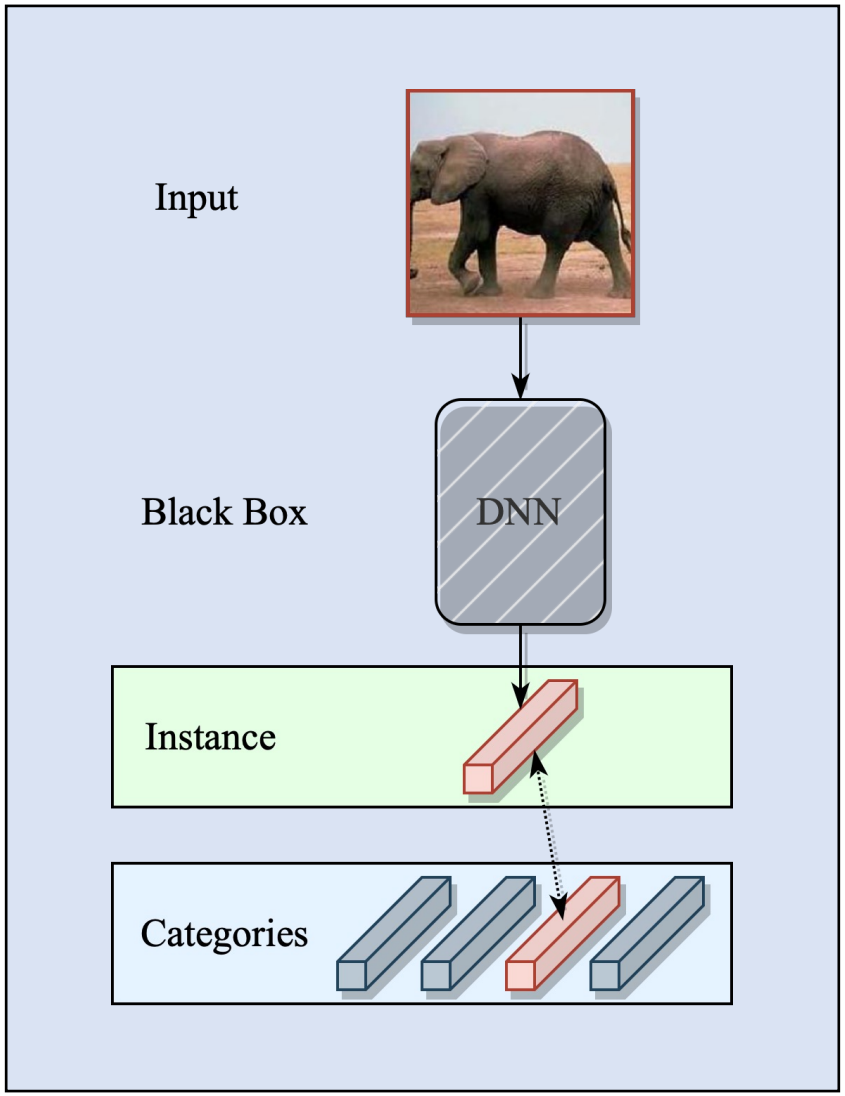
Schema Inference for Interpretable Image Classification

ICLR 2023

Haofei Zhang, Mengqi Xue, Xiaokang Liu,
Kaixuan Chen, Jie Song, and Mingli Song

Zhejiang University, Shanghai Institute for Advanced Study of Zhejiang University

Background



Canonical DL Scheme of Image Classification

Represent an input image as an instance-level embedding vector by a stack of non-linear layers (e.g., Conv, MHSA)

Perceive low-level patterns & high-level semantics

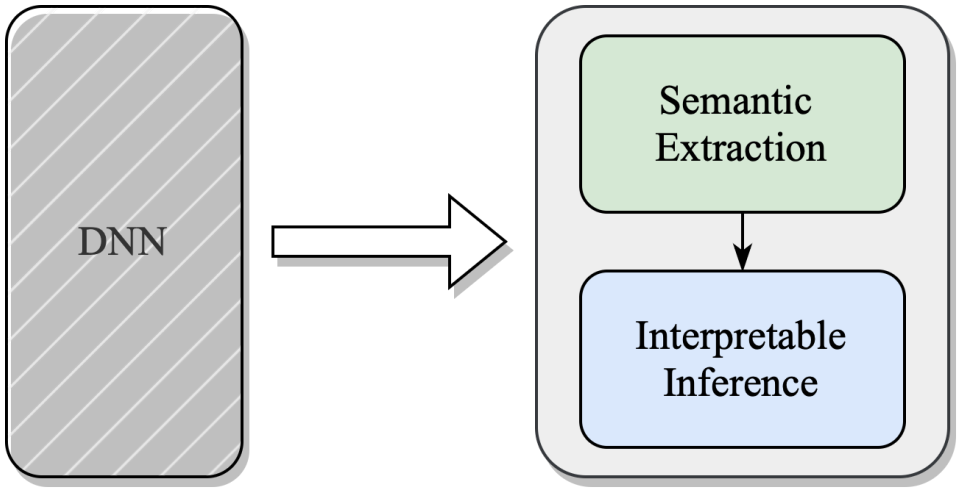
Compute inner-product similarities between the instance-level embedding and category-level embeddings (FC)

