Few-shot fungi classification

Computer Vision

Antonio Cordeiro
Lorenzo Ugolini

SAPIENZA
UNIVERSITÀ DI ROMA

Table of Contents

- Introduction & Paper
- The Project pt. 1
- Background
- The Project pt. 2
- Conclusion

Introduction

This work is based on the Kaggle competition **FungiCLEF25**

Few-shot recognition of fungi species using real-world observational data.

- multiple photographs of the same specimen
- metadata
- satellite imagery
- meteorological variables

Automatic species recognition aids mycologists, scientists, and nature enthusiasts in identifying species in the wild.

Efficiently predict species with limited resources and handle many classes, some of which have just a few recorded observations.

20% of verified observations involve rare or under-recorded species

FungiTastic: A Multi-Modal Dataset and Benchmark for Image Categorization[1]

Biological problems provide a natural, challenging setting for benchmarking image classification methods:

- seasonality/evolution i.e. domain shift
- few samples for rare species
- “unknown” class option

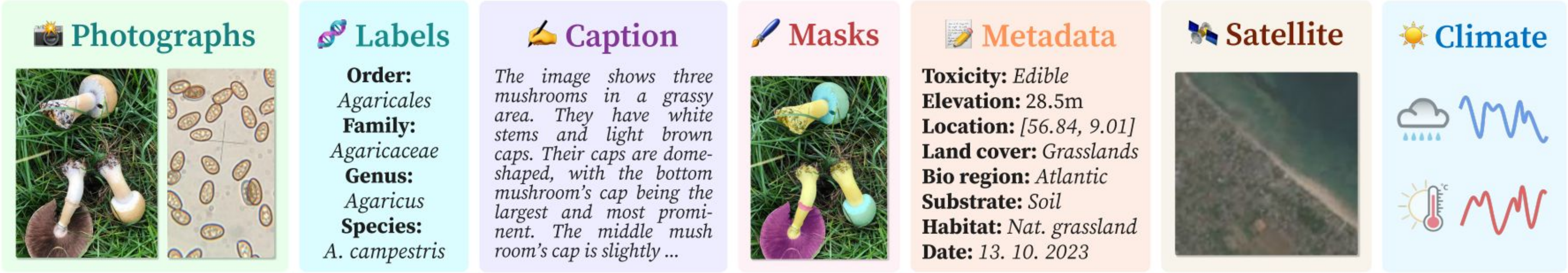
FungiTastic is a complex, real-world, multi-modal dataset collected over 20 years

Labelled and curated by experts

- about 350k multimodal observations
- 6k fine-grained categories (species)

[1] https://arxiv.org/pdf/2408.13632

Each observation



Order: *Agaricales*
Family: *Agaricaceae*
Genus: *Agaricus*
Species: *A. campestris*

The image shows three mushrooms in a grassy area. They have white stems and light brown caps. Their caps are dome-shaped, with the bottom mushroom's cap being the largest and most prominent. The middle mush room's cap is slightly ...

Toxicity: *Edible*
Elevation: 28.5m
Location: *[56.84, 9.01]*
Land cover: *Grasslands*
Bio region: *Atlantic*
Substrate: *Soil*
Habitat: *Nat. grassland*
Date: *13. 10. 2023*

- one or more photos of an observed specimen,occasionally a microscopic image of spores
- textual captions, observation metadata , geospatial data
- climatic time-series data
- for a subset (~70k photos), body part segmentation masks

Existing datasets problems

Datasets for classification of data originating in nature are typically artificially sampled, solely image-based, and focused on traditional image classification. Many popular datasets also suffer from specific limitations that compromise their generalizability and robustness. Common issues include:

- Lack of Multi-Modal Data
- Biases in Data Representation
- Biases in Data Representation
- Labeling Errors and Quality Control

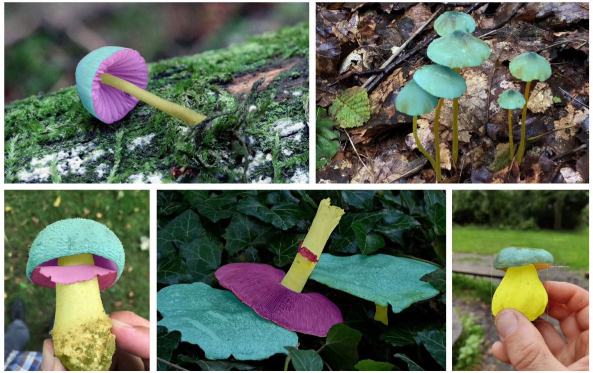
The FungiTastic Benchmark

- FungiTastic:
    - around 346k observations of 4,507 species
- FungiTastic–Mini (FungiTastic–M)
    - all observations belonging to 6 hand-picked genera (e.g., Russula, Boletus, Amanita, Clitocybe, Agaricus, and Mycena)
    - comprises 67,848 images (36,287 observations) of 253 species
- FungiTastic–FS
    - species with less than 5 observations in the training set, which were removed from the main (FungiTastic) dataset
    - contains 6,391 observations encompassing 12,015 images of a total of 2,427 species

