THE OHIO STATE UNIVERSITY

Hierarchical Classification with Confidence using Generalized Logits

Jim Davis[1], **Tong Liang[1]**, James Enouen[1], Roman Ilin[2]

[1]Ohio State University, Columbus OH 43210

[2]AFRL/RYAP, Wright-Patterson AFB OH 45433

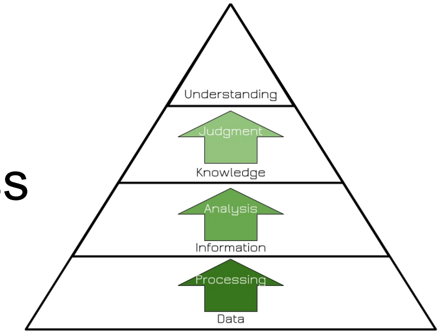
This work is supported by the U.S. Air Force Research Laboratory

CVL Computer Vision Lab

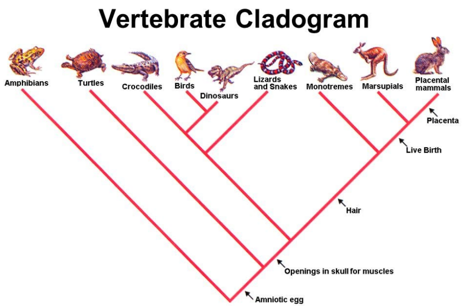
AIR FORCE RESEARCH LABORATORY

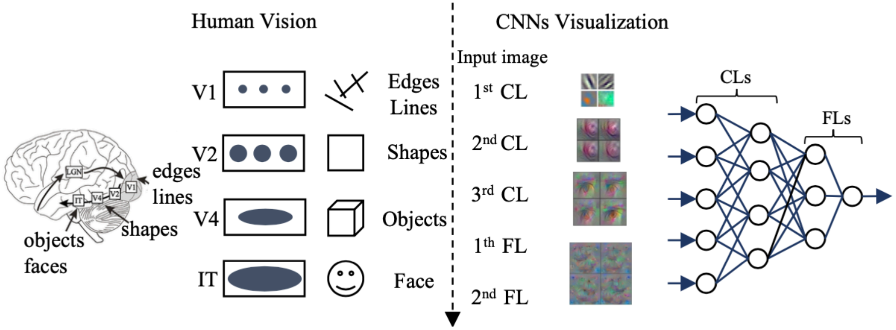
Hierarchical Reasonings

- Hierarchical reasoning prevails in many areas
  - Hierarchical cognitive process
  - Taxonomy in biology
  - Natural hierarchical representation of visual features in our brain