Computing procedure of visual representations

Learned category-specific embeddings

Opaque to humans

Motivation



Schema Inference

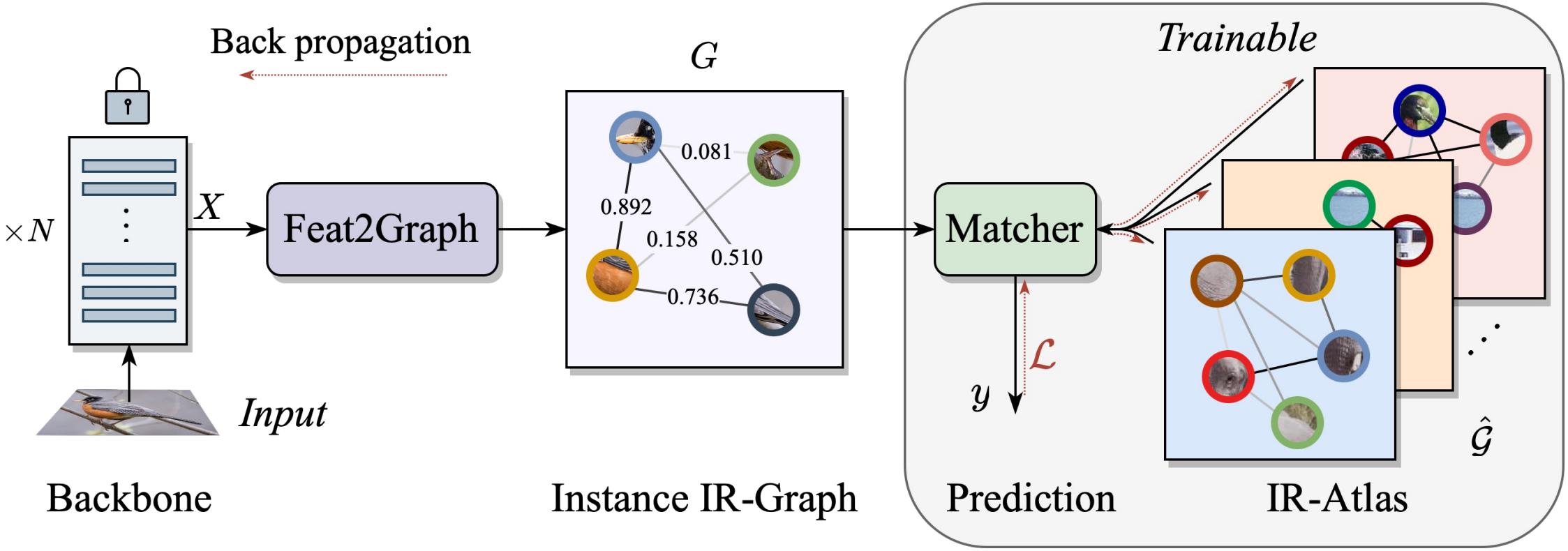
Semantic Extraction

Represent an input image as a relation graph of the semantics (i.e., instance-level IR-Graph)

Interpretable Inference

Compute graph similarities between the instance IR-Graph and category-level IR-Atlas

Method



The pipeline of our method

Method

Feat2Graph

Discretization

$$\text{Ingredient}(x) = \underset{i \in \{1,\dots,M\}}{\arg\min} \|x - \omega_i\|_2$$

$$\omega_i \in \Omega \text{ (from } k\text{-means clustering)}$$

IR-Graph

Feat2Vertex

$$\lambda_v = \alpha_1 \lambda_v^{\text{CLS}} + \alpha_2 \lambda_v^{\text{bag}}$$