Additional Observation Data

- Tabular metadata
- Remote sensing data
- Meteorological data
- Body part segmentation masks
- Image captions

Its stem is thick and light brown, with a hint of green at the base. The smaller mushroom on the right has a similar light brown cap, but its rim is more pronounced and has a white, almost translucent appearance. This gives it a delicate, lacy look. The stem of this mushroom is thinner and lighter in color compared to ......



- Body part segmentation masks
- Multi-band remote sensing data
- Meteorological data
- Image captions
- Location-related metadata

FungiTastic Benchmarks

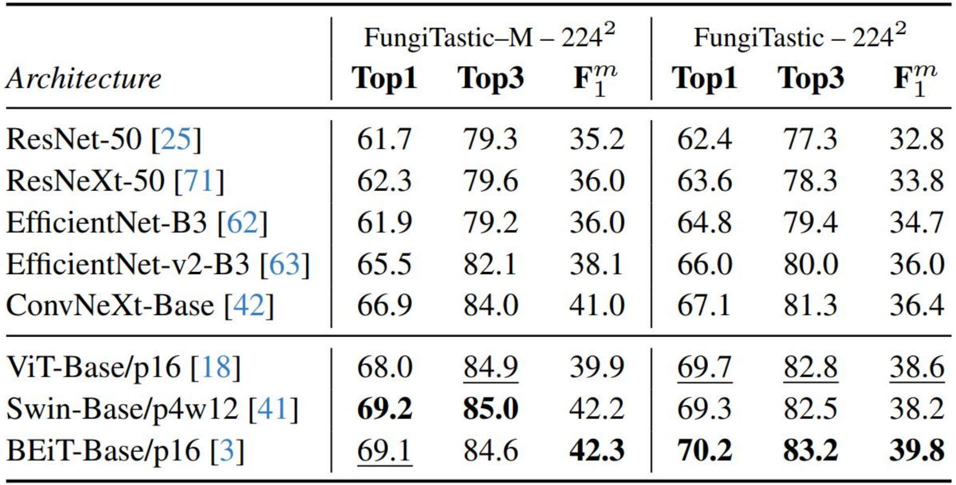
Evaluation of various fundamental computer vision and machine learning problems

- **Closed-set classification**: standard identification. Difficult due to a long-tailed distribution (many rare species) and high intra-class variance. F_1^m*-score*
- **Open-set classification**: recognize a species it has never seen before and label it as "unknown". *AUC* and TNR^{95}
- **Chronological classification**: each observation has a timestamp, allowing the study of species distribution changes over time; a model can access all observations with timestamps $t' < t$
- **Segmentation**: enables analysis of species-specific morphological and environmental relationships and reveals ecological and morphological patterns across locations. *mIoU* and *mAP*
- **Few-shot classification**: learn to identify a species from fewer than 5 examples, crucial to recognize rare fungi. F_1^m*-score* and *Top3 accuracy*

Evaluation of classification networks is typically based on the 0–1 loss function.
In practice not all errors are equal

Closed-set Image Classification

Trained a variety of state-of-the-art CNN architectures to establish some baselines for closed-set classification on the FungiTastic and FungiTastic–M.



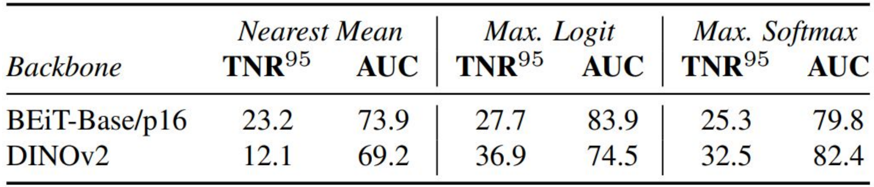
	FungiTastic–M – 224^2			FungiTastic – 224^2		
Architecture	**Top1**	**Top3**	$\mathbf{F}_1^m$	**Top1**	**Top3**	$\mathbf{F}_1^m$
ResNet-50 [25]	61.7	79.3	35.2	62.4	77.3	32.8
ResNeXt-50 [71]	62.3	79.6	36.0	63.6	78.3	33.8
EfficientNet-B3 [62]	61.9	79.2	36.0	64.8	79.4	34.7
EfficientNet-v2-B3 [63]	65.5	82.1	38.1	66.0	80.0	36.0
ConvNeXt-Base [42]	66.9	84.0	41.0	67.1	81.3	36.4
ViT-Base/p16 [18]	68.0	84.9	39.9	69.7	82.8	38.6
Swin-Base/p4w12 [41]	**69.2**	**85.0**	42.2	69.3	82.5	38.2
BEiT-Base/p16 [3]	69.1	84.6	**42.3**	**70.2**	**83.2**	**39.8**