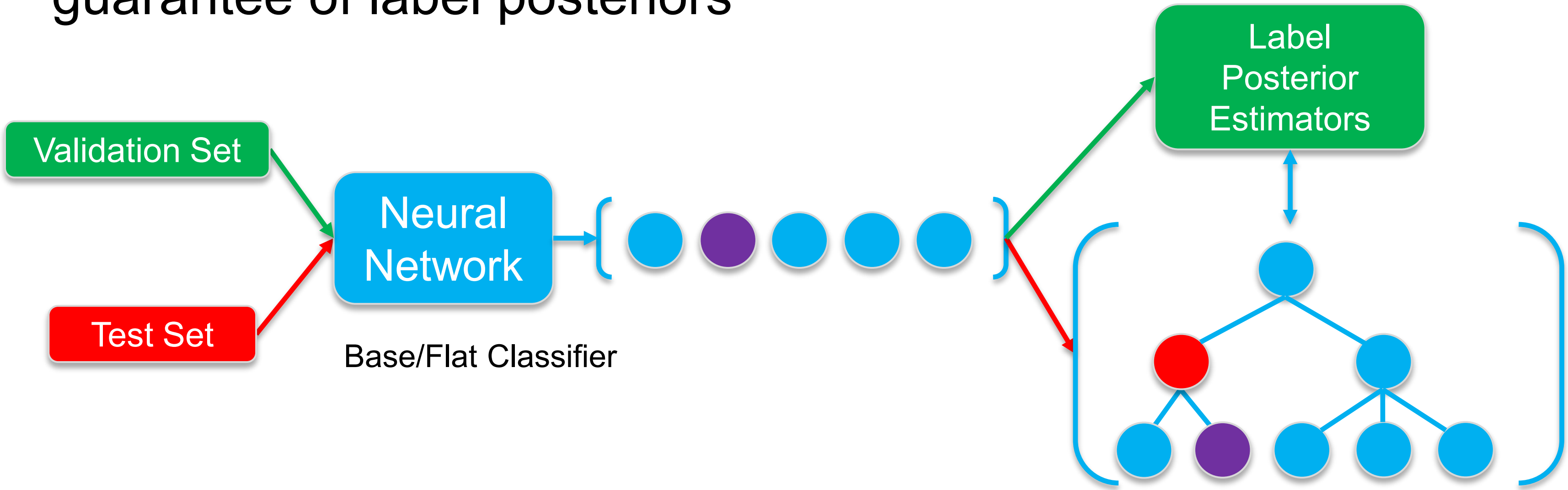
Cognitive Hierarchy





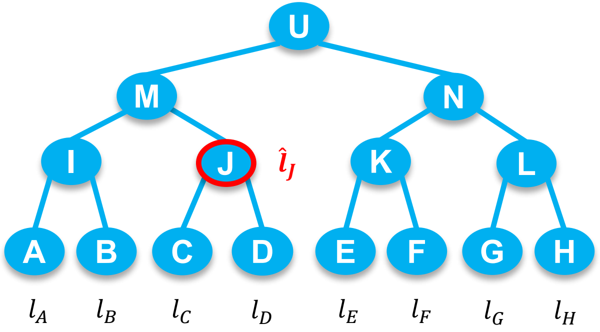
Overview of Our Approach

- Serves as a post processing step after a given base model
- Estimates label posteriors from base model logits output of the validation set
- Produces hierarchical predictions to test set with statistical guarantee of label posteriors



Generalized Logits

- Generalized logit = derived logit for a non-terminal class
- Computed from base classifier's softmax value of s_J



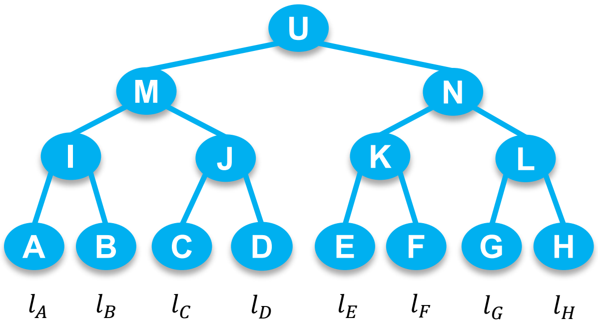
$$s_J = s_C + s_D = \frac{e^{l_C}}{e^{l_C} + \sum_{k \neq C} e^{l_k}} + \frac{e^{l_D}}{e^{l_D} + \sum_{k \neq D} e^{l_k}} = \frac{\sum_{i \in \{C,D\}} e^{l_i}}{\sum_{i \in \{C,D\}} e^{l_i} + \sum_{k \notin \{C,D\}} e^{l_k}} \triangleq \frac{e^{\hat{l}_J}}{e^{\hat{l}_J} + \sum_{k \notin J} e^{l_k}}$$

$\hat{l}_J$ is the **generalized logit** of class/node J

$$\hat{l}_J = \ln\left(e^{\hat{l}_J}\right) = \ln\Big(\sum_{i \in \{C,D\}} e^{l_i}\Big)$$

Inference

- Start with initial terminal label hypothesis
  - Option 1: Select via argmax of base classifier's logits
  - Option 2: <u>Select via argmax of estimated terminal label posteriors</u>

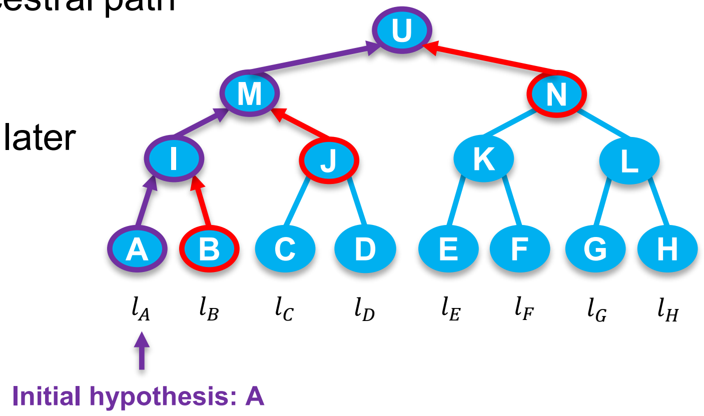


- Examine confidence of initial terminal label hypothesis
  - Is it above a given confidence threshold?
  - If YES, return that label
  - If NO, examine the remaining ancestral classes until meeting the threshold (customized)
    - Root node of the hierarchy has posterior of 1.0

Inference: An Example

- Consider the binary tree shown on the right, evaluate the ancestral path label posteriors:

$\mathrm{P}(\mathrm{A}|\mathcal{L}_A)$ $\mathcal{L}_A$ is the generalized logit vector to be introduced later