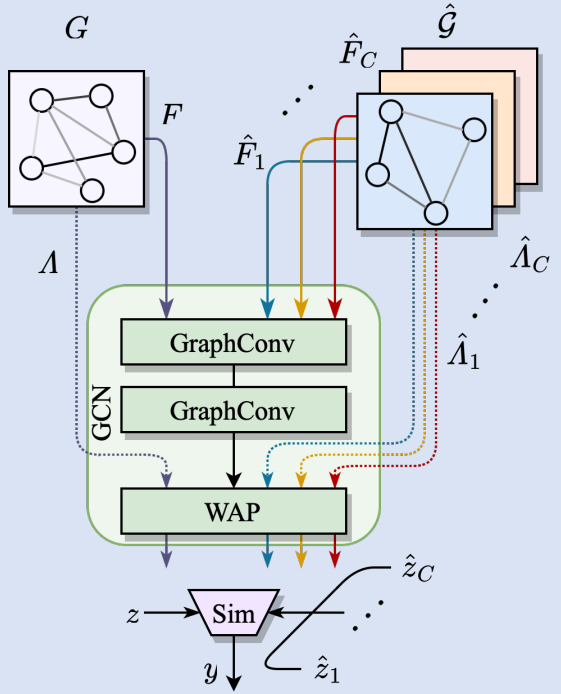
- λ_v^{CLS}: *Attention to CLS token*
- λ_v^{bag}: *Count of occurrences*

Feat2Edge

$$e_{u,v} = \beta_1 e_{u,v}^{\text{attn}} + \beta_2 e_{u,v}^{\text{adj}}$$

- $e_{u,v}^{\text{attn}}$: *Pair-wise attention*
- $e_{u,v}^{\text{adj}}$: *Adjacency*

