Loss in Layers

Evaluating hierarchical loss functions for image classification (gr. 28)

Pierrick Bournez, Paul-Arno Lamarque, Kerrian Le Caillec, Samuel Sithakoul, Paul Tabbara

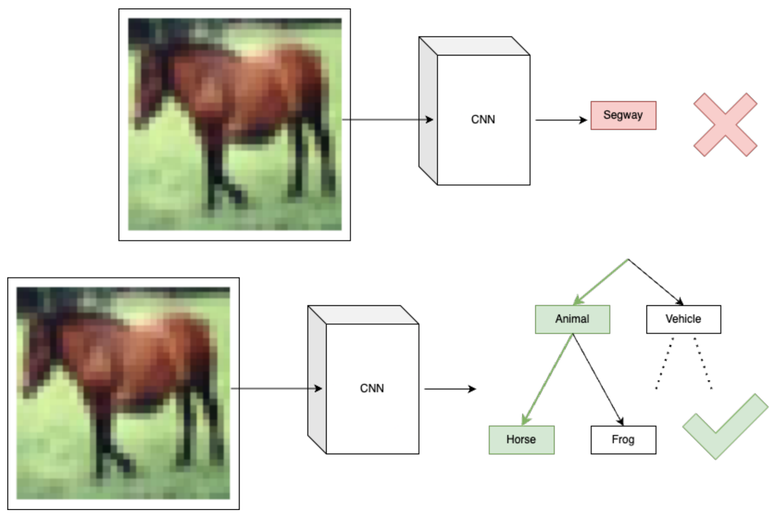
CentraleSupélec, Université Paris-Saclay, France – Supervisor: Alix Chazottes

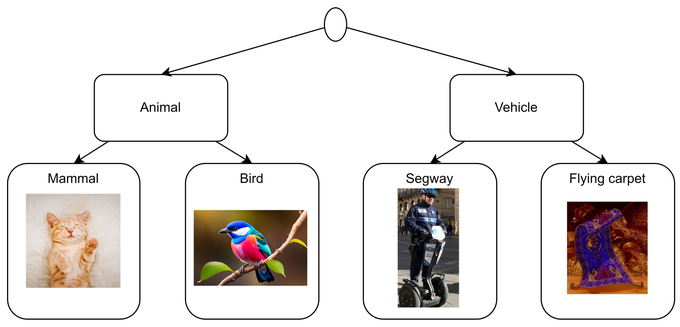
CentraleSupélec

SCAN ME

Introduction

- In Deep Learning, automatic image classification can suffer from unstructured labels and loss functions.
- Introducing **semantic knowledge** in data labels, via a **custom tree-based loss**, could reduce critical misclassifications.
- Our goal is to assess if this new **model-agnostic** method performs better than existing solutions on the **CIFAR-10** dataset.





Example of a hierarchical tree

A closer look at the losses

Classical loss functions

CrossEntropy Loss (CE)

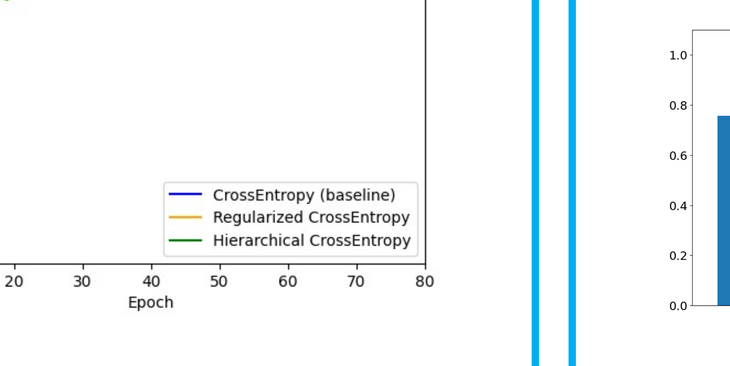
Most commonly used loss function for classification problems.
The model outputs, p_i, must match the labels y_i.
To minimize this loss, p_i must be close to 1 where $y_i = 1$.

$$\ell_{CE}(\mathbf{y}, \mathbf{p}) = -\sum_{i=1}^{N} y_i \log(p_i)$$

Regularized CrossEntropy (r-CE)

Overly confident predictions are penalized to avoid overfitting.

$$\ell_{rCE}(\mathbf{y}, \mathbf{p}) = \ell_{CE}(\mathbf{y}, \mathbf{p}) + \lambda \sum_{i=1}^{N} p_i \log(p_i)$$

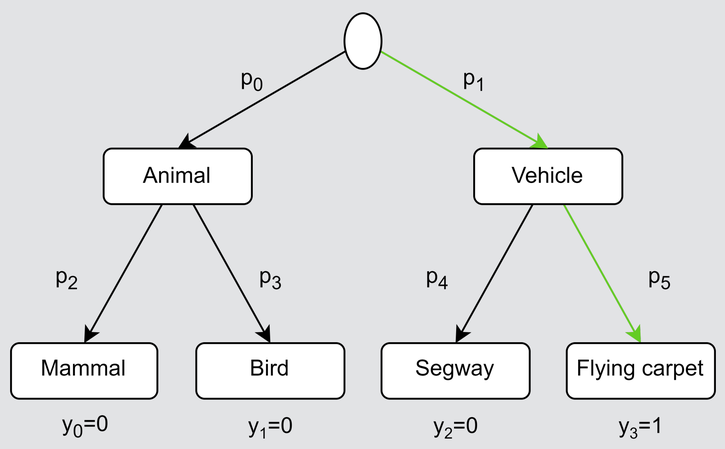


Hierarchical loss function

Hierarchical CrossEntropy (h-CE)

Leverages knowledge about the classes.
Intermediate labels are predicted.
Let $\mathbb{J}_i = \{\text{nodes on path from root to } i\}$.

$$\ell_H(\mathbf{y}, \mathbf{p}) = -\sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \in \mathbb{J}_i}}^{N} y_i \log(p_j) + \sum_{\substack{i=1 \\ i \text{ not a leaf}}}^{N} \prod_{\substack{j=1 \\ j \in \mathbb{J}_i}}^{N} p_j$$

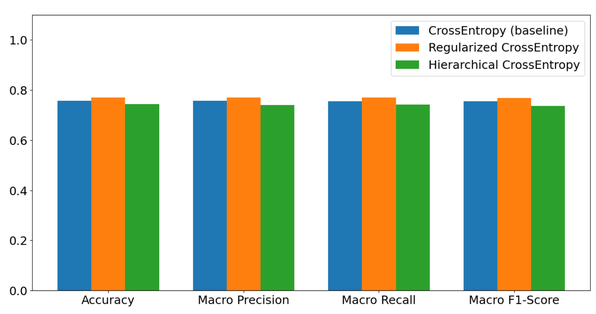


Results during training

Training with the three loss functions yields similar accuracy profiles

Results after training

- Models trained until convergence
- Metrics evaluated on 10 test folds
- **h-CE < CE ≈ r-CE**
- Mann-Whitney U-tests do **not show any significant differences**



Conclusions

No significant improvements over the baseline were observed.

Hypothesis: the **dataset** is too simple to leverage the hierarchical information.

Possible axes of further research:

- Explore more **challenging** datasets (CIFAR-100, Cityscapes)
- Use more **complex models** (ResNets, transformers)