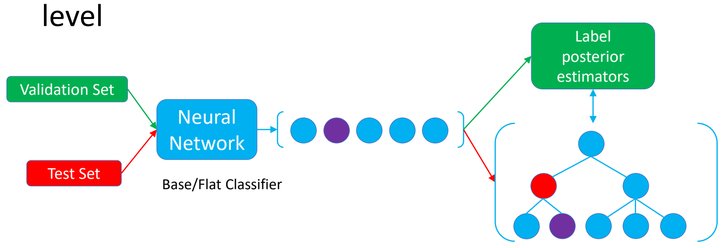
Hierarchical Classification with Confidence using Generalized Logits

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

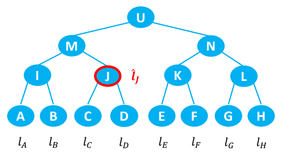
1. Why Hierarchical Classification?

- Utilize extra knowledge of the semantic relationships among labels of flat classification problem
- Offer statistical confidence guarantee of hierarchical prediction instead of forcing the prediction at the terminal level
  - More informative label when confident
  - Less informative label when not confident at deeper level



2. Generalized Logits

- Goal: Aggregate logit information to derive logit of non-terminal classes
- Derive 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

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

3. Inference: An Example

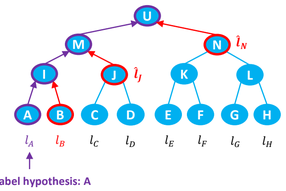
- Start with initial terminal label hypothesis of class A
- Evaluate the posteriors of classes on the Ancestral path

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

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

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

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



- 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

4. Extension to Non-binary Tree

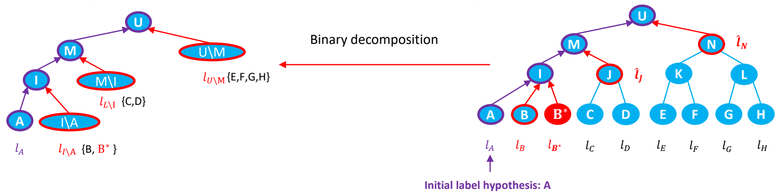
- Add a node B* under class I

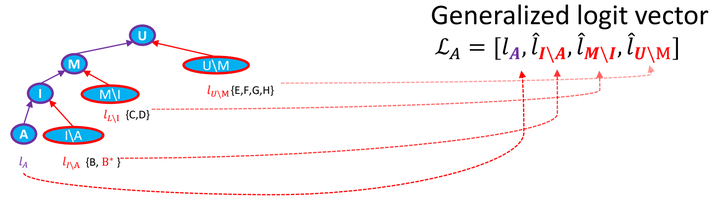
$$\mathrm{P}(\mathrm{A}|\mathcal{L}_A)$$

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

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

$$\mathrm{P}(\boldsymbol{U}|\boldsymbol{\mathcal{L}_A}) = \mathrm{P}(\mathrm{M}|\mathcal{L}_A) + P(E \cup F \cup G \cup H|\mathcal{L}_A) = \mathrm{P}(\mathrm{M}|\mathcal{L}_A) + P(U \backslash \mathrm{M}|\mathcal{L}_A) \triangleq 1.0$$





Generalized logit vector

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

5. Results

- Comparing with related bottom-up hierarchical classification methods on ImageNet-Animal dataset (more results in the paper)
  - Highest IC-Reform -> corrects most mistakes made by base classifier
  - Highest overall accuracy
  - Lowest avg-sIG -> predictions closer to root comparing to other methods

	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

6. Conclusion

- Estimation of label posteriors using compact vector of generalized logits
- Bottom-up probabilistic inference framework
- Label generalization based on semantic hierarchy that can correct most of the mistakes made by the flat classifiers