Graph Matcher

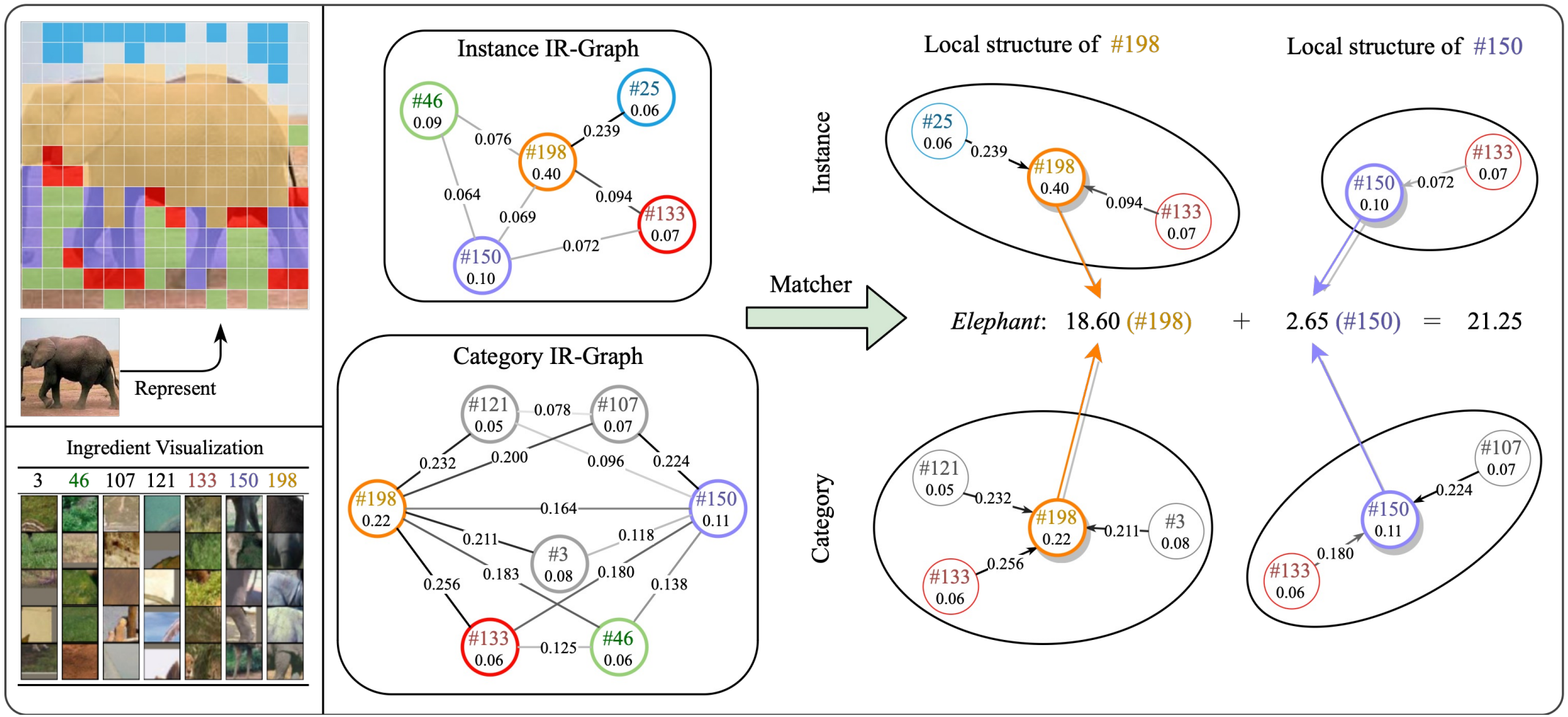


Optimization Target

$$\mathcal{L} = \mathcal{L}_{ce} + \gamma_v \mathcal{L}_v + \gamma_e \mathcal{L}_e$$

Sparsity constrains of IR-Atlas

Method



An example of how an instance IR-Graph is matched to the class imagination

Method

Interpretability of the graph matcher

Theorem 1

For a shallow GCN module, graph similarity can be approximated by

$$s = \sum_{\phi \in \Phi} \hat{\lambda}_\phi \lambda_\phi \hat{f}_\phi^{(L_G)} f_\phi^{(L_G)^\top}.$$

Particularly, if $L_G = 0$ and $\alpha_1 = 0$, our method is equivalent to BoVW with a linear classifier.

Corollary 1

*$s_{u,v}$ is significant if $\dot{\mathcal{N}}_{\hat{G}}(u)$ and $\dot{\mathcal{N}}_G(v)$ have shared vertices, relative large edge weight product $\hat{e}_{u,i} e_{v,j}$, and vertex weight product $\hat{\lambda}_u \lambda_v$, which means u and v have similar **local structure**, particularly in the case that u and v are the same vertex.*

Experiments

Backbone	Method	CIFAR-10			CIFAR-100			Caltech-101		
		Acc	#param.	FLOPs	Acc	#param.	FLOPs	Acc	#param.	FLOPs
-	BoVW-SIFT	20.20	-	-	-	-	-	-	-	-
DeiT-Tiny	Base	96.69	5.53M	1.27G	82.88	5.54M	1.27G	92.53	5.54M	1.27G
	Backbone-FC	94.58	1.93K	1.06G	76.74	19.3K	1.06G	76.71	19.5K	1.06G
	BoVW-Deep	94.95	10.3K	1.06G	75.45	103K	1.06G	60.82	104K	1.06G
	BagNet	70.27	5.73M	1.10G	42.90	5.84M	1.10G	72.54	5.85M	1.10G
	SchemaNet	95.92	396K	1.19G	77.11	105M	2.40G	83.24	106M	2.40G
	SchemaNet-Init	**95.96**	234K	1.06G	**78.45**	573K	1.06G	**87.57**	564K	1.06G
DeiT-Small	Base	97.77	21.7M	4.63G	87.65	21.7M	4.63G	94.96	21.7M	4.63G
	Backbone-FC	96.62	3.85K	3.87G	82.44	38.5K	3.87G	86.94	38.9K	3.87G
	BoVW-Deep	95.61	10.3K	3.89G	77.39	103K	3.89G	65.54	104K	3.89G
	BagNet	83.90	22.1M	3.95G	50.49	22.2M	3.95G	78.13	22.2M	3.95G
	SchemaNet	**97.42**	396K	4.00G	82.21	105M	5.21G	84.66	106M	5.21G
	SchemaNet-Init	97.35	235K	3.89G	**82.46**	591K	3.89G	**90.09**	589K	3.89G
DeiT-Base	Base	98.41	85.8M	17.6G	89.17	85.9M	17.6G	95.83	85.9M	17.6G
	Backbone-FC	97.04	7.69K	14.7G	**81.66**	76.9K	14.7G	88.20	77.7K	14.7G
	BoVW-Deep	95.50	10.3K	14.7G	72.23	103K	14.7G	66.17	104K	14.7G
	BagNet	90.71	86.6M	14.9G	67.84	86.8M	14.9G	86.55	86.8M	14.9G
	SchemaNet	**97.26**	396K	14.8G	79.26	105M	16.1G	81.12	106M	16.1G
	SchemaNet-Init	97.07	235K	14.7G	79.36	606K	14.7G	**90.72**	593K	14.7G