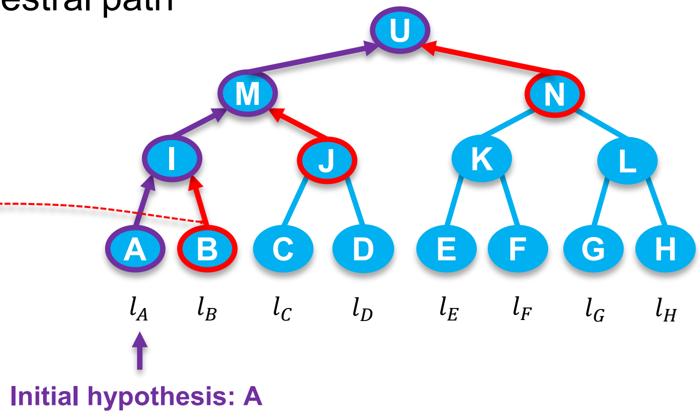


Inference: An Example

- Consider the binary tree shown on the right, evaluate the ancestral path label posteriors:

$$\mathrm{P}(\mathrm{A}|\mathcal{L}_A)$$

$$\mathrm{P}(\mathrm{I}|\mathcal{L}_A) = P(A \cup \boldsymbol{B}|\mathcal{L}_A) = P(A|\mathcal{L}_A) + \boldsymbol{P}(\boldsymbol{B}|\mathcal{L}_A)$$



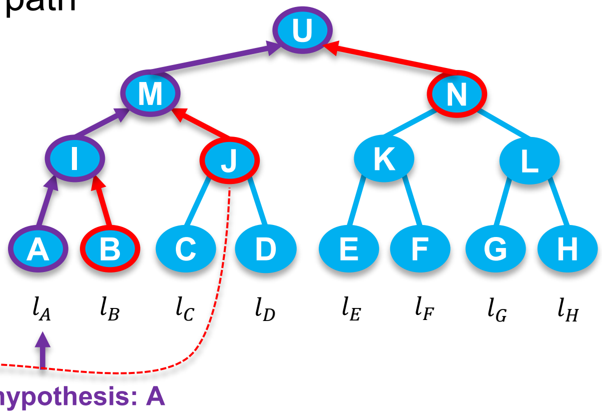
Inference: An Example

- Consider the binary tree shown on the right, evaluate the ancestral path label posteriors:

$$\mathrm{P}(\mathrm{A}|\mathcal{L}_A)$$

$$\mathrm{P}(\mathrm{I}|\mathcal{L}_A) = P(A \cup \boldsymbol{B}|\mathcal{L}_A) = P(A|\mathcal{L}_A) + \boldsymbol{P}(\boldsymbol{B}|\mathcal{L}_A)$$

$$P(M|\mathcal{L}_A) = P(I \cup \boldsymbol{J}|\mathcal{L}_A) = \mathrm{P}(\mathrm{I}|\mathcal{L}_A) + \mathbf{P}(\boldsymbol{C} \cup \boldsymbol{D}|\mathcal{L}_A)$$



Inference: An Example

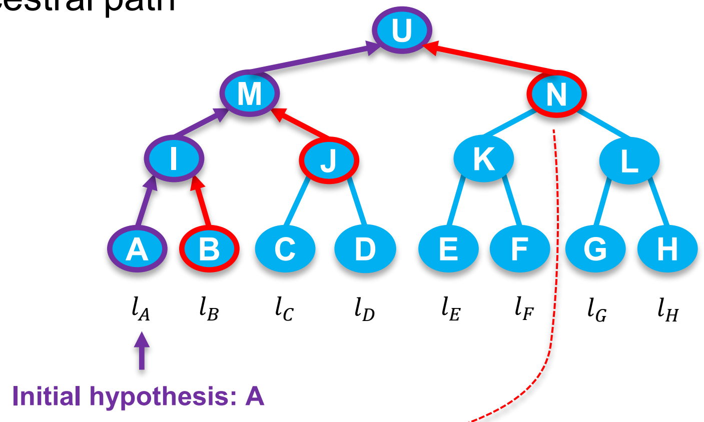
- Consider the binary tree shown on the right, evaluate the ancestral path label posteriors:

$$\mathrm{P}(\mathrm{A}|\mathcal{L}_A)$$

$$\mathrm{P}(\mathrm{I}|\mathcal{L}_A) = P(A \cup \boldsymbol{B}|\mathcal{L}_A) = P(A|\mathcal{L}_A) + \boldsymbol{P(B|\mathcal{L}_A)}$$

$$P(M|\mathcal{L}_A) = P(I \cup \boldsymbol{J}|\mathcal{L}_A) = \mathrm{P}(\mathrm{I}|\mathcal{L}_A) + \mathbf{P}(\boldsymbol{C} \cup \boldsymbol{D}|\mathcal{L}_A)$$

$$P(U|\mathcal{L}_A) = P(M \cup \boldsymbol{N}|\mathcal{L}_A) = \mathrm{P}(\mathrm{M}|\mathcal{L}_A) + \boldsymbol{P(E \cup F \cup G \cup H|\mathcal{L}_A)}$$