Open-set Image Classification

Open-set classification as a binary decision-making problem, where the model determines whether a new image belongs to a known class or a novel class. We evaluate several approaches for open-set classification:

- Maximum Softmax Probability (MSP)
- Maximum Logit Score (MLS)
- Nearest Mean Score (NM)

Used features and logits from the BEiT-Base closed-set classifier baseline, trained on the full dataset (i.e., FungiTastic) and compare the fully-supervised model with generic features from a pretrained DINOv2 model.



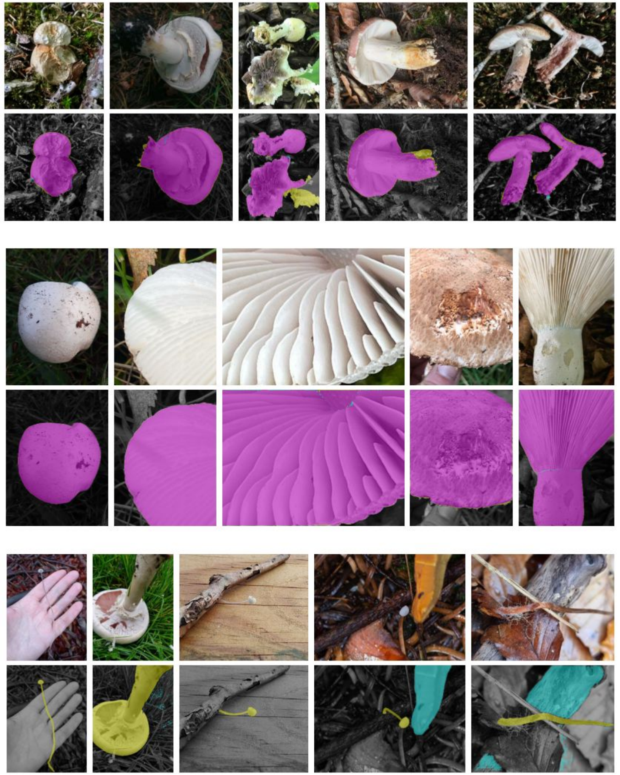
	Nearest Mean		*Max. Logit*		*Max. Softmax*	
Backbone	**TNR**^{95}	**AUC**	**TNR**^{95}	**AUC**	**TNR**^{95}	**AUC**
BEiT-Base/p16	23.2	73.9	27.7	83.9	25.3	79.8
DINOv2	12.1	69.2	36.9	74.5	32.5	82.4

Segmentation

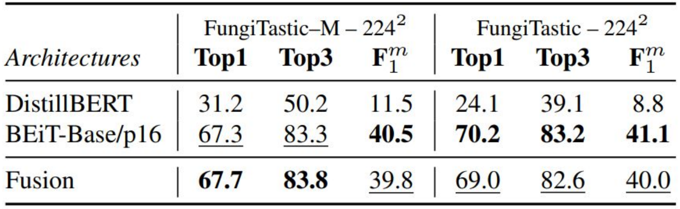
A zero-shot baseline for foreground-background binary segmentation of fungi is evaluated on the FungiTastic–M dataset.

1. *GroundingDINO* is prompted with the text ‘mushroom’ and outputs a set of instance-level bounding boxes
2. The bounding boxes from the first step are used as prompts for the SAM segmentation model

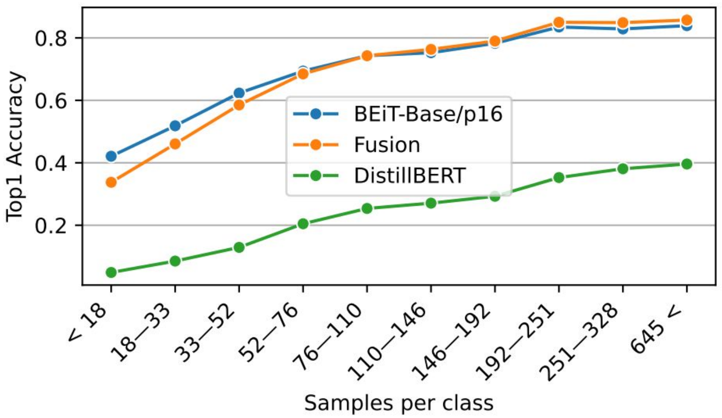
The baseline method achieved an average perimage IoU of 89.36%. While the model exhibits strong zero-shot performance, it sometimes fails to detect mushrooms.



Vision-Language Fusion



Architectures	FungiTastic–M – 224^2			FungiTastic – 224^2		
	Top1	**Top3**	$\mathbf{F}_1^m$	**Top1**	**Top3**	$\mathbf{F}_1^m$
DistillBERT	31.2	50.2	11.5	24.1	39.1	8.8
BEiT-Base/p16	67.3	83.3	**40.5**	**70.2**	**83.2**	**41.1**
Fusion	**67.7**	**83.8**	39.8	69.0	82.6	40.0



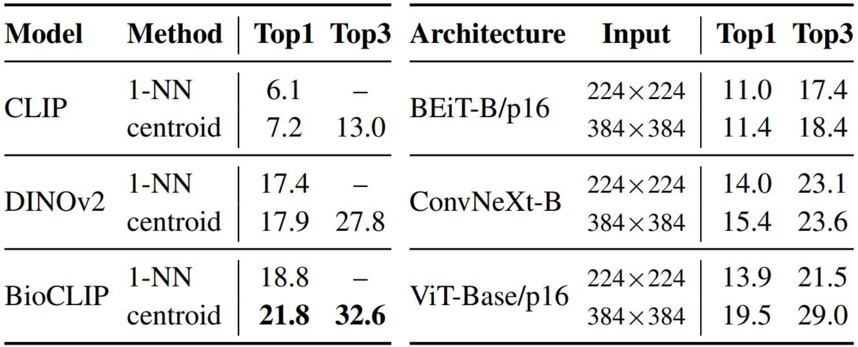
To evaluate the relevance of the available textual data for species classification, they provide baselines that use a sequence classification variant of the lightweight DistilBERT model trained as a classifier on textual descriptions only.

The model was trained using the standard cross-entropy loss, with logits obtained from a classification head applied to the pooled features of the class token in DistilBERT.

Few-shot Image Classification

1. Standard classifier training
2. nearest-neighbor classification
3. centroid prototype classification

For 2. and 3. deep embeddings are extracted from large-scale pre-trained vision models, namely CLIP, BioCLIP and DINOv2.



Model	Method	Top1	Top3
CLIP	1-NN	6.1	–
	centroid	7.2	13.0
DINOv2	1-NN	17.4	–
	centroid	17.9	27.8
BioCLIP	1-NN	18.8	–
	centroid	**21.8**	**32.6**

Architecture	Input	Top1	Top3
BEiT-B/p16	224×224	11.0	17.4
	384×384	11.4	18.4
ConvNeXt-B	224×224	14.0	23.1
	384×384	15.4	23.6
ViT-Base/p16	224×224	13.9	21.5
	384×384	19.5	29.0

Table of Contents

- Introduction & Paper
- The Project pt. 1
- Background
- The Project pt. 2
- Conclusion

Starting Point

Following the paper approach:

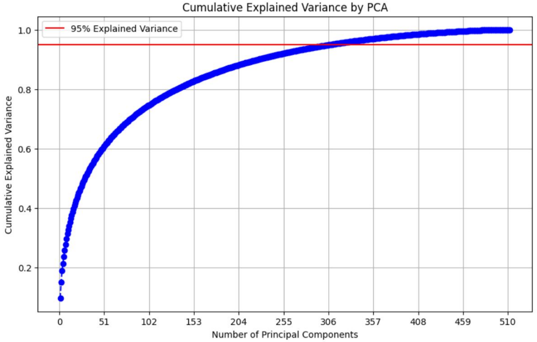
- Compute embeddings with bioCLIP
- Calculate Prototypes
- Classify finding the closest prototype

→

top-1 accuracy ≃ 15%
top-3 accuracy ≃ 25%

However, we can do better! (they thought…)

First improvements



Dimensionality reduction with PCA

top-1 accuracy ≃ 14%
top-3 accuracy ≃ 23%

Use metadata:

