Hierarchical Classification with Confidence using Generalized Logits

Jim Davis¹, Tong Liang¹, James Enouen¹, Roman Ilin²

¹Ohio State University, Columbus OH 43210
²AFRL/RYAP, Wright-Patterson AFB OH 45433

This work is supported by the U.S. Air Force Research Laboratory
Hierarchical reasoning prevails in many areas
- Hierarchical cognitive process
- Taxonomy in biology
- Natural hierarchical representation of visual features in our brain
• 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 $J$

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

• Start with initial terminal label hypothesis
  – Option 1: Select via argmax of base classifier’s logits
  – Option 2: Select via argmax of estimated terminal label posteriors

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

\[ P(A|\mathcal{L}_A) \] 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:

\[
P(A|\mathcal{L}_A) \\
\]

\[
P(I|\mathcal{L}_A) = P(A \cup B|\mathcal{L}_A) = P(A|\mathcal{L}_A) + P(B|\mathcal{L}_A)
\]
Consider the binary tree shown on the right, evaluate the ancestral path label posteriors:

\[
P(A|\mathcal{L}_A)
\]

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

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

Initial hypothesis: A
Consider the binary tree shown on the right, evaluate the ancestral path label posteriors:

\[
P(A|\mathcal{L}_A)
\]

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

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

\[
P(U|\mathcal{L}_A) = P(M \cup N|\mathcal{L}_A) = P(M|\mathcal{L}_A) + P(E \cup F \cup G \cup H|\mathcal{L}_A)
\]
Consider the binary tree shown on the right, evaluate the ancestral path label posteriors:

\[ P(A|\mathcal{L}_A) \]

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

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

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

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:

\[
P(A|\mathcal{L}_A)
\]

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

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

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

• L1 normalization before inference:

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

Initial hypothesis: A
Extension to Non-Binary Tree

- Adding a node $B^*$ under node $I$

\[
P(A|\mathcal{L}_A)
\]
\[
P(I|\mathcal{L}_A) = P(A|\mathcal{L}_A) + P(B \cup B^*|\mathcal{L}_A) = P(A|\mathcal{L}_A) + P(I\setminus A|\mathcal{L}_A)
\]
\[
P(M|\mathcal{L}_A) = P(I|\mathcal{L}_A) + P(C \cup D|\mathcal{L}_A) = P(I|\mathcal{L}_A) + P(M\setminus I|\mathcal{L}_A)
\]
\[
P(U|\mathcal{L}_A) = P(M|\mathcal{L}_A) + P(E \cup F \cup G \cup H|\mathcal{L}_A) = P(M|\mathcal{L}_A) + P(U\setminus M|\mathcal{L}_A) \triangleq 1.0
\]
Generalized Logit Feature Vector

- Adding a node $B^*$ under node $I$

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

Generalized logit vector
$$\mathcal{L}_A = [l_A, \hat{i}_{I \setminus A}, \hat{i}_{M \setminus I}, \hat{i}_{U \setminus M}]$$
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
• 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

<table>
<thead>
<tr>
<th>Ground Truth</th>
<th>Base Prediction</th>
<th>Hierarchical Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-Corrupt</td>
<td>A</td>
<td>A</td>
</tr>
</tbody>
</table>
• 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

<table>
<thead>
<tr>
<th></th>
<th>Ground Truth</th>
<th>Base Prediction</th>
<th>Hierarchical Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-Corrupt</td>
<td>A</td>
<td>A</td>
<td>N</td>
</tr>
<tr>
<td>IC-Reform</td>
<td>A</td>
<td>D</td>
<td>M</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>Ground Truth</th>
<th>Base Prediction</th>
<th>Hierarchical Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-Corrupt</td>
<td>A</td>
<td>A</td>
<td>N</td>
</tr>
<tr>
<td>IC-Reform</td>
<td>A</td>
<td>D</td>
<td>M</td>
</tr>
<tr>
<td>C-Withdrawn</td>
<td>A</td>
<td>A</td>
<td>U</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Ground Truth</th>
<th>Base Prediction</th>
<th>Hierarchical Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-Corrupt</td>
<td>A</td>
<td>N</td>
</tr>
<tr>
<td>IC-Reform</td>
<td>A</td>
<td>M</td>
</tr>
<tr>
<td>C-Withdrawn</td>
<td>A</td>
<td>U</td>
</tr>
<tr>
<td>IC-Withdrawn</td>
<td>A</td>
<td>U</td>
</tr>
</tbody>
</table>

Diagram:

```
            U
           / \
          /   \ 
         /     \
        /       \
       /         \
      /           \
     /             \
    /               \
   /                 \n  /                   \n /                     \n/                       U
```

