# Images in a Hierarchy: Monarchy or Anarchy?

Moritz Neeb

March 17, 2016

# Goals of this talk

- Learn about an interesting concept:
  Hierarchical Classification
- Have some fun: Algorithms with Friends.

# Outline

Definition Hierarchical Image Classification

How to evaluate a model?

How to find a model?

# Classification: a model predicting labels

Find a *model f*
for *labeled data* $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
s.t. a certain *loss* is minimized:

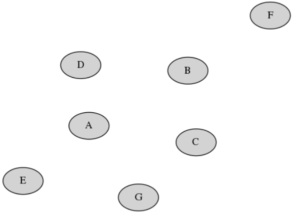
$$\min \sum_i L(y_i, \hat{y}_i)$$

where

$$\hat{y}_i := f(x_i)$$

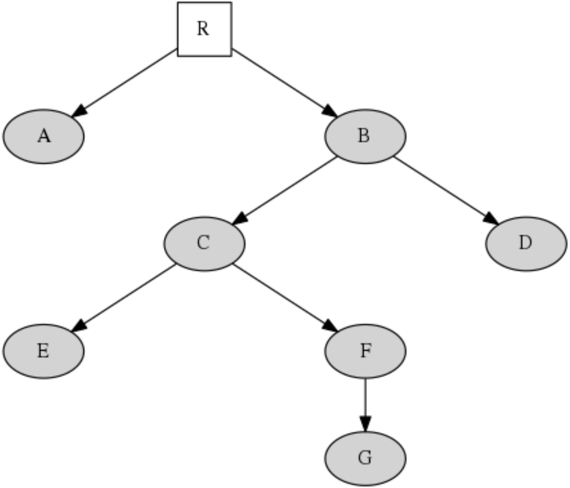
and

$$\hat{y}_i, y_i \in \{0, 1, \ldots, k\} := Y \qquad \forall i$$

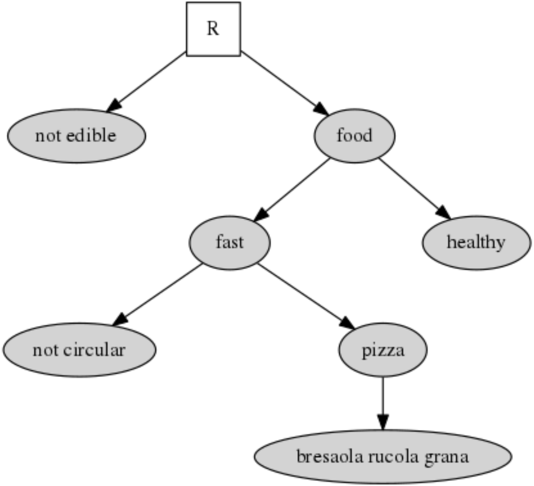


Y

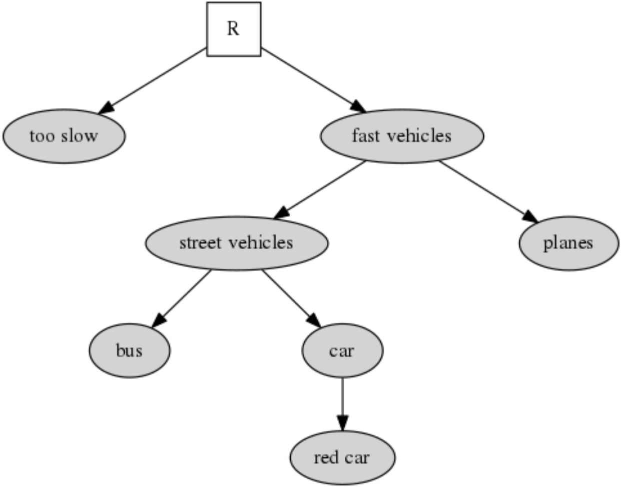
# A hierarchy . . .



# A hierarchy . . .


R
not edible
food
fast
healthy
not circular
pizza
bresaola rucola grana

# A hierarchy . . .


R
too slow
fast vehicles
street vehicles
planes
bus
car
red car

# . . . is an anti-reflexive partially ordered set

Let the labels $Y$ be a finite set with the following properties:

- There is only one greatest element, the root $R$
- $\forall y_i, y_j \in Y : y_i \prec y_j \Rightarrow y_j \nprec y_i$ (asymmetry)
- $\forall y_i \in Y : y_i \nprec y_i$ (anti-reflexivity)
- $\forall y_i, y_j, y_k \in Y : y_i \prec y_j \wedge y_j \prec y_k \Rightarrow y_i \prec y_k$ (transitivity)

