Topical Classification of Images in Wikipedia

Matheus V. Bernat

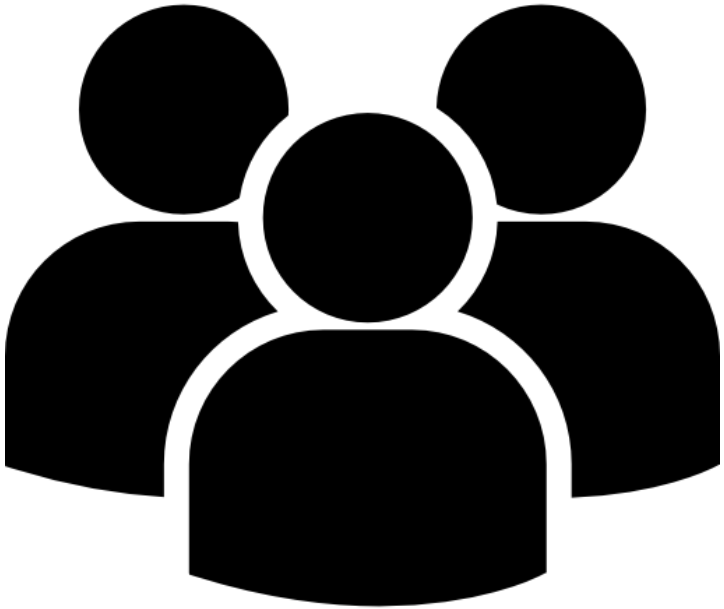
Pavlo Melnyk & Per-Erik Forssén (LiU), Miriam Redi (Wikimedia Foundation), Tiziano Piccardi & Bob West (DLAB)

Image classification, ok,
but what is a ***topic***?

Topic ≙ general concept that can englobe several visual objects

Thanks, now I know what a ***topic*** is. Let’s classify them!

Motivation



814*M* unique users,
with **10*B*** monthly **page views**

21,000 : 1

38*k* monthly
active **editors**

https://stats.wikimedia.org/#/en.wikipedia.org

Motivation

Need to automate!

Improve article quality
Welcome new volunteers
Check grammar
Add images
...

21,000 : 1

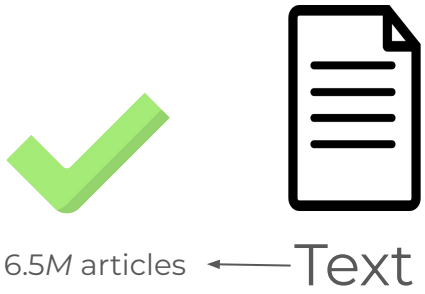
814*M* unique users,
with **10*B*** monthly **page views**

38*k* monthly
active **editors**

https://stats.wikimedia.org/#/en.wikipedia.org

But automate ***what***?

Motivation



Automation system: ORES

- Topic classification
- Quality score
- Vandalism detection

1 https://en.wikipedia.org/wiki/Wikipedia:Size_comparisons#:~:text=Currently%2C%20the%20English%20Wikipedia%20alone,million%20articles%20in%20309%20languages.
2 WIT dataset. https://dl.acm.org/doi/pdf/10.1145/3404835.3463257

Text

Vandalism detection on en.wiki

✍️ 100*k* daily edits 🕵️‍♀️🏃‍♀️ 10 revisions per minute

⏱️⏱️⏱️⏱️⏱️ 166 labour hours per day

=

🧍🧍🧍🧍🧍🧍🧍🧍🧍

83 volunteers working 2 hours a day

🤖 **ORES** doing 90% of revisions

⏱️ 16.6 labour hours per day

=

🧍 8 volunteers working 2 hours a day

Motivation

6.5*M* articles ← Text

Automation system: ORES

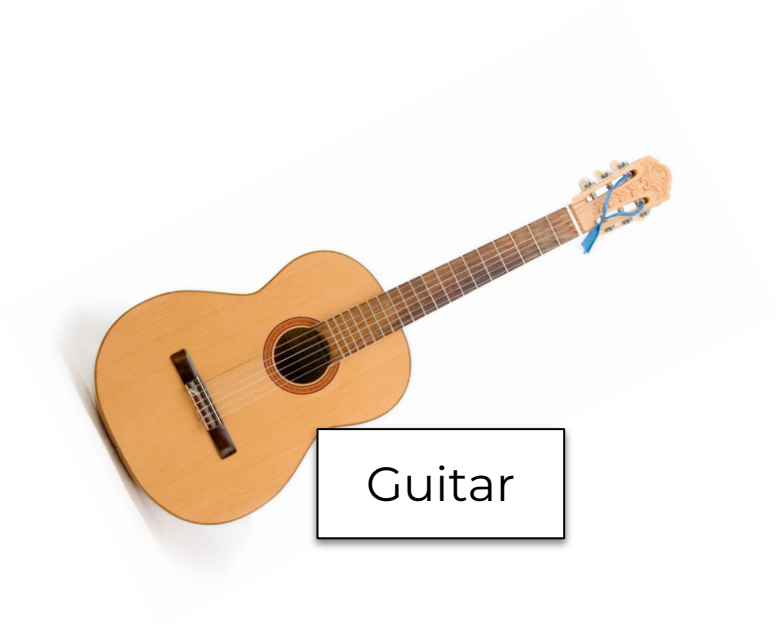
- Topic classification
- Quality score
- Vandalism detection

3.9*M* images ← Images

1 https://en.wikipedia.org/wiki/Wikipedia:Size_comparisons#:~:text=Currently%2C%20the%20English%20Wikipedia%20alone,million%20articles%20in%20309%20languages.
2 WIT dataset. https://dl.acm.org/doi/pdf/10.1145/3404835.3463257

Problem: Images are lacking labels

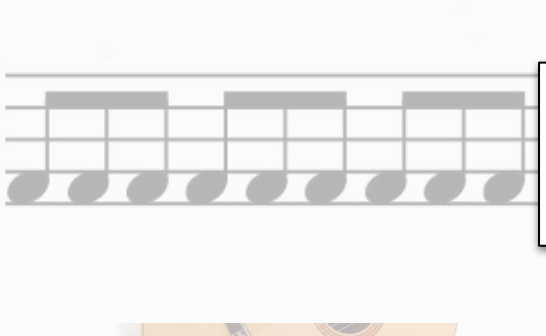
Why don’t we use a standard object classifier?



Guitar

Microphone

Why don’t we use a standard object classifier?



Music!

Guit

Problem: Images are lacking ***topical*** labels

Goal: Build a custom ***topical*** classifier!

Applications

Research

- Find visual knowledge gaps
- Explain reader interaction

Practical

- Detect visual vandalism
- Suggest images

Improve quality of Wikipedia content

RECAP:

Wikipedia is huge, needs task automation

Text 📜 tasks are automated with ORES ✔️, while **image** 🖼️ tasks are missing automation ❌

To automate image 🖼️ tasks we need to classify images with ***topical*** labels

Ok, what’s the matter?

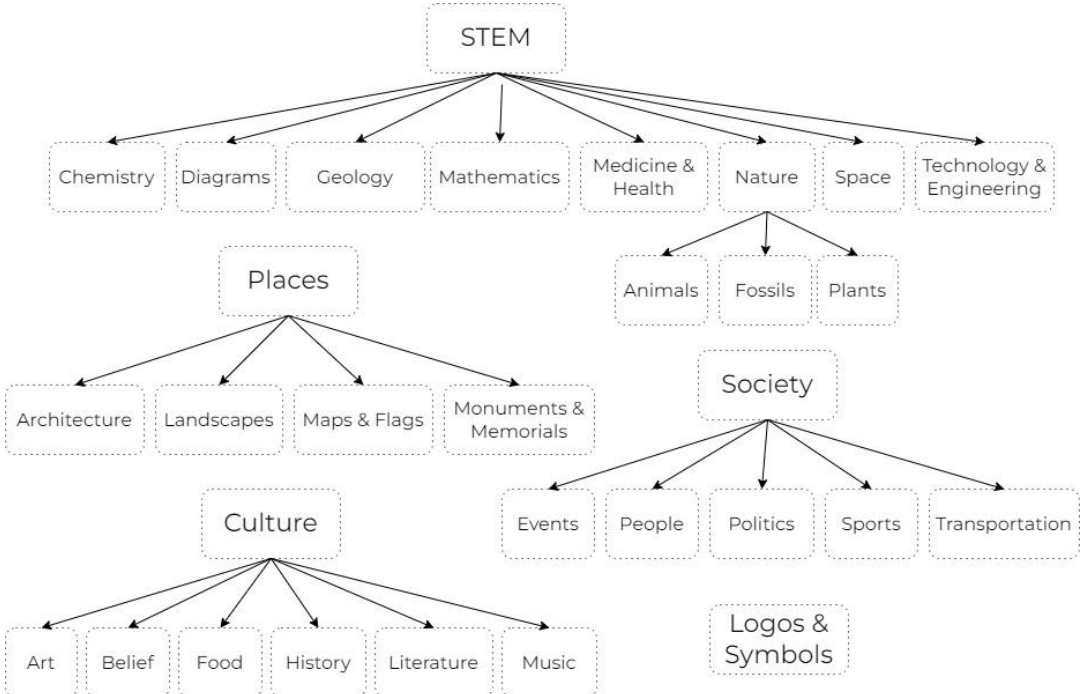
Challenges of constructing this topical classifier – 1