Experiments

Comparison results on ImageNet

Dataset	Base	Backbone-FC	BoVW-Deep	BagNet	SchemaNet	SchemaNet-Init
mini-ImageNet	95.35	90.04	90.24	85.64	89.40	**91.02**
ImageNet-1k	79.90	69.23	58.95	68.92	-	**74.05**

Adversarial attack results

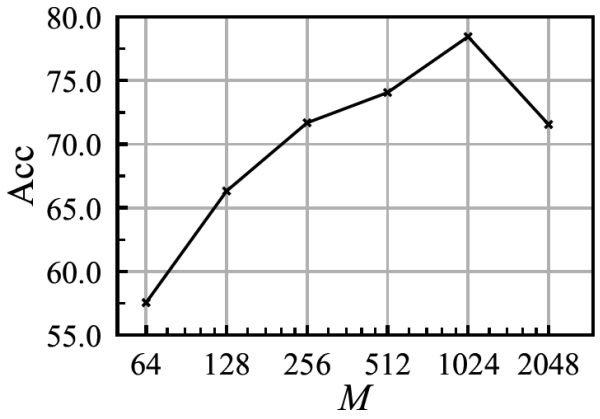
Dataset	Backbone-FC	BoVW-Deep	BagNet	SchemaNet-Init
ImageNet-A	5.41	7.36	5.53	**13.05**
ImageNet-R	31.21	24.17	30.93	**33.46**

Extending SchemaNet to unseen tasks on Caltech-101

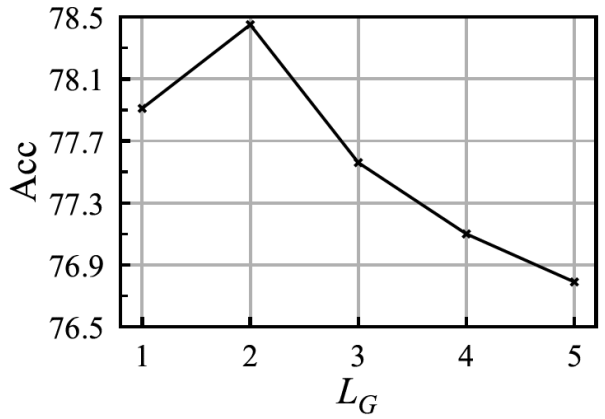
	Average	Base				
Base	97.57	97.57	**Task 1**			
+Task 1	94.68	95.95	93.69	**Task 2**		
+Task 2	93.36	95.14	93.06	92.11	**Task 3**	
+Task 3	93.63	95.55	93.06	92.47	93.75	**Task 4**
+Task 4	92.37	95.14	92.74	91.76	90.42	91.49

Experiments

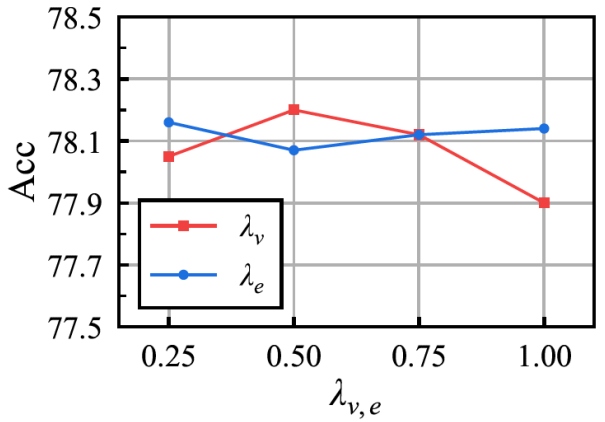
Ablation study and sensitivity analysis on Caltech-101



Visual vocabulary size



GCN layers



GCN layers

Time costing (ms) of the components in SchemaNet with input batch size of 64

Dataset	Backbone	Feat2Vertex	Feat2Edge	Matcher
CIFAR-10	4.89	7.00	19.8	1.95
CIFAR-100	4.31	9.50	101.7	1.30
Caltech-101	4.12	8.12	111.7	1.42

Conclusion

- We propose a novel inference paradigm, named schema inference towards resembling human deductive reasoning of associating the abstract concept image with the specific sense impression.

- The graph vertices are visual semantics represented by common feature vectors from DNN's intermediate layer and the edges indicate the vertex interactions characterized by semantic similarity and spatial adjacency, which facilitate capturing the compositional contributions to the predictions.

- Theoretical analysis and experimental results on several benchmarks demonstrate the superiority and interpretability of schema inference.

ICLR

Schema Inference for Interpretable Image Classification

Thank You!

VIPA
Visual Intelligence and Pattern Analysis

For more interesting works,
please visit our homepage:
https://www.vipazoo.cn