# Classification: a model predicting labels

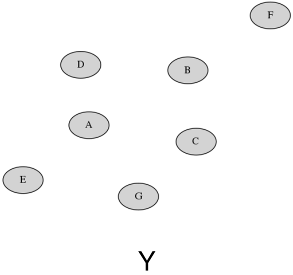
$$\min \sum_i L(y_i, \hat{y}_i)$$

where

$$\hat{y}_i := f(x_i)$$

and

$$\hat{y}_i, y_i \in \{0, 1, \ldots, k\} := Y \qquad \forall i$$



# Hierarchical Classification:
# a model predicting hierarchical labels

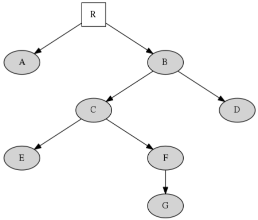
$$\min \sum_i L(y_i, \hat{y}_i)$$

where

$$\hat{y}_i := f(x_i)$$

and

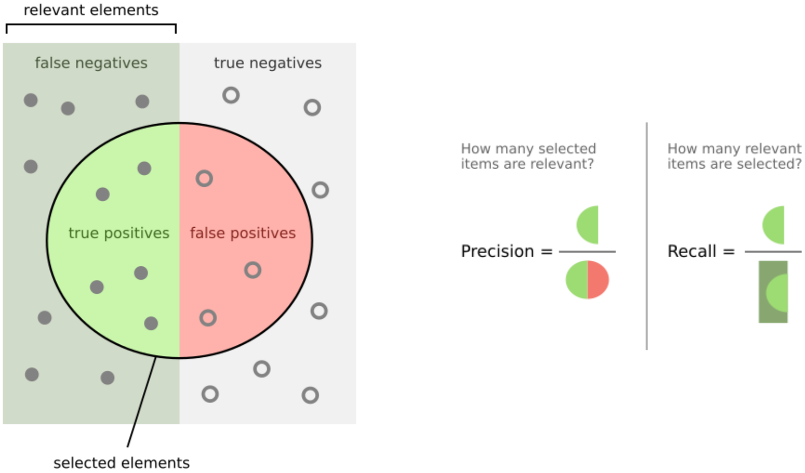
$$\hat{y}_i, y_i \in \{0, 1, \ldots, k\} := Y \qquad \forall i$$



Y

**where $Y$ has hierarchy properties**

# Standard Evaluation Metrics: Precision & Recall

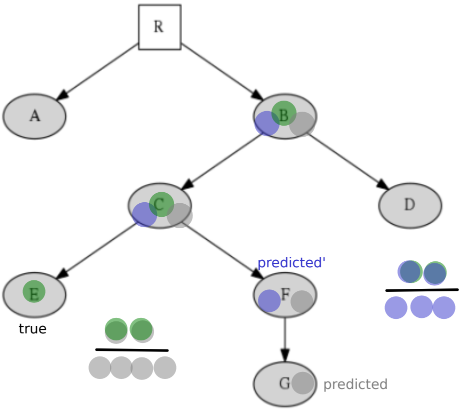


# What can go wrong?

- The relation is not considered
- This can lead to terrible mistakes:
    - classifying a red car as too slow or
    - pizza bressaola rucola grana as not edible.

# Hierarchical evaluation metrics: Precision

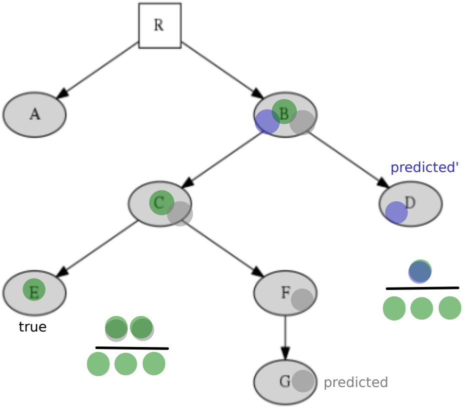
Punish incorrect steps



$$\frac{|\text{path}_T \cap \text{path}_P|}{|\text{path}_P|}$$

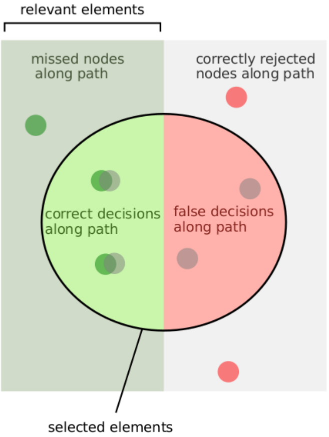
# Hierarchical evaluation metrics: Recall