1. Ground truth labels are ***predictions*** from parallel work (not human ground truth)

Topical classification

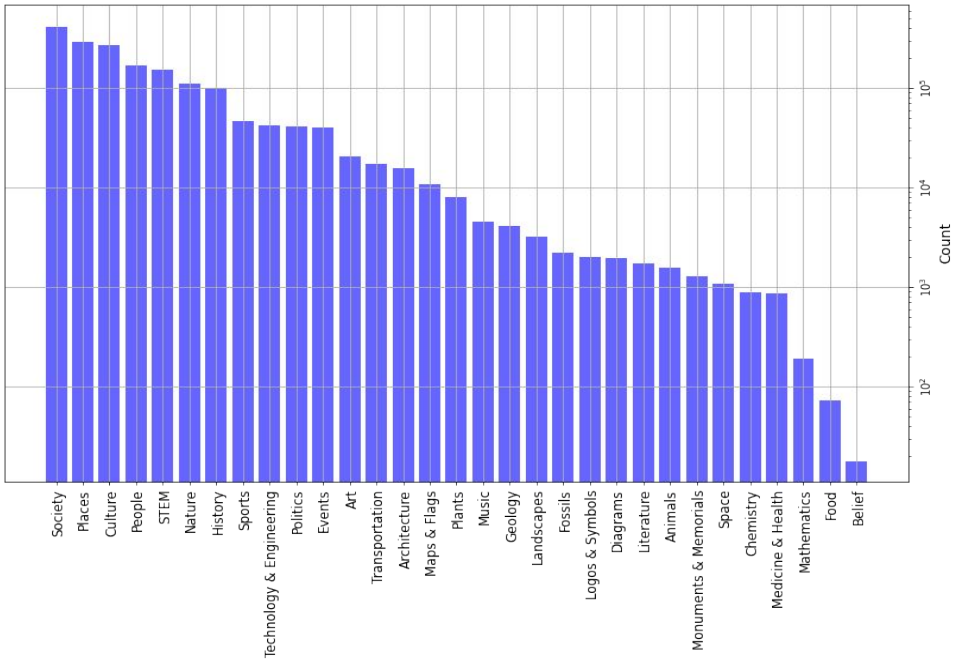
Taxonomy

Classification



Challenges of constructing this topical classifier – 2

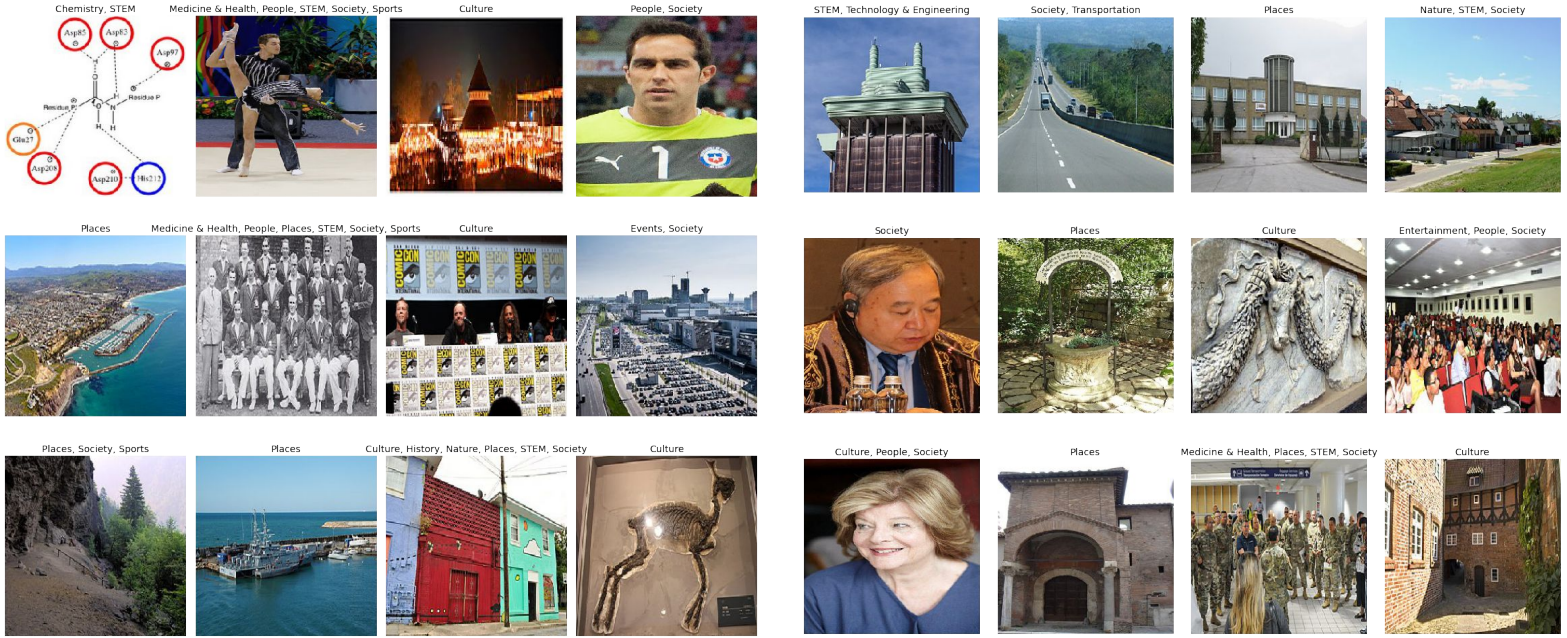
2. Heavily imbalanced label distribution



log-scale!

Challenges of constructing this topical classifier – 3

3. Peculiar image distribution



Challenges of constructing this topical classifier – 4

4. Classify in ***topics*** rather than *objects*

What's the **scope** of the thesis 🤔?

RQ1: How to mitigate the consequences of the imbalanced data?

RQ2: How does a *hierarchical* classification perform compared to a *flat* one?

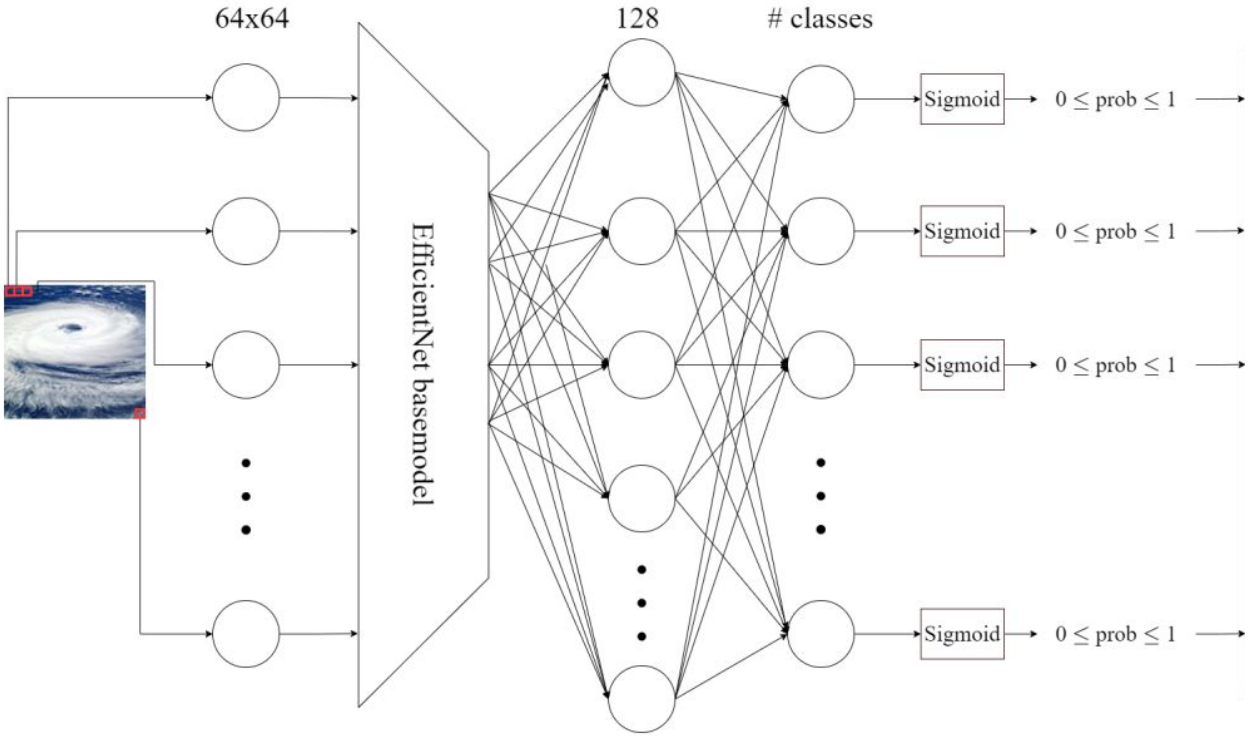
RQ3: Using the developed model, what insights can be drawn from images of Wikipedia?