Ground Truth: A B C D E F G H
Base Prediction: A B C D E F G H
Hierarchical Prediction: A B C D E F G H
• 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 $N_i$

\[
avg-sIG(N_i) = \frac{1}{M} \sum_{i=1}^{M} sIG(N_i)
\]
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>
<thead>
<tr>
<th>Ungulate</th>
<th>Primate</th>
<th>Carnivore</th>
<th>Mammal</th>
<th>Vertebrate</th>
<th>Reptile</th>
<th>Bird</th>
<th>Fish</th>
<th>Invertebrate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Even-toed ungulate</td>
<td>Monkey</td>
<td></td>
<td>Canine</td>
<td>Placental</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Feline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Dog</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experiments: ImageNet-Animal

- Highest IC-Reform
- Highest overall accuracy

<table>
<thead>
<tr>
<th></th>
<th>Base</th>
<th>Proposed</th>
<th>Deng et al. 2012</th>
<th>Davis et al. 2019</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confidence</td>
<td>-</td>
<td>90%</td>
<td>90%</td>
<td>90%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>95%</td>
<td>95%</td>
<td>95%</td>
</tr>
<tr>
<td>C-Corrupt</td>
<td>-</td>
<td>.00</td>
<td>.03</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.00</td>
<td>.01</td>
<td>.00</td>
</tr>
<tr>
<td>IC-Reform</td>
<td>-</td>
<td>.96</td>
<td>.48</td>
<td>.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.98</td>
<td>.72</td>
<td>.70</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.85</td>
<td>.99</td>
<td>.89</td>
<td>.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.00</td>
<td>.95</td>
<td>.95</td>
</tr>
<tr>
<td>avg-sIG</td>
<td>0.85</td>
<td>.30</td>
<td>.78</td>
<td>.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.20</td>
<td>.69</td>
<td>.68</td>
</tr>
<tr>
<td>C-Withdrawn</td>
<td>-</td>
<td>.07</td>
<td>.00</td>
<td>.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.13</td>
<td>.01</td>
<td>.01</td>
</tr>
<tr>
<td>IC-Withdrawn</td>
<td>-</td>
<td>.09</td>
<td>.00</td>
<td>.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.14</td>
<td>.06</td>
<td>.05</td>
</tr>
</tbody>
</table>
Experiments: ImageNet-Animal

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

<table>
<thead>
<tr>
<th></th>
<th>Base</th>
<th>Proposed</th>
<th>Deng et al. 2012</th>
<th>Davis et al. 2019</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confidence</td>
<td>-</td>
<td>90%</td>
<td>90%</td>
<td>90%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>95%</td>
<td>95%</td>
<td>95%</td>
</tr>
<tr>
<td>C-Corrupt</td>
<td>-</td>
<td>.00</td>
<td>.03</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.00</td>
<td>.01</td>
<td>.00</td>
</tr>
<tr>
<td>IC-Reform</td>
<td>-</td>
<td>.96</td>
<td>.48</td>
<td>.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.98</td>
<td>.72</td>
<td>.70</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.85</td>
<td>.99</td>
<td>.89</td>
<td>.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.00</td>
<td>.95</td>
<td>.95</td>
</tr>
<tr>
<td>avg-sIG</td>
<td>0.85</td>
<td>.30</td>
<td>.78</td>
<td>.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.20</td>
<td>.69</td>
<td>.68</td>
</tr>
<tr>
<td>C-Withdrawn</td>
<td>-</td>
<td>.07</td>
<td>.00</td>
<td>.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.13</td>
<td>.01</td>
<td>.01</td>
</tr>
<tr>
<td>IC-Withdrawn</td>
<td>-</td>
<td>.09</td>
<td>.00</td>
<td>.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.14</td>
<td>.06</td>
<td>.05</td>
</tr>
</tbody>
</table>
### Visual Examples Across Datasets

<table>
<thead>
<tr>
<th>ground truth</th>
<th>apple</th>
<th>automobile</th>
<th>ankle boot</th>
<th>lobster</th>
</tr>
</thead>
<tbody>
<tr>
<td>flat prediction</td>
<td>apple</td>
<td>automobile</td>
<td>ankle boot</td>
<td>lobster</td>
</tr>
<tr>
<td>‘Hierarchical Prediction’</td>
<td>‘Produce’</td>
<td>‘Unknown’</td>
<td>‘Footwear’</td>
<td>‘Unknown’</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>hammeredhead</th>
<th>bag</th>
<th>bear</th>
<th>deer</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘Fish’</td>
<td>‘Unknown’</td>
<td>‘Placental’</td>
<td>‘Unknown’</td>
</tr>
</tbody>
</table>

**Generalized**

- **C-Withdrawn**

- **IC-Reform**

- **IC-Withdrawn**
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!