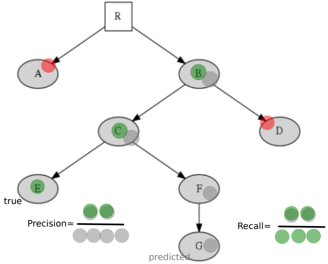
Punish missed correct steps

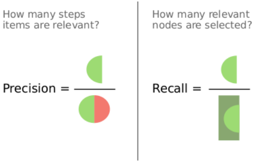


$$\frac{|\text{path}_T \cap \text{path}_P|}{|\text{path}_T|}$$

# Connection between flat & hierarchical metrics


relevant elements
missed nodes along path
correctly rejected nodes along path
correct decisions along path
false decisions along path
selected elements




How many steps items are relevant?
Precision =
How many relevant nodes are selected?
Recall =

# Hierarchical evaluation: (to) sum it up

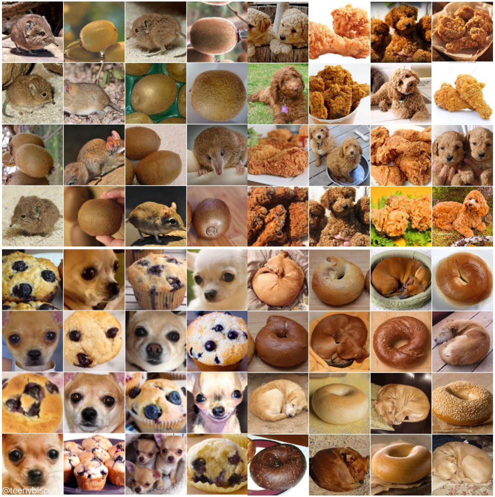
- Precision: Punish incorrect steps

$$hP := \frac{\sum_i |P_i \cap T_i|}{\sum_i |P_i|} \qquad fP = \frac{TP}{TP + FP}$$

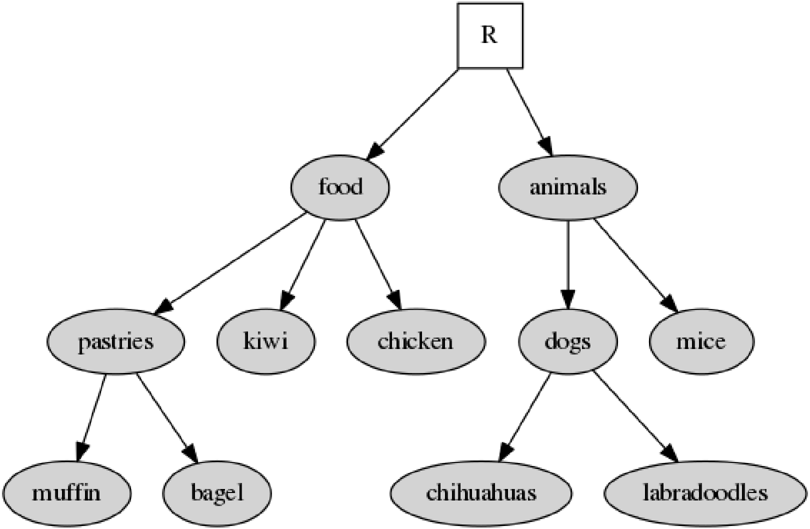
- Recall: Punish missed correct steps

$$hR := \frac{\sum_i |P_i \cap T_i|}{\sum_i |T_i|} \qquad fR = \frac{TP}{TP + FN}$$

# Sometimes it’s just difficult.

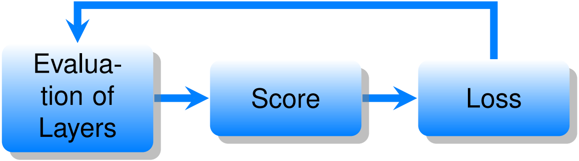


# Towards a solution



The NN can be divided into two steps:

- Feature Generation ($\hat{=}$ Forward Propagation)
- Loss function ($\Rightarrow$ Backpropagation/Optimization)



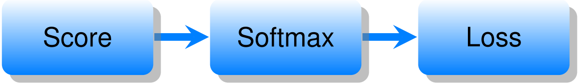
Evaluation of Layers

Score

Loss

Let’s adapt the latter.