Methods

- ➔ Big data ➔ use **deep learning**
- ➔ Many parameters ➔ use **transfer learning**
- ➔ An may have multiple labels ➔ output **independent probabilities**, use binary cross-entropy loss function
- ➔ Hierarchical data ➔ experiment with **hierarchical model**
- ➔ Imbalanced distribution makes model underperform on rarer labels ➔ apply **mitigation techniques**

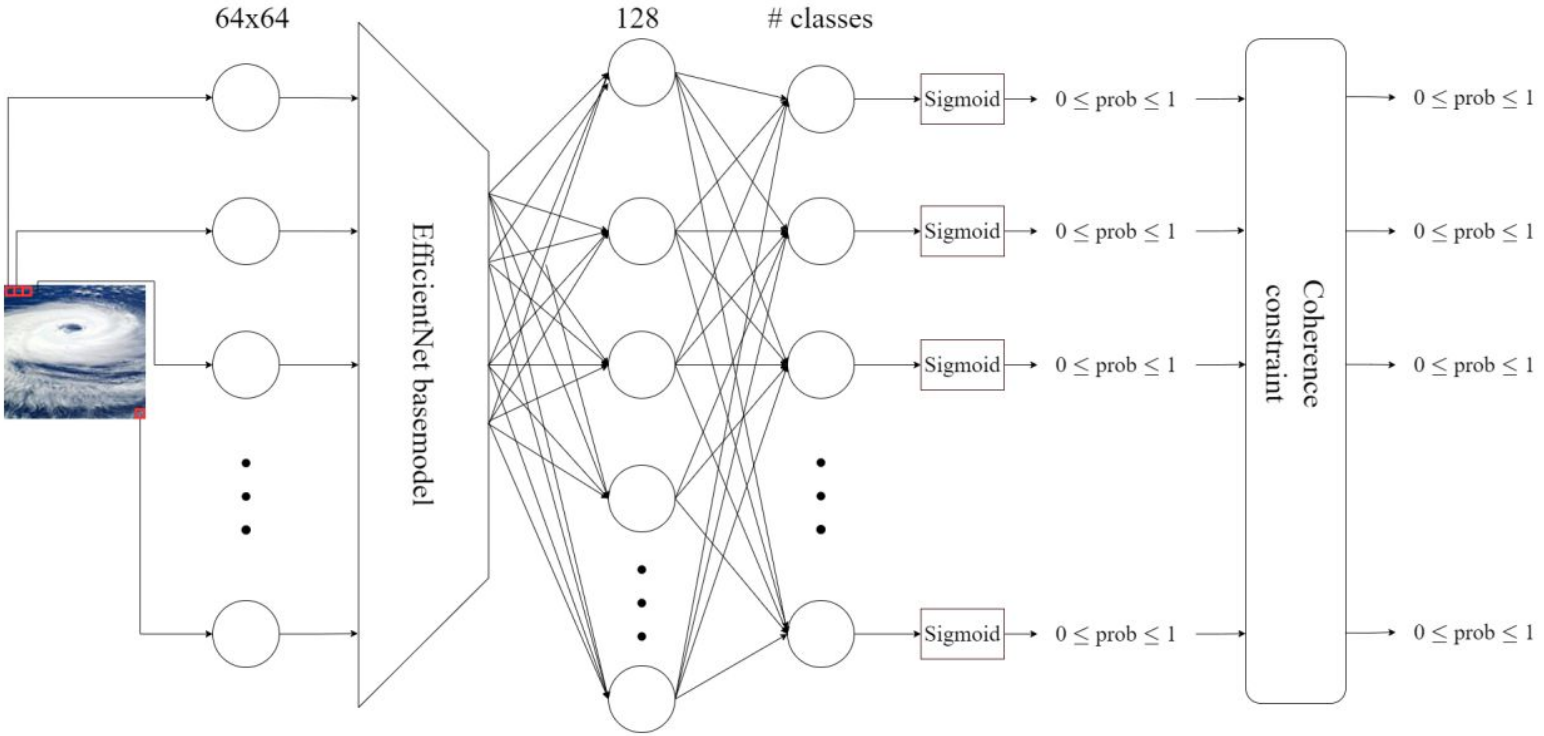
RQ1

Baseline model (flat)



RQ1

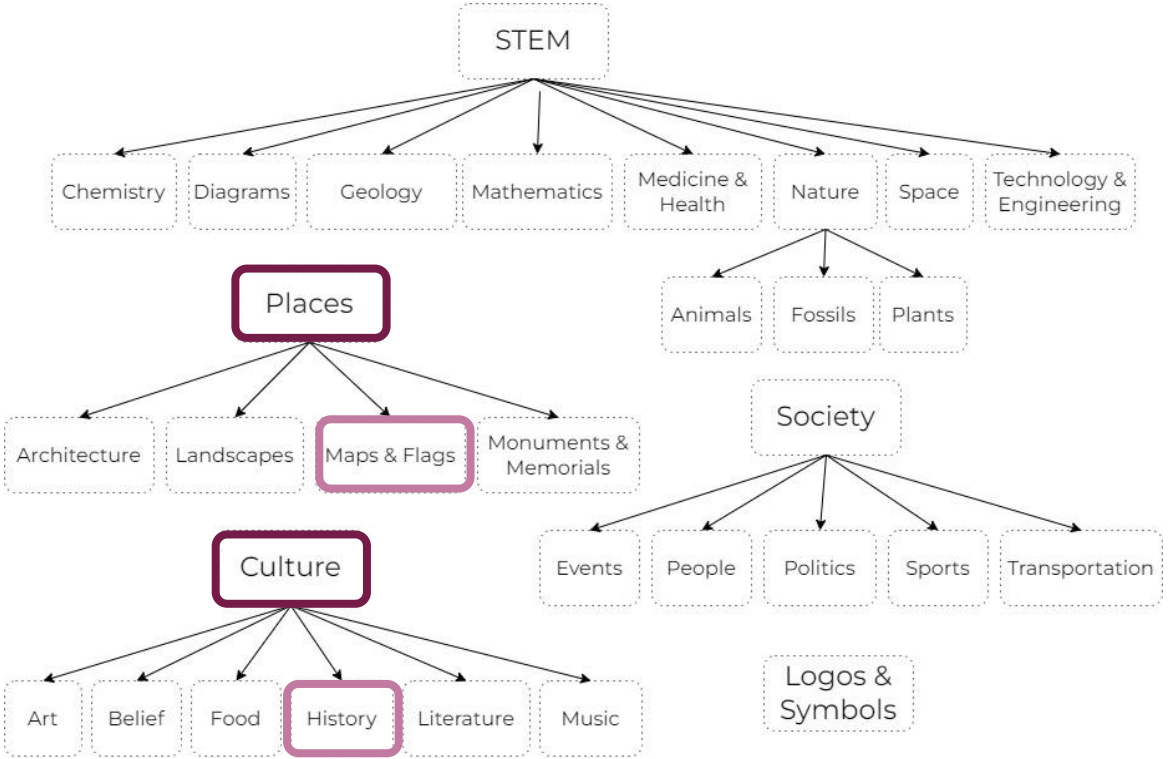
Baseline model (hierarchical)



RQ1: How to mitigate the consequences of the imbalanced data?

RQ1

First: ***flat*** or ***hierarchical*** data?



RQ1

First: ***flat*** or ***hierarchical*** data?

Table 4.1: *AUC(PR)* when evaluating the trained models on a held-out test data. Training is performed with flat and hierarchical models, with flat or hierarchical data.