Inference: An Example

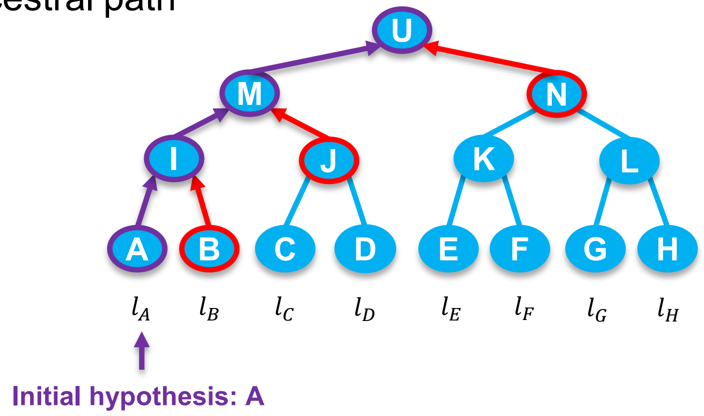
- Consider the binary tree shown on the right, evaluate the ancestral path label posteriors:

$$P(A|\mathcal{L}_A)$$

$$P(I|\mathcal{L}_A) = P(A \cup \boldsymbol{B}|\mathcal{L}_A) = P(A|\mathcal{L}_A) + \boldsymbol{P(B|\mathcal{L}_A)}$$

$$P(M|\mathcal{L}_A) = P(I \cup \boldsymbol{J}|\mathcal{L}_A) = P(I|\mathcal{L}_A) + \boldsymbol{P(C \cup D|\mathcal{L}_A)}$$

$$P(U|\mathcal{L}_A) = P(M \cup \boldsymbol{N}|\mathcal{L}_A) = P(M|\mathcal{L}_A) + \boldsymbol{P(E \cup F \cup G \cup H|\mathcal{L}_A)}$$



- L1 normalization before inference:

$$P(A|\mathcal{L}_A) + P(B|\mathcal{L}_A) + P(C \cup D|\mathcal{L}_A) + P(E \cup F \cup G \cup H|\mathcal{L}_A) \triangleq 1.0$$

Inference: An Example

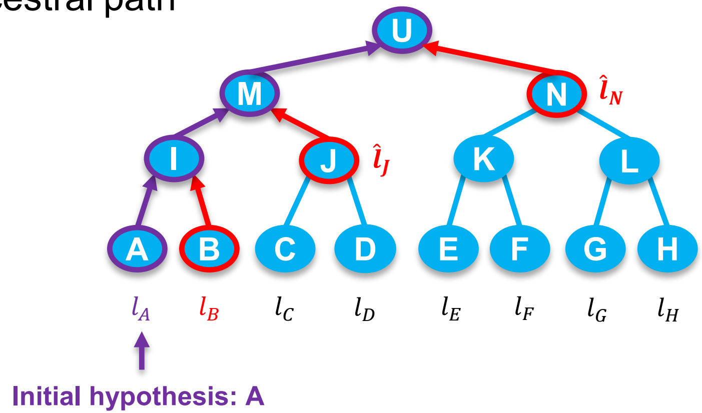
- Consider the binary tree shown on the right, evaluate the ancestral path label posteriors:

$$\mathrm{P}(\mathrm{A}|\mathcal{L}_A)$$

$$\mathrm{P}(\mathrm{I}|\mathcal{L}_A) = P(A \cup \boldsymbol{B}|\mathcal{L}_A) = P(A|\mathcal{L}_A) + \boldsymbol{P(B|\mathcal{L}_A)}$$

$$P(M|\mathcal{L}_A) = P(I \cup \boldsymbol{J}|\mathcal{L}_A) = \mathrm{P}(\mathrm{I}|\mathcal{L}_A) + \boldsymbol{P(C \cup D|\mathcal{L}_A)}$$

$$P(U|\mathcal{L}_A) = P(M \cup \boldsymbol{N}|\mathcal{L}_A) = \mathrm{P}(\mathrm{M}|\mathcal{L}_A) + \boldsymbol{P(E \cup F \cup G \cup H|\mathcal{L}_A)}$$



- L1 normalization before inference:

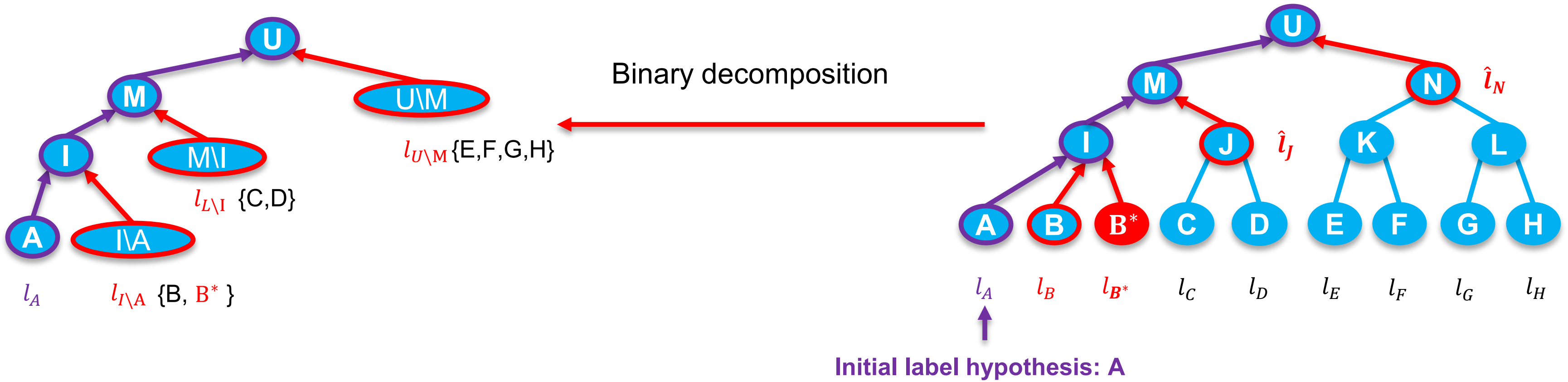
$$\mathrm{P}(\mathrm{A}|\mathcal{L}_A) + \mathrm{P}(\mathrm{B}|\mathcal{L}_A) + P(C \cup D|\mathcal{L}_A) + P(E \cup F \cup G \cup H|\mathcal{L}_A) \triangleq 1.0$$