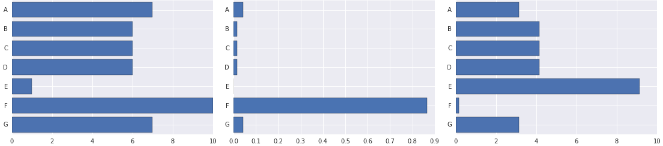
# Crossentropy-Loss is based on Softmax



Score

Softmax

Loss



$$f = \begin{pmatrix} \mathrm{A}_{\mathrm{score}} \\ \mathrm{B}_{\mathrm{score}} \\ \dots \\ \mathrm{G}_{\mathrm{score}} \end{pmatrix} \qquad P(y_i|f) = \frac{\exp(f_{y_i})}{\sum_k \exp(f_k)} \qquad -\log(p(y_i^T))$$

# Augmented Softmax

Standard Softmax:

$$P(y_i|f) = \frac{exp(f_{y_i})}{\sum_{k=1}^{K} exp(f_k)}$$

let's introduce *S*, a semantic relatedness matrix:

$$P(y_i|f, S) = \frac{\sum_{r=1}^{K} S_{y_i,r} exp(f_r)}{\sum_{r=1}^{K} \sum_{k=1}^{K} S_{k,r} exp(f_k)}$$

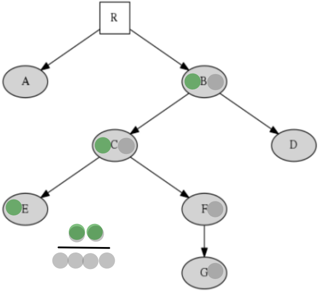
# Hierarchical loss needs a distance measure

Define the distance as follows:

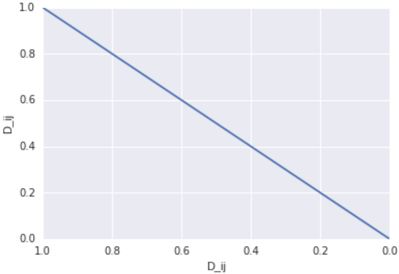
$$D_{ij} := \frac{|\mathrm{path}_i \cap \mathrm{path}_j|}{\max(|\mathrm{path}_i|, |\mathrm{path}_j|)} \hat{=} \min(hP, hR)$$

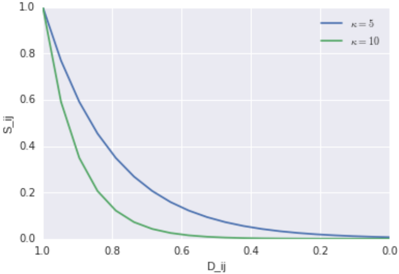
we get for example:

$$D_{EG} = \frac{|\mathrm{path\ E} \cap \mathrm{path\ G}|}{\max(|\mathrm{path\ E}|, |\mathrm{path\ G}|)}$$
$$= \frac{|\{B, C\}|}{|\{B, C, F, G\}|} = \frac{1}{2}$$



# Semantic relatedness: Faster decaying

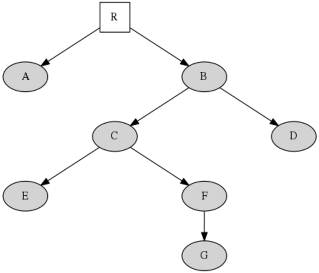


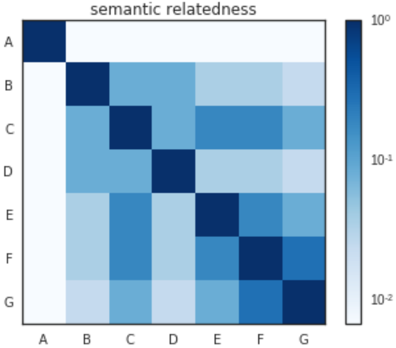


$$D_{ij} := \frac{|\text{path}_i \cap \text{path}_j|}{\max(|\text{path}_i|, |\text{path}_j|)}$$

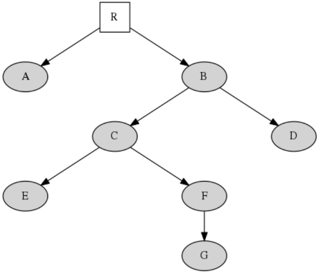
$$S := \exp\left(-\kappa\left(1 - D\right)\right)$$