		MODEL	
		Flat	**Hierarchical**
DATA	**Flat**	0.271	0.179
	Hierarchical	**0.303**	**0.210**

→ ***Hierarchical* data** overperforms *flat* data: ✓

- the data is enriched with correct labels;
- the ratio of positive samples in the data increase.

Ways to mitigate imbalance: **data** and **algorithm**

RQ1

Techniques – data-level

Table 4.2: Performance metrics when using data-level techniques for mitigating imbalance.

Classifier	*AUC(PR)*	Precision | Recall		
		All labels	Top 5	Rest
Baseline	**0.254**	**0.408** | 0.142	**0.652** | 0.448	**0.362** | 0.083
Naïve augmentation	0.200	0.345 | 0.098	0.590 | 0.276	0.280 | 0.053
Under-sampling	0.233	0.367 | 0.139	0.646 | 0.457	0.313 | 0.078
Oversampling + augmentation	0.158	0.198 | **0.203**	0.488 | **0.613**	0.142 | **0.124**

- → **Naive augmentation is bad**, especially on rarer labels. ❌
- → **Under-sampling keep 80% retains 90% performance** can be useful for saving resources. 🤔
- → **Oversampling + augmentation is bad**, contrary to what is said in literature. ❌

RQ1

Techniques – data-level

➔ Expected that resampling algorithms (under- and oversampling) underperform due to the high label concurrence in the data (0.51 *SCUMBLE*).

RQ1

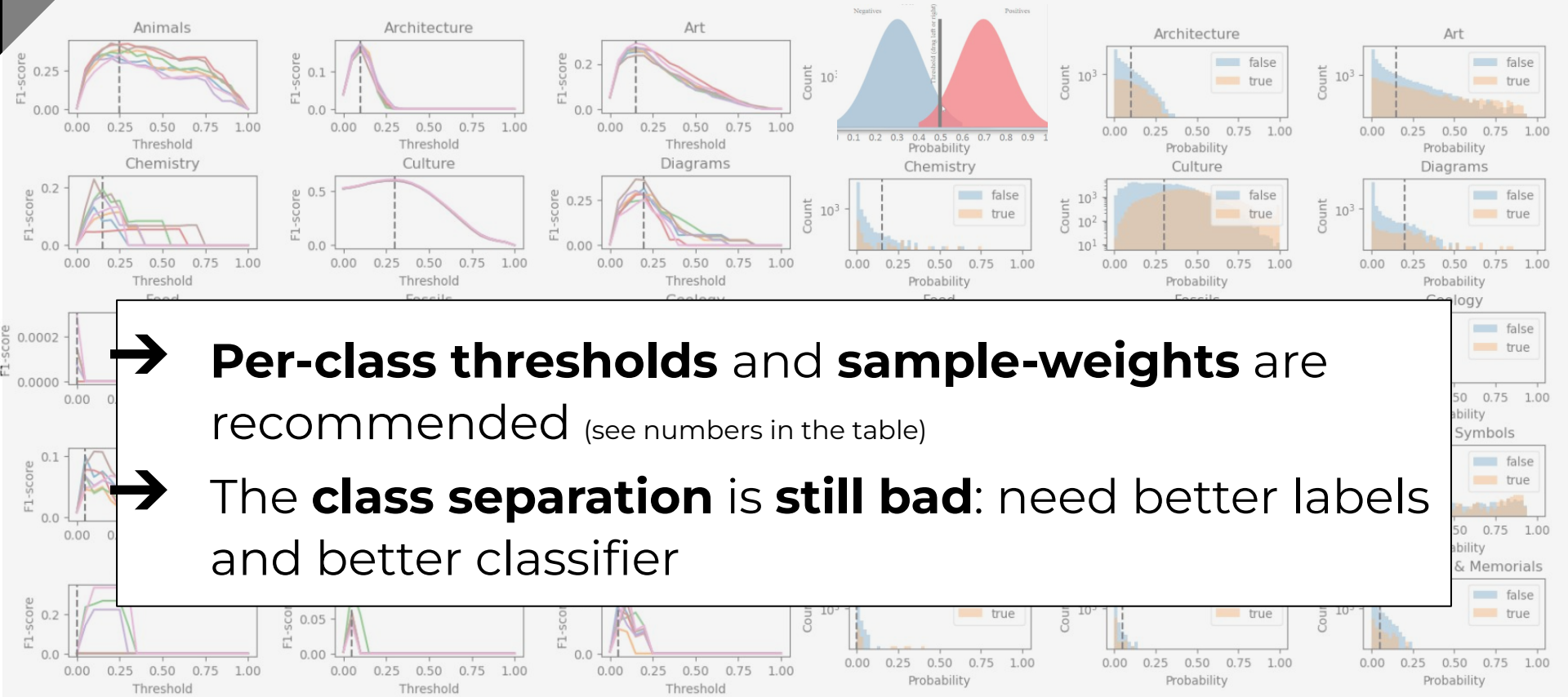
Techniques – algorithm-level

Table 4.3: Performance metrics when using algorithm-level techniques for mitigating imbalance.

Loss	*AUC(PR)*	Precision | Recall		
		All labels	Top 5	Rest
Baseline	0.254	0.408 | 0.142	0.668 | 0.487	0.358 | 0.076
Per-class thresholds	0.254	0.233 | **0.527**	0.488 | **0.800**	0.184 | **0.474**
Focal loss	0.252	0.373 | 0.144	0.668 | 0.489	0.316 | 0.078
Class-weights	0.208	0.267 | 0.185	0.556 | 0.486	0.211 | 0.127
Sample-weights	**0.263**	**0.420** | 0.156	**0.674** | 0.517	**0.371** | 0.087

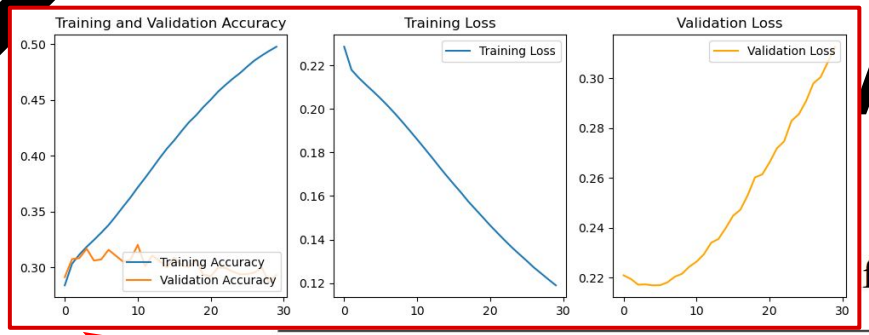
- ➔ Per-class thresholds boost recall (and overall F1-score) 🤔
- ➔ Focal-loss didn't have any greater effect ❌
- ➔ Class-weights decreased *AUC(PR)* and precision of rarer classes ❌
- ➔ Sample-weights is the only to increase *AUC(PR)* ✅

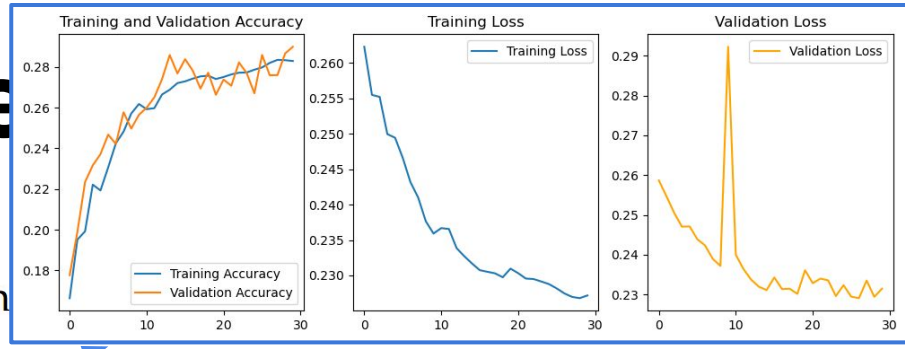
RQ1



- → **Per-class thresholds** and **sample-weights** are recommended (see numbers in the table)
- → The **class separation** is **still bad**: need better labels and better classifier