A **B** **J** **N**

Extension to Non-Binary Tree

- Adding a node $\boldsymbol{B}^*$ under node I

$$\mathrm{P}(\mathrm{A}|\mathcal{L}_A)$$
$$\mathbf{P}(\boldsymbol{I}|\boldsymbol{\mathcal{L}_A}) = P(A|\mathcal{L}_A) + P(B \cup \boldsymbol{B}^*|\mathcal{L}_A) = P(A|\mathcal{L}_A) + P(I\backslash \mathrm{A}|\mathcal{L}_A)$$
$$\mathbf{P}(\mathbf{M}|\boldsymbol{\mathcal{L}_A}) = \mathrm{P}(\mathrm{I}|\mathcal{L}_A) + P(C \cup D|\mathcal{L}_A) = \mathrm{P}(\mathrm{I}|\mathcal{L}_A) + P(M\backslash \mathrm{I}|\mathcal{L}_A)$$
$$\mathbf{P}(\boldsymbol{U}|\boldsymbol{\mathcal{L}_A}) = \mathrm{P}(\mathrm{M}|\mathcal{L}_A) + P(E \cup F \cup G \cup H|\mathcal{L}_A) = \mathrm{P}(\mathrm{M}|\mathcal{L}_A) + P(U\backslash \mathrm{M}|\mathcal{L}_A) \triangleq 1.0$$



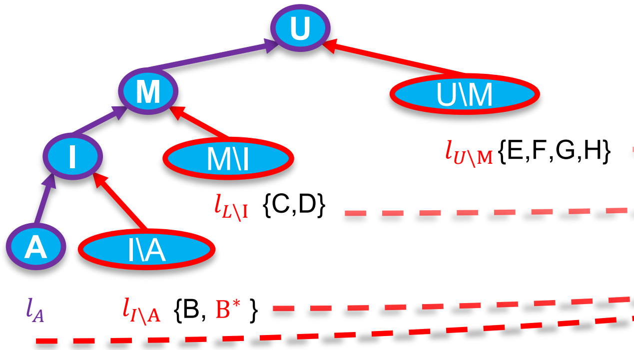
Generalized Logit Feature Vector

- Adding a node $\boldsymbol{B^*}$ under node I

$$\mathrm{P}(\mathrm{A}|\mathcal{L}_A)$$
$$\mathbf{P}(\boldsymbol{I}|\boldsymbol{\mathcal{L}_A}) = P(A|\mathcal{L}_A) + P(B \cup \boldsymbol{B}^*|\mathcal{L}_A) = P(A|\mathcal{L}_A) + P(I\backslash \mathrm{A}|\mathcal{L}_A)$$
$$\mathbf{P}(\mathbf{M}|\boldsymbol{\mathcal{L}_A}) = \mathrm{P}(\mathrm{I}|\mathcal{L}_A) + P(C \cup D|\mathcal{L}_A) = \mathrm{P}(\mathrm{I}|\mathcal{L}_A) + P(M\backslash \mathrm{I}|\mathcal{L}_A)$$
$$\mathbf{P}(\boldsymbol{U}|\boldsymbol{\mathcal{L}_A}) = \mathrm{P}(\mathrm{M}|\mathcal{L}_A) + P(E \cup F \cup G \cup H|\mathcal{L}_A) = \mathrm{P}(\mathrm{M}|\mathcal{L}_A) + P(U\backslash \mathrm{M}|\mathcal{L}_A) \triangleq 1.0$$

Generalized logit vector

$$\mathcal{L}_A = [l_{\boldsymbol{A}}, \hat{l}_{\boldsymbol{I}\backslash \boldsymbol{A}}, \hat{l}_{\boldsymbol{M}\backslash \boldsymbol{I}}, \hat{l}_{\boldsymbol{U}\backslash \mathrm{M}}]$$