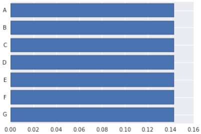
# Semantic relatedness: Illustration



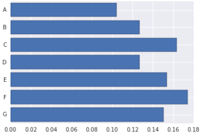


# Applying augmented Softmax: initial case





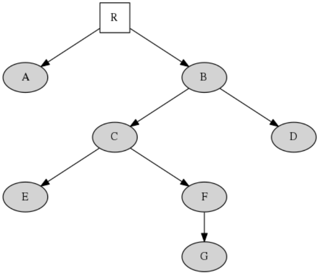
Standard Softmax

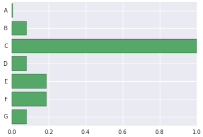


Augmented Softmax

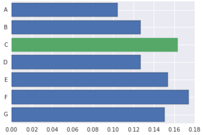
$$P(y_i|f, S) = \frac{\sum_{r=1}^{K} S_{y_i,r} exp(f_r)}{\sum_{r=1}^{K} \sum_{k=1}^{K} S_{k,r} exp(f_k)}$$

# Applying augmented Softmax: initial case





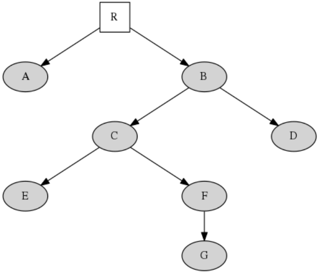
$S_C$



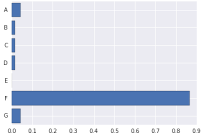
Augmented Softmax

$$P(y_i|f,S) = \frac{\sum_{r=1}^{K} S_{y_i,r} exp(f_r)}{\sum_{r=1}^{K} \sum_{k=1}^{K} S_{k,r} exp(f_k)}$$

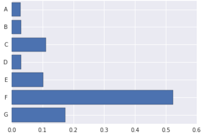
# Higher certainty spreads along tree



$$P(y_i|f,S) = \frac{\sum_{r=1}^{K} S_{y_i,r} exp(f_r)}{\sum_{r=1}^{K} \sum_{k=1}^{K} S_{k,r} exp(f_k)}$$

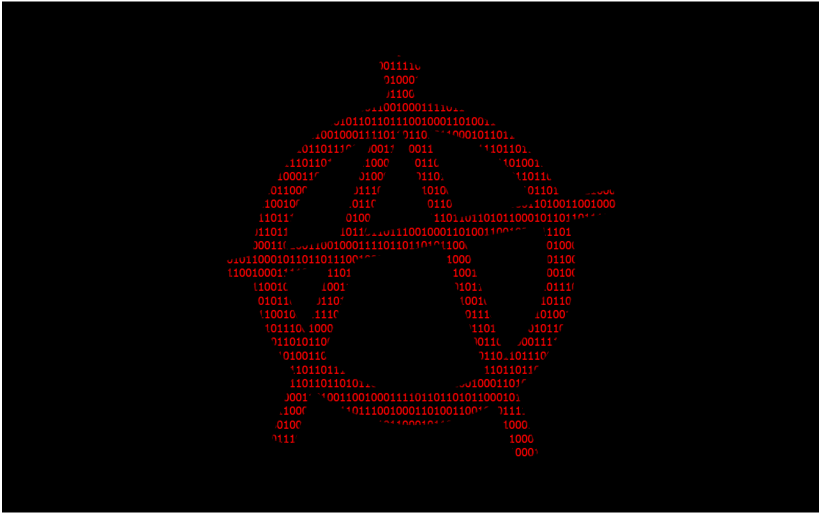


Standard Softmax



Augmented Softmax

# Anarchy is Order?



# This was only a glimpse

- Questions answered:
  - How predictors/models can be compared
  - How Neural Networks can be guided

# This was only a glimpse

- Questions answered:
  - How predictors/models can be compared
  - How Neural Networks can be guided
- Open Questions
  - Does this work? :)
  - Especially: Can human-crafted hierarchies guide the optimization?
  - How to train: What data to use as input?
  - Can features be re-used along the hierarchy?
  - How to deal with semantically overlapping concepts?

# References

Articles:

- Silla, Freitas: A survey of hierarchical classification across different application domains, Data Mining and Knowledge Discovery `http://dl.acm.org/citation.cfm?id=1937884`
- Zhao, Fei-Fei, Xing: Large-Scale Category Structure Aware Image Categorization, NIPS 2011

  `http://vision.stanford.edu/pdf/NIPS2011_0730.pdf`
- we didn't have time for the architecture-adaption:
  Yan et al.: HD-CNN: Hierarchical Deep Convolutional Neural Network for Large Scale Visual Recognition

  `http://arxiv.org/abs/1410.0736`

# Other image sources

- Precision Recall Diagram `https://en.wikipedia.org/wiki/Precision_and_recall#/media/File:Precisionrecall.svg`
- Dogs vs. Food `https://twitter.com/teenybiscuit/`
- Anarchy is Order. "I won't follow these citing rules! ;)"

# Thanks for listening.

Q? A!