RQ1





We

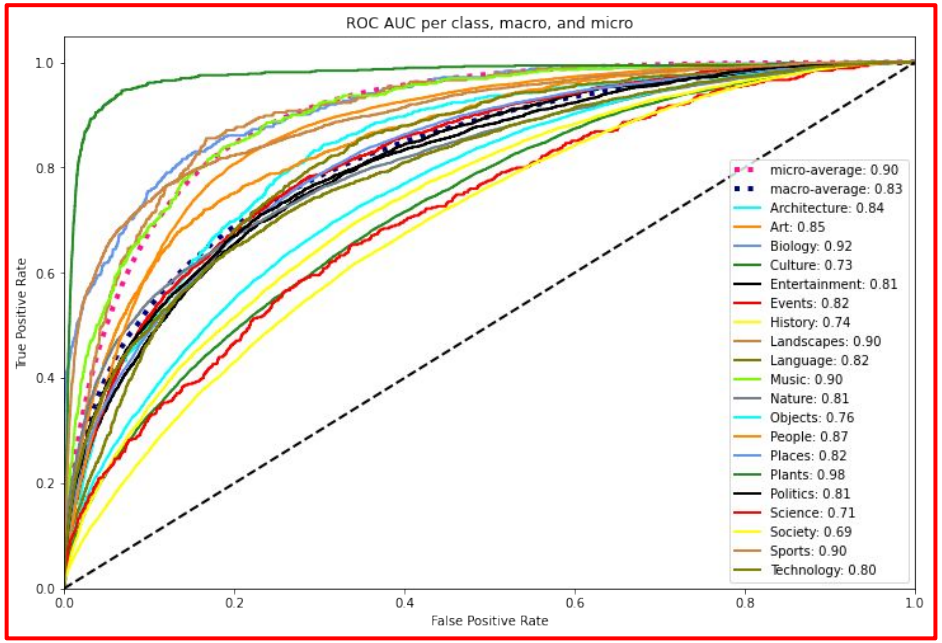
fferen

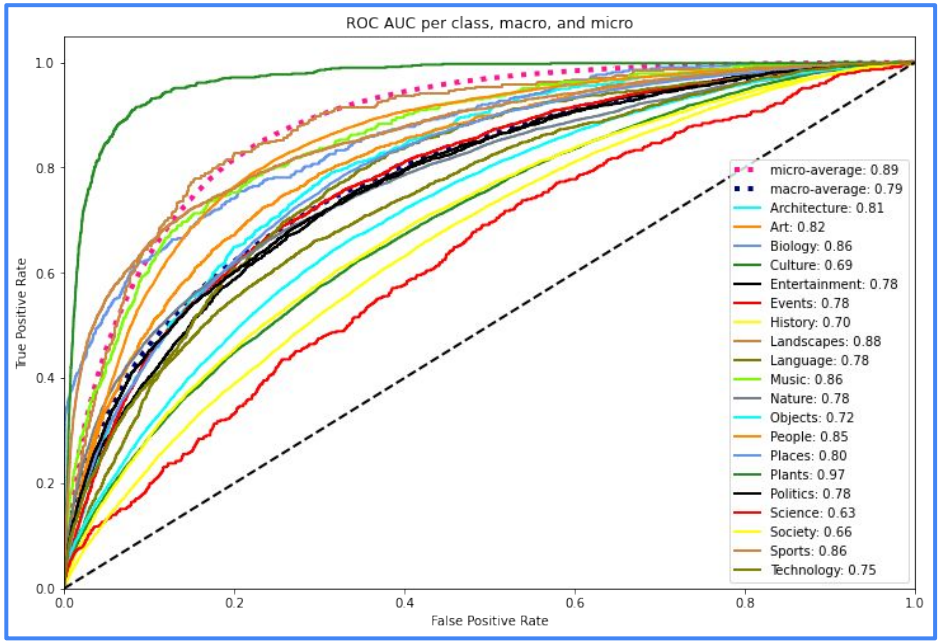
Initial weights	Labels per image	F1-score		
		All labels	Top 5	Rest
ImageNet	**1.024**	**0.267**	**0.450**	**0.206**
Random	0.972	0.192	0.415	0.118

➔ Transfer learning from **ImageNet** yields better results but **overfits early**

🤔

RQ1





→ **Bias** from ImageNet pre-training is **not** observed

(Plants still the greatest AUC, followed by Biology, People, Music, etc)

RQ1

Fine-tuning all layers

Table 4.3: Effects of fine-tuning the 339-layers base model; F1-scores

Trainable layers	Trainable params	Labels per image	F1-scores		
			All labels	Top 5	Rest
0	723,604	0.922	0.186	0.386	0.119
3	1,222,036	0.976	0.209	0.414	0.141
10	2,340,076	0.912	0.195	0.399	0.127
70	5,722,684	1.064	0.241	0.441	0.175
100	6,900,894	1.079	0.242	0.438	0.176
200	8,187,424	1.132	0.252	0.450	0.186
339	8,424,598	**1.165**	**0.262**	**0.469**	**0.193**

→ **Fine-tuning all layers** is clearly **good**: consistent increase of metrics when training more layers.

RQ1

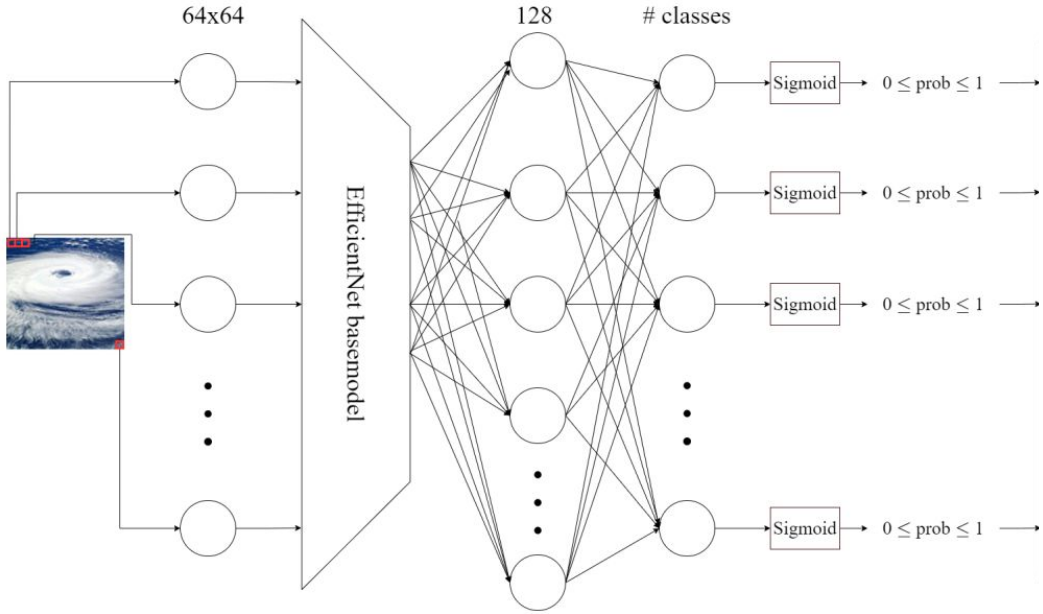
RQ1: How to mitigate the consequences of the imbalanced data?

Answer: Sample-weighting, per-class thresholds, using parameters pre-trained on ImageNet. Heuristic undersampling can be used to retain performance while requiring less data.

RQ2: How does a hierarchical classification perform compared to a flat one?

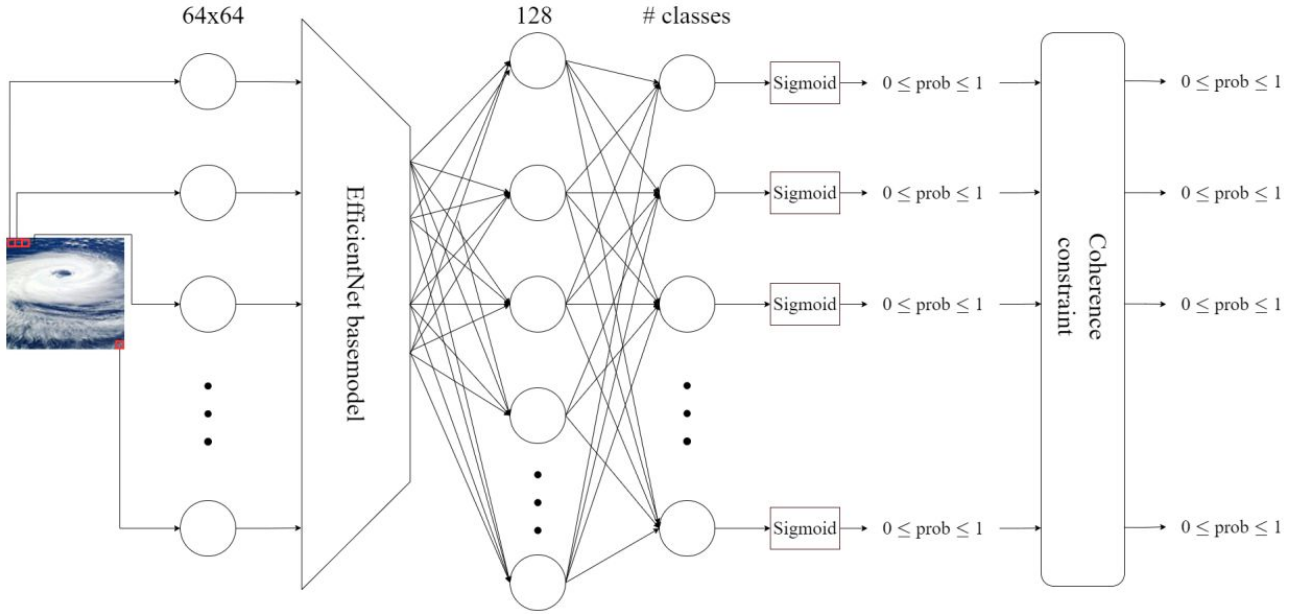
RQ1

Baseline model (flat)



RQ1

Baseline model (hierarchical)



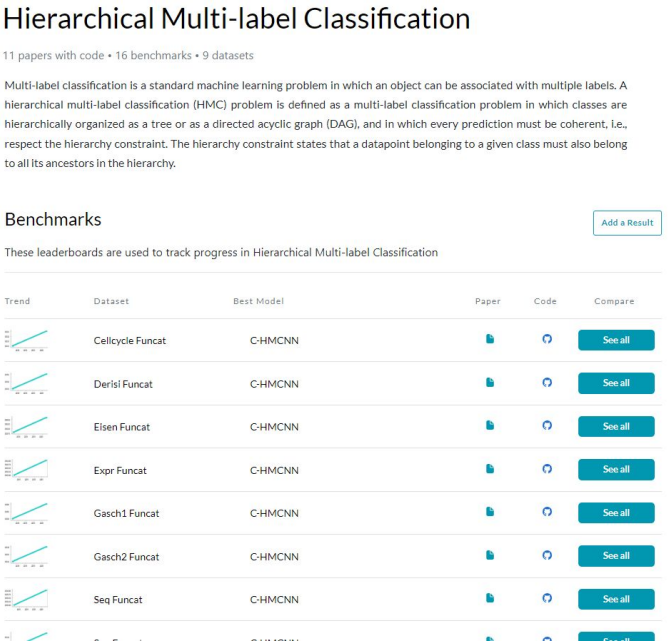
➔ The coherence constraint block performs a matrix operation enforcing any parent labels to be given equal probability as all its children.

Coherent Hierarchical Multi-Label Classification Networks, Giunchiglia et al. (2021)

RQ1

Baseline model (hierarchical)

➔ *CHMCNN(h)*: state-of-the-art on hierarchical multilabel classification



Hierarchical Multi-label Classification

11 papers with code • 16 benchmarks • 9 datasets

Multi-label classification is a standard machine learning problem in which an object can be associated with multiple labels. A hierarchical multi-label classification (HMC) problem is defined as a multi-label classification problem in which classes are hierarchically organized as a tree or as a directed acyclic graph (DAG), and in which every prediction must be coherent, i.e., respect the hierarchy constraint. The hierarchy constraint states that a datapoint belonging to a given class must also belong to all its ancestors in the hierarchy.

Benchmarks

Add a Result

These leaderboards are used to track progress in Hierarchical Multi-label Classification

Trend	Dataset	Best Model	Paper	Code	Compare
	Cellcycle Funcat	C-HMCNN			See all
	Derisi Funcat	C-HMCNN			See all
	Elsen Funcat	C-HMCNN			See all
	Expr Funcat	C-HMCNN			See all
	Gasch1 Funcat	C-HMCNN			See all
	Gasch2 Funcat	C-HMCNN			See all
	Seq Funcat	C-HMCNN			See all

Coherent Hierarchical Multi-Label Classification Networks, Giunchiglia et al. (2021)

RQ2

Hierarchical classification

Table 4.6: Flat vs. hierarchical models.

Model	*AUC(PR)*	Precision / Recall					
		All labels		Top 5		Rest	
Flat	**0.310**	**0.518**	**0.190**	0.686	**0.535**	**0.486**	**0.124**
Hierarchical	0.237	0.401	0.132	**0.723**	0.421	0.339	0.076

➔ The *flat* model performs significantly better than the *hierarchical* one, proving wrong our initial hypothesis.

RQ2

RQ2: How does a hierarchical classification perform compared to a flat one?

Answer: The *hierarchical* model performs significantly worse than the *flat* one. Believed reason: the hierarchical model is not adapted for image datasets as WIT. It had previously only been used on *functional genomics* datasets.

RQ3: Using the best developed model, what insights can be drawn from images of Wikipedia?

RQ3

Data study – model performance

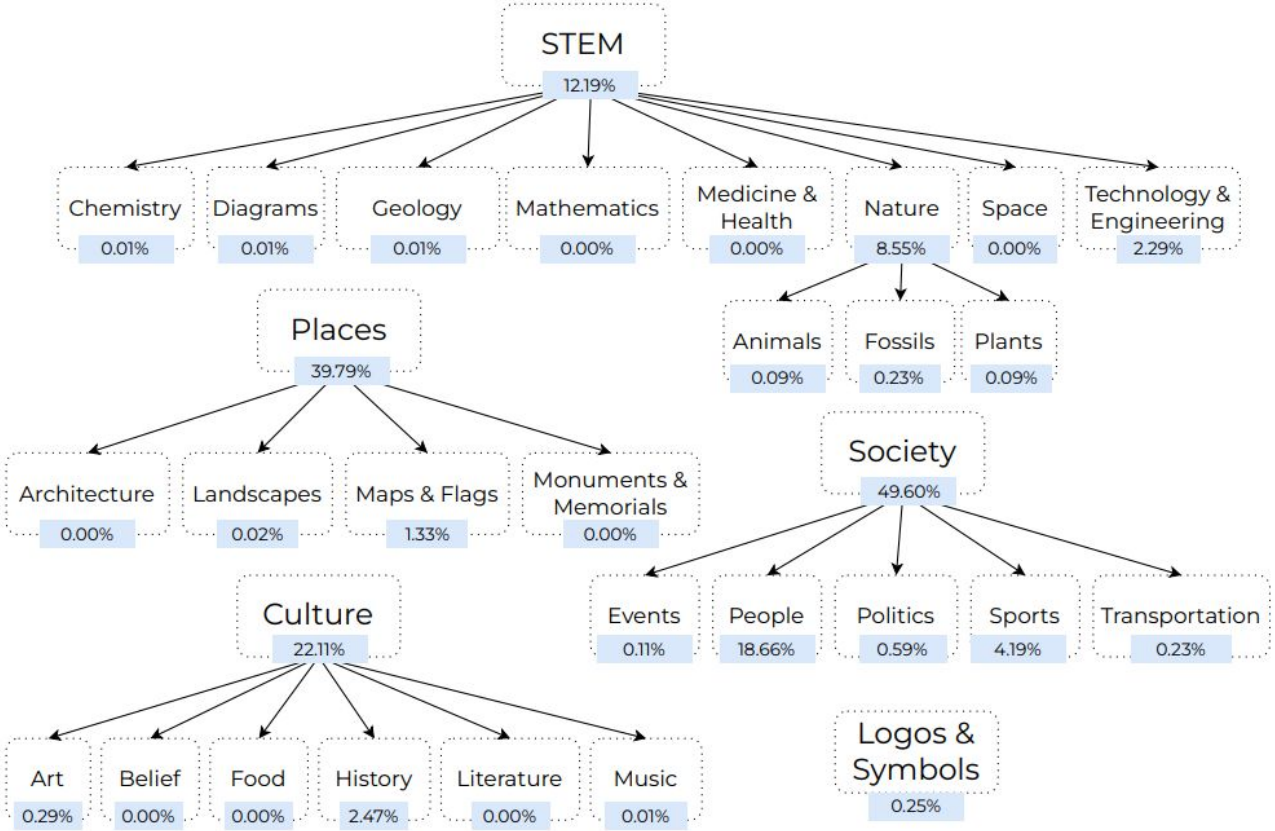
➔ When predicting on the training set itself (760*k* images), we get *(*not as bad as it looks)*

Table 5.1: Performance metrics on predictions on the training data.

Classifier	***AUC(PR)***	**Precision | Recall**		
		All labels	Top 5	Rest
Flat	0.349	0.525 | 0.229	0.722 | 0.578	0.487 | 0.162
Flat (per-class thresh)	0.349	0.335 | 0.524	0.601 | 0.784	0.283 | 0.474

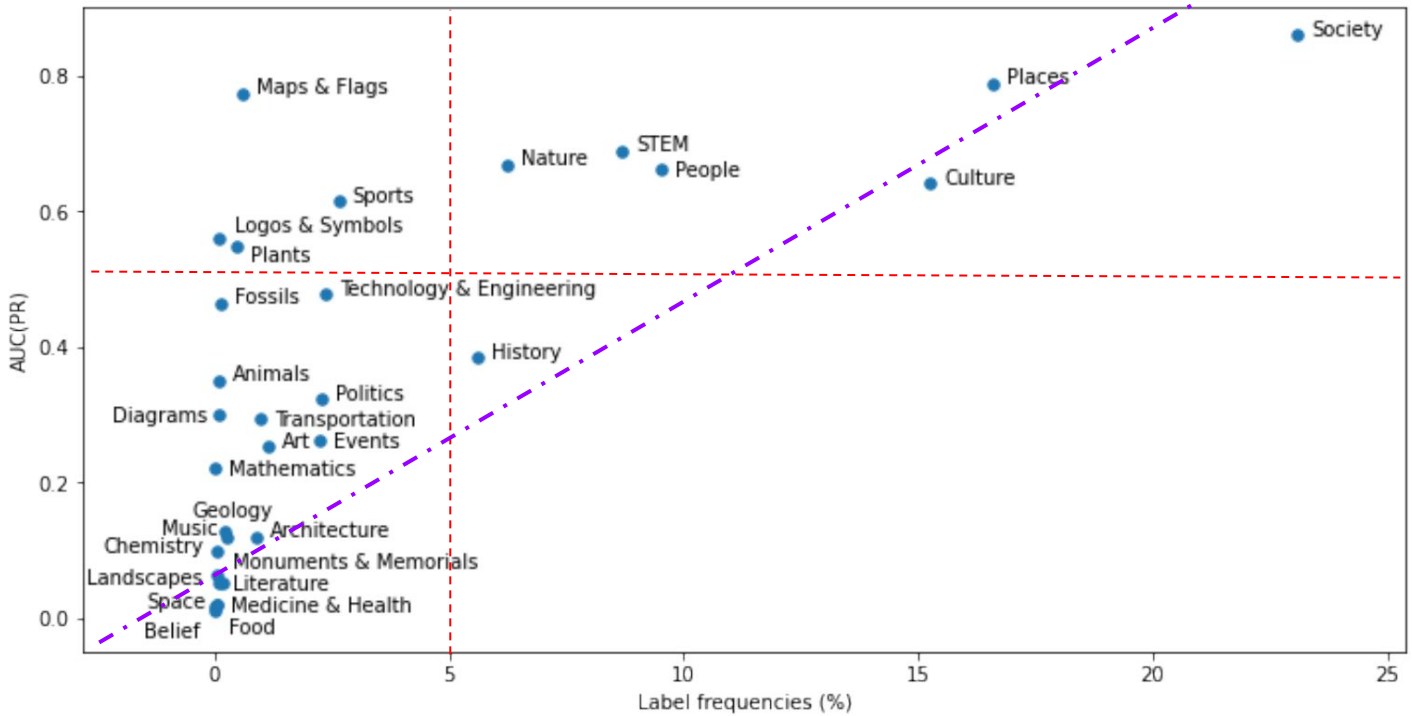
RQ3

Data study – predicted label distribution



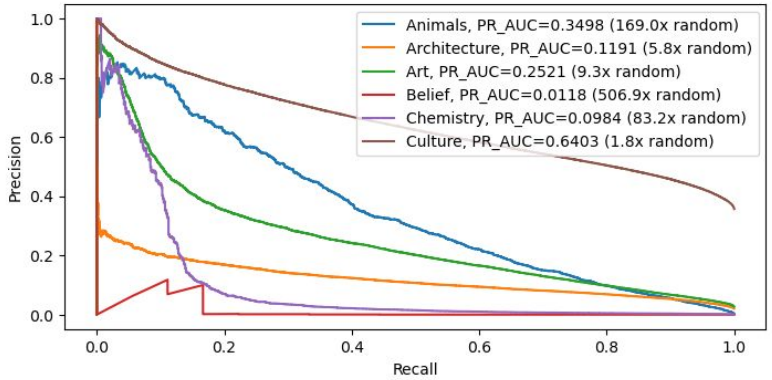
RQ3

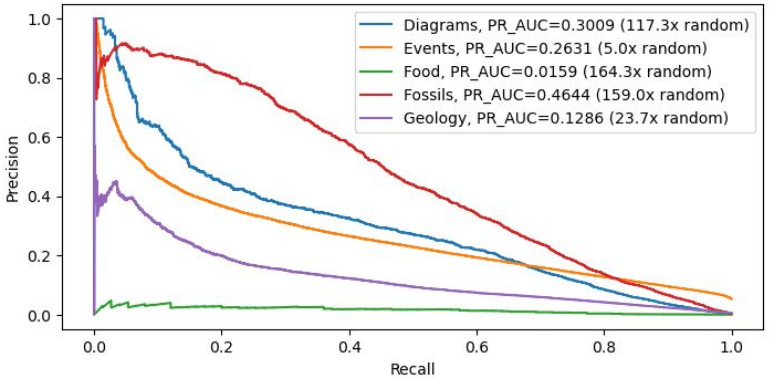
Data study – label frequency & performance

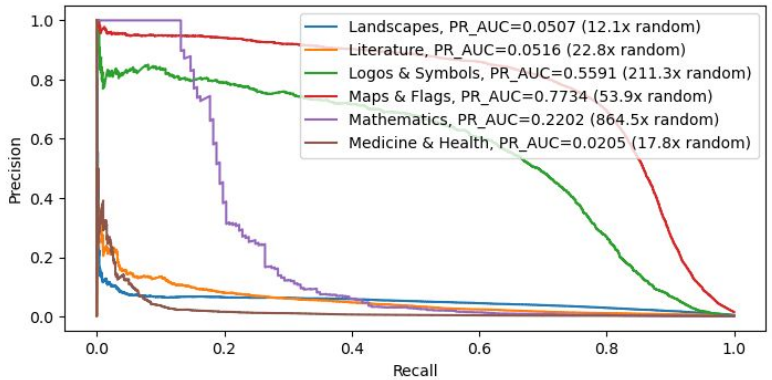


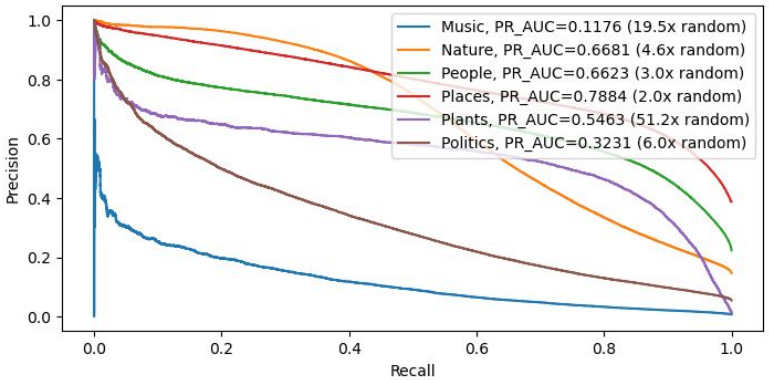
RQ3

Data study – PR-curves









RQ3

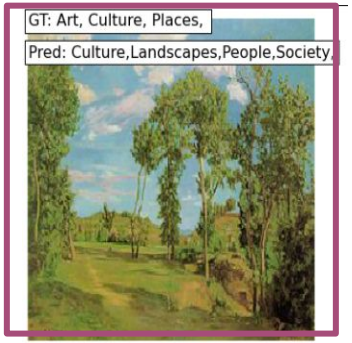
RQ3: Using the best developed model, what insights can be drawn from images of Wikipedia?

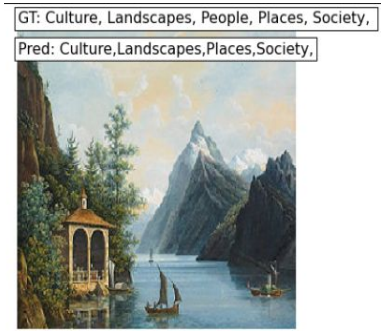
Answer: The predicted label distribution as given by the image. An interesting insight is the outlier performance of labels such as Maps & Flags, Logos & Symbols, Sports, and Plants. These are either due to more signal in the label or due to bias from pre-training.

Proof of use-case

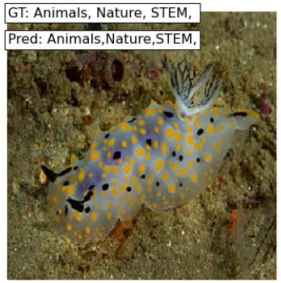
➔ Sometimes the predicted labels capture correct labels missing in the ground truth

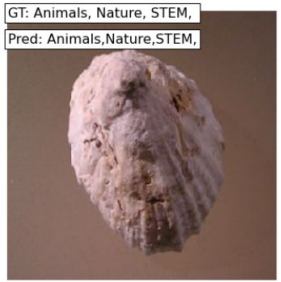
Landscape





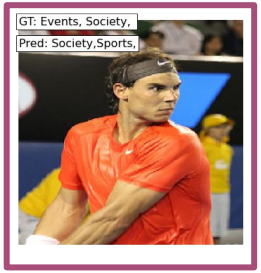
Animals

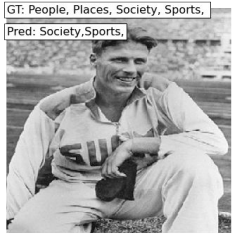




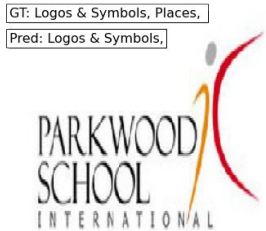
Politics

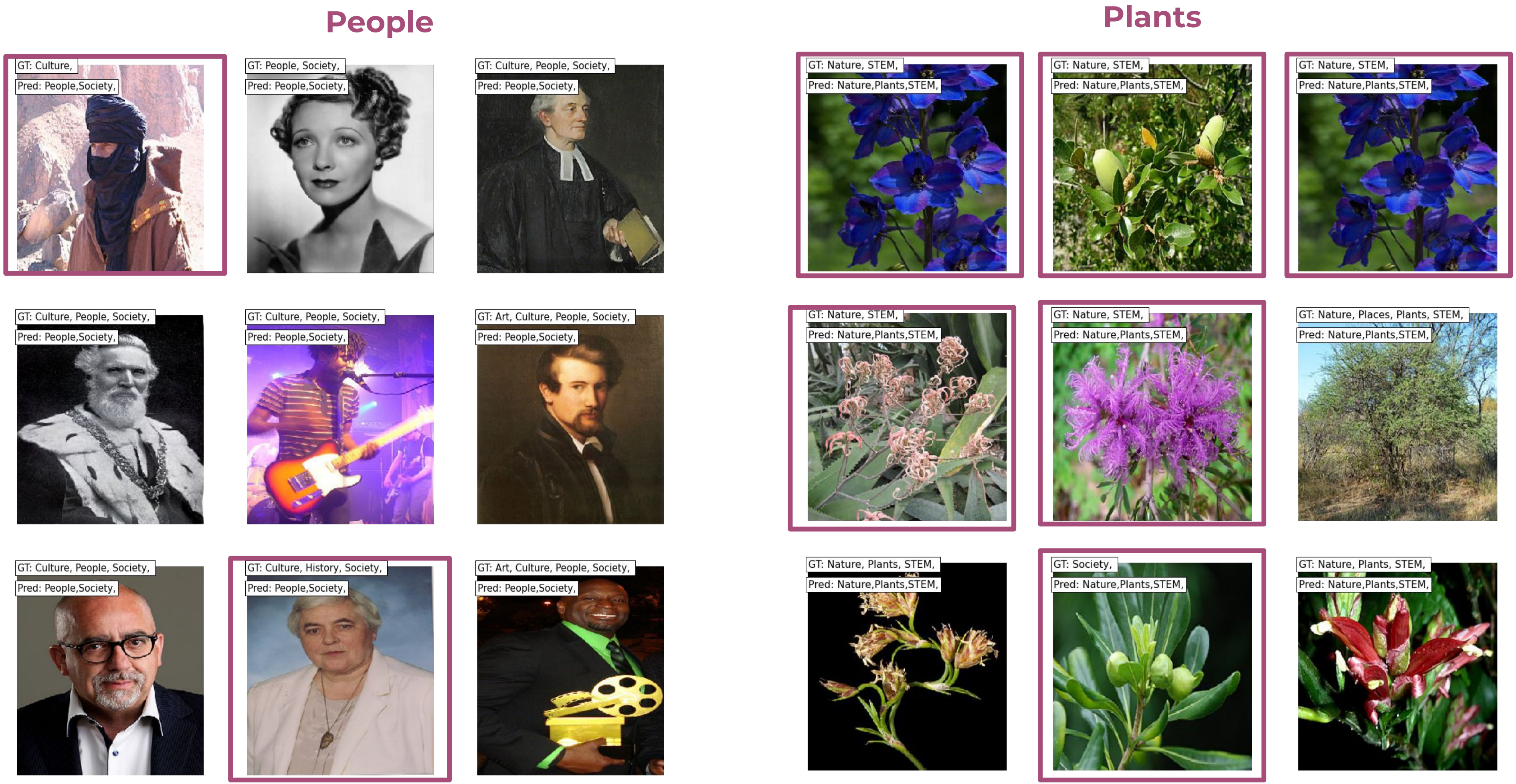
Sports





Logos & symbols




People
GT: Culture,
Pred: People,Society,
GT: People, Society,
Pred: People,Society,
GT: Culture, People, Society,
Pred: People,Society,
GT: Culture, People, Society,
Pred: People,Society,
GT: Culture, People, Society,
Pred: People,Society,
GT: Art, Culture, People, Society,
Pred: People,Society,
GT: Culture, People, Society,
Pred: People,Society,
GT: Culture, History, Society,
Pred: People,Society,
GT: Art, Culture, People, Society,
Pred: People,Society,
Plants
GT: Nature, STEM,
Pred: Nature,Plants,STEM,
GT: Nature, STEM,
Pred: Nature,Plants,STEM,
GT: Nature, STEM,
Pred: Nature,Plants,STEM,
GT: Nature, STEM,
Pred: Nature,Plants,STEM,
GT: Nature, STEM,
Pred: Nature,Plants,STEM,
GT: Nature, Places, Plants, STEM,
Pred: Nature,Plants,STEM,
GT: Nature, Plants, STEM,
Pred: Nature,Plants,STEM,
GT: Society,
Pred: Nature,Plants,STEM,
GT: Nature, Plants, STEM,
Pred: Nature,Plants,STEM,

Discussion

➔ One could ask: if you already have the ground truth labels providing the topic distribution of Wikipedia, why would you want to train a model to predict the topic distribution of Wikipedia? The advantages of the model compared to the ground truth labels are:

- Infer correct labels not present in the ground truth labels, thus increasing recall
- Return topical *distribution*, rather than only *prediction*

Discussion

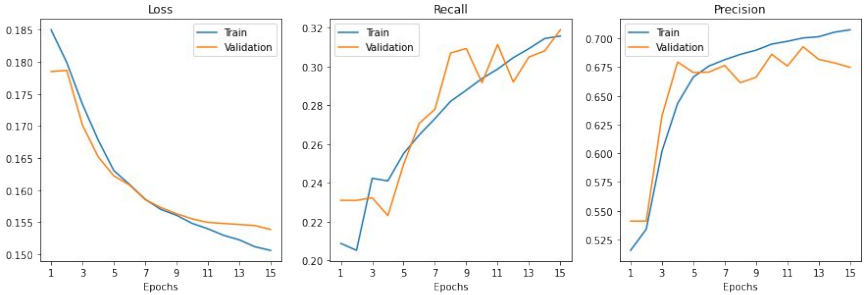
➔ The predicted distribution should only give us intuitions of the topical distribution of images in Wikipedia; the data separation and the classifier are not yet good enough to call the predictions a picture of the reality.

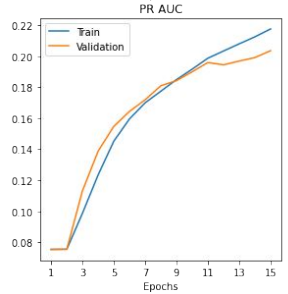
Future work

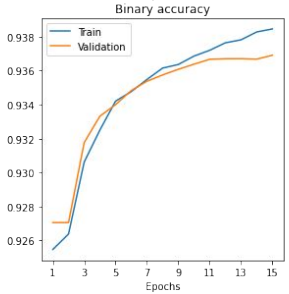
- ➔ Try with more classical methods for hierarchical classification, e.g., XGBoost (used in ORES)
- ➔ Compare the performance of the model with human-labeled ground truth
- ➔ Using filename as a feature for multimodal prediction

Thanks!

Hierarchical model, hierarchical data







Flat model, hierarchical data