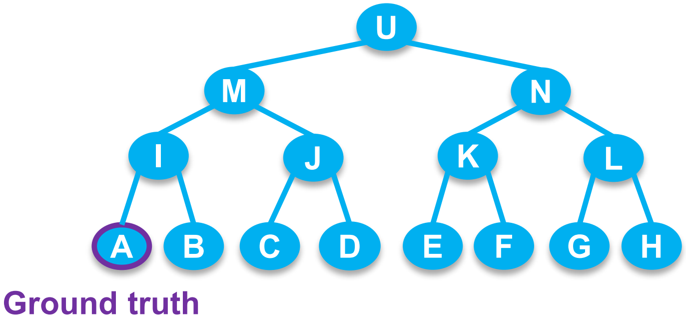


Experiments

- We conducted experiments on 4 datasets with
  - ImageNet-Animal (398 terminal classes) [Davis et al., 2019]
  - CIFAR100 (100 terminal classes) [Krizhevsky, 2009]
  - CIFAR10 (10 terminal classes) [Krizhevsky, 2009]
  - Fashion-MNIST (10 terminal classes) [Xiao et al., 2017]
- Semantic hierarchy for each dataset is derived from WordNet [Davis et al., 2019]
- Compared our method with the two most related works
  - [Deng et al., 2012] employed optimization of tradeoff between accuracy and label specificity
  - [Davis et al., 2019] proposed a non-parametric histogram binning approach

Metrics

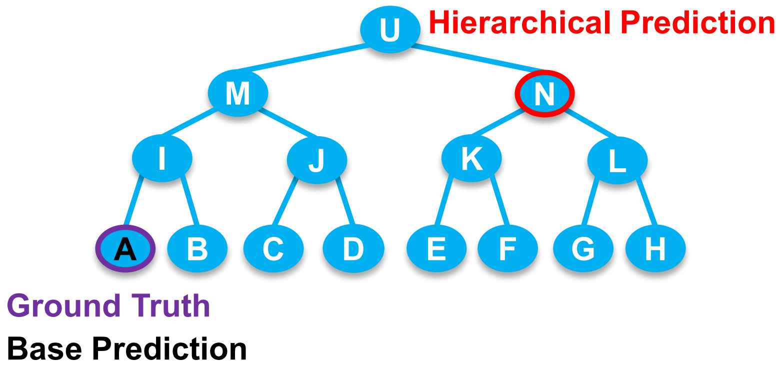
- Hierarchical Classification Metrics based on originally correct (**C**) and originally incorrect (**IC**) predictions



Metrics

- Hierarchical Classification Metrics based on originally correct (**C**) and originally incorrect (**IC**) predictions
  - **C-Corrupt**: The fraction of original correct terminal predictions relabeled to incorrect labels

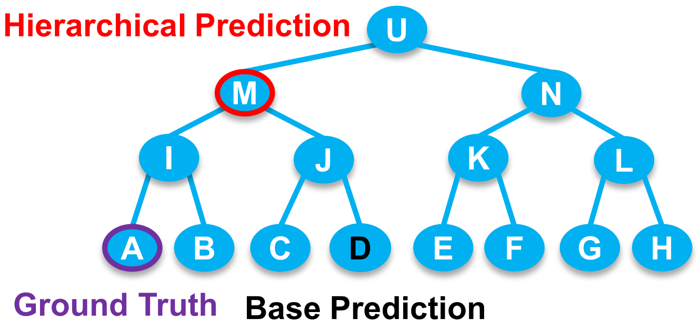
	Ground Truth	Base Prediction	Hierarchical Prediction
C-Corrupt	A	A	N



Metrics

- Hierarchical Classification Metrics based on originally correct (**C**) and originally incorrect (**IC**) predictions
  - C-Corrupt: The fraction of original correct terminal predictions relabeled to incorrect labels
  - **IC-Reform**: The fraction of original incorrect terminal predictions generalized to correct labels

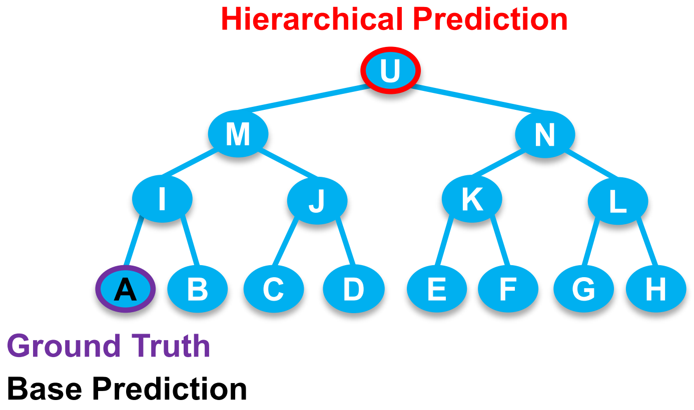
	Ground Truth	Base Prediction	Hierarchical Prediction
C-Corrupt	A	A	N
IC-Reform	**A**	**D**	**M**



Metrics

- Hierarchical Classification Metrics based on originally correct (**C**) and originally incorrect (**IC**) predictions
  - ▼ – C-Corrupt: The fraction of original correct terminal predictions relabeled to incorrect labels
  - ▲ – IC-Reform: The fraction of original incorrect terminal predictions generalized to correct labels
  - ▼ – **C-Withdrawn**: The fraction of original correct terminal predictions relabeled to root

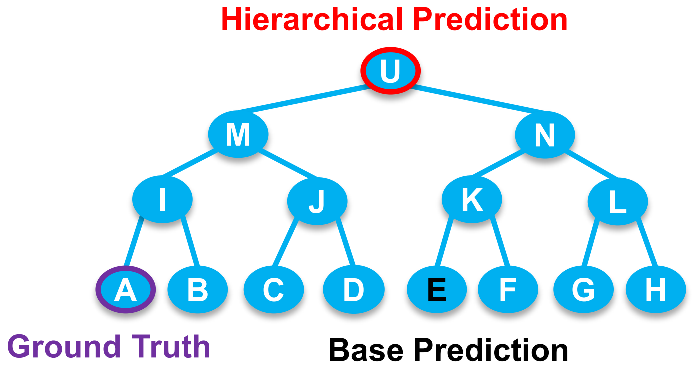
	Ground Truth	Base Prediction	Hierarchical Prediction
C-Corrupt	A	A	N
IC-Reform	A	D	M
C-Withdrawn	**A**	**A**	**U**



Metrics

- Hierarchical Classification Metrics based on originally correct (**C**) and originally incorrect (**IC**) predictions
  - ▼ – C-Corrupt: The fraction of original correct terminal predictions relabeled to incorrect labels
  - ▲ – IC-Reform: The fraction of original incorrect terminal predictions generalized to correct labels
  - ▼ – C-Withdrawn: The fraction of original correct terminal predictions relabeled to root
  - ▼ – **IC-Withdrawn**: The fraction of original incorrect terminal predictions relabeled to root

	Ground Truth	Base Prediction	Hierarchical Prediction
C-Corrupt	A	A	N
IC-Reform	A	D	M
C-Withdrawn	A	A	U
IC-Withdrawn	**A**	**E**	**U**



Metrics

- Hierarchical Classification Metrics:
  - **Accuracy**: the fraction of correct hierarchical predictions (root is considered correct)
  - **Average scaled Information Gain (avg-sIG)**: corresponds to average depth of label generalizations in terms of Information Gain [Deng et al., 2012]

$$sIG(N_i) = \frac{log_2|T| - log_2(|\downarrow(N_i)|)}{log_2|T|}$$

$|T|$ is the total number of terminal classes

$|\downarrow(N_i)|$ Is the number of terminal descendants of class N_i

$$avg\text{-}sIG(N_i) = \frac{1}{M}\sum_{i=1}^{M} sIG(N_i)$$

Experiments: ImageNet-Animal

- ImageNet-Animal derived from WordNet
  - Due to space limit, the lower part of the tree is omitted
  - # in (**#**) indicates the number of terminal classes at the branch

<table>
<tr><td colspan="30">Unknown</td></tr>
<tr><td colspan="27">Vertebrate</td><td colspan="3">Invertebrate</td></tr>
<tr><td colspan="16">Mammal</td><td colspan="4">Reptile</td><td colspan="4">Bird</td><td colspan="2">Fish</td><td>...</td><td colspan="2">Arthropod</td><td>...</td></tr>
<tr><td colspan="15">Placental</td><td>...</td><td colspan="3">Diapsid</td><td>...</td><td colspan="2">Aquatic</td><td>Oscine</td><td>...</td><td>Teleost
fish</td><td>...</td><td></td><td>Insect</td><td>...</td><td></td></tr>
<tr><td colspan="2">Ungulate</td><td colspan="2">Primate</td><td colspan="10">Carnivore</td><td>...</td><td></td><td>Snake</td><td>Lizard</td><td>...</td><td></td><td>Wading
bird</td><td>...</td><td>...</td><td colspan="4"></td><td>...</td><td colspan="2"></td></tr>
<tr><td>Even-toed
ungulate</td><td>...</td><td>Monkey</td><td>...</td><td colspan="8">Canine</td><td>Feline</td><td>...</td><td colspan="2"></td><td>...</td><td>...</td><td colspan="12"></td></tr>
<tr><td></td><td></td><td>...</td><td></td><td colspan="7">Dog</td><td>...</td><td>...</td><td colspan="7"></td><td>...</td><td colspan="9"></td></tr>
<tr><td>...</td><td colspan="3"></td><td colspan="4">Hunting</td><td colspan="2">Working</td><td>...</td><td colspan="19"></td></tr>
<tr><td colspan="4"></td><td>Hound</td><td>Terrier</td><td>Sporting</td><td></td><td>Shepherd</td><td>...</td><td colspan="20"></td></tr>
<tr><td colspan="4"></td><td>...</td><td>...</td><td>...</td><td></td><td>...</td><td colspan="21"></td></tr>
<tr><td>(15)</td><td>(2)</td><td>(13)</td><td>(7)</td><td>(19)</td><td>(26)</td><td>(17)</td><td>(1)</td><td>(12)</td><td>(18)</td><td>(25)</td><td>(12)</td><td>(13)</td><td>(15)</td><td>(17)</td><td>(6)</td><td>(17)</td><td>(11)</td><td>(3)</td><td>(5)</td><td>(16)</td><td>(8)</td><td>(11)</td><td>(24)</td><td>(10)</td><td>(6)</td><td>(8)</td><td>(27)</td><td>(20)</td><td>(14)</td></tr>
</table>

Experiments: ImageNet-Animal

- Highest IC-Reform
- Highest overall accuracy

	Base	Proposed		Deng et al. 2012		Davis et al. 2019	
Confidence	-	90%	95%	90%	95%	90%	95%
C-Corrupt	-	.00	.00	.03	.01	.00	.00
IC-Reform	-	**.96**	**.98**	.48	.72	.67	.70
Accuracy	0.85	**.99**	**1.00**	.89	.95	.95	.95
avg-sIG	0.85	.30	.20	.78	.69	.71	.68
C-Withdrawn	-	.07	.13	.00	.01	.01	.01
IC-Withdrawn	-	.09	.14	.00	.06	.03	.05

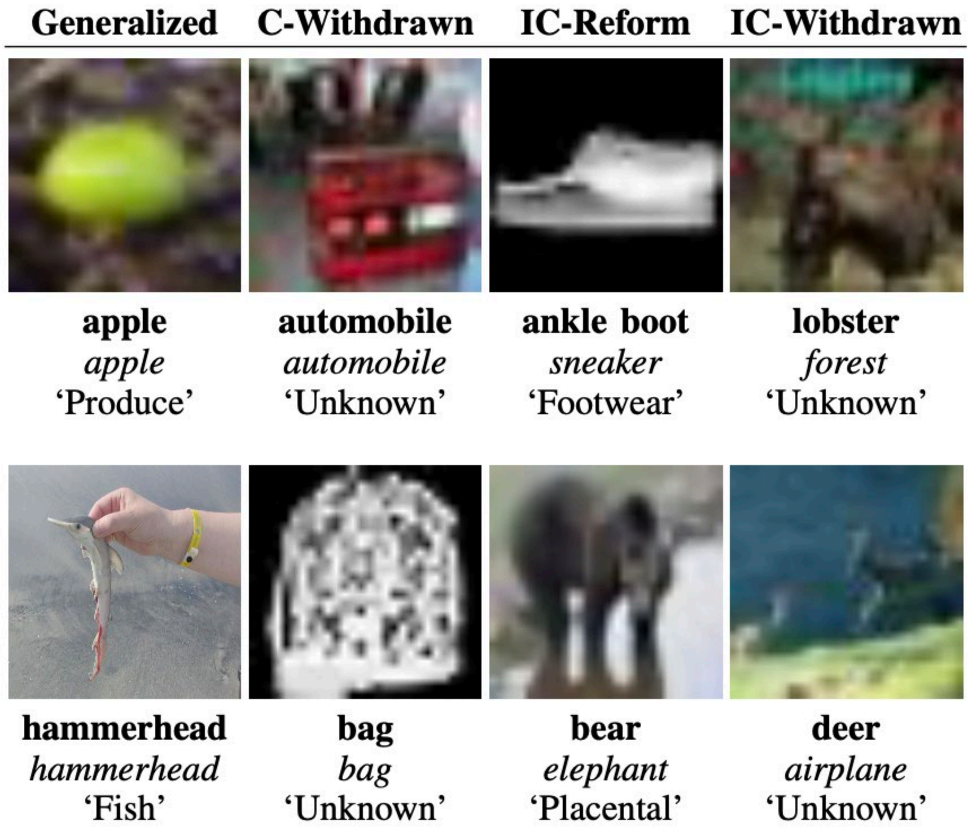
Experiments: ImageNet-Animal

- Highest IC-Reform
- Highest overall accuracy
- Lowest avg-sIG corresponding to high withdrawns

	Base	Proposed		Deng et al. 2012		Davis et al. 2019	
Confidence	-	90%	95%	90%	95%	90%	95%
C-Corrupt	-	.00	.00	.03	.01	.00	.00
IC-Reform	-	.96	.98	.48	.72	.67	.70
Accuracy	0.85	.99	1.00	.89	.95	.95	.95
avg-sIG	0.85	**.30**	**.20**	.78	.69	.71	.68
C-Withdrawn	-	**.07**	**.13**	.00	.01	.01	.01
IC-Withdrawn	-	**.09**	**.14**	.00	.06	.03	.05

Visual Examples Across Datasets

ground truth
flat prediction
'Hierarchical Prediction'



Conclusion

- Estimation of label posteriors using generalized logits
  - Efficient and compact conditional vector
  - Mitigate issues of lack of validation data
- Label generalization based on semantic hierarchy
  - Bottom-up probabilistic inference framework
  - Able to correct mistakes made by the flat base classifier

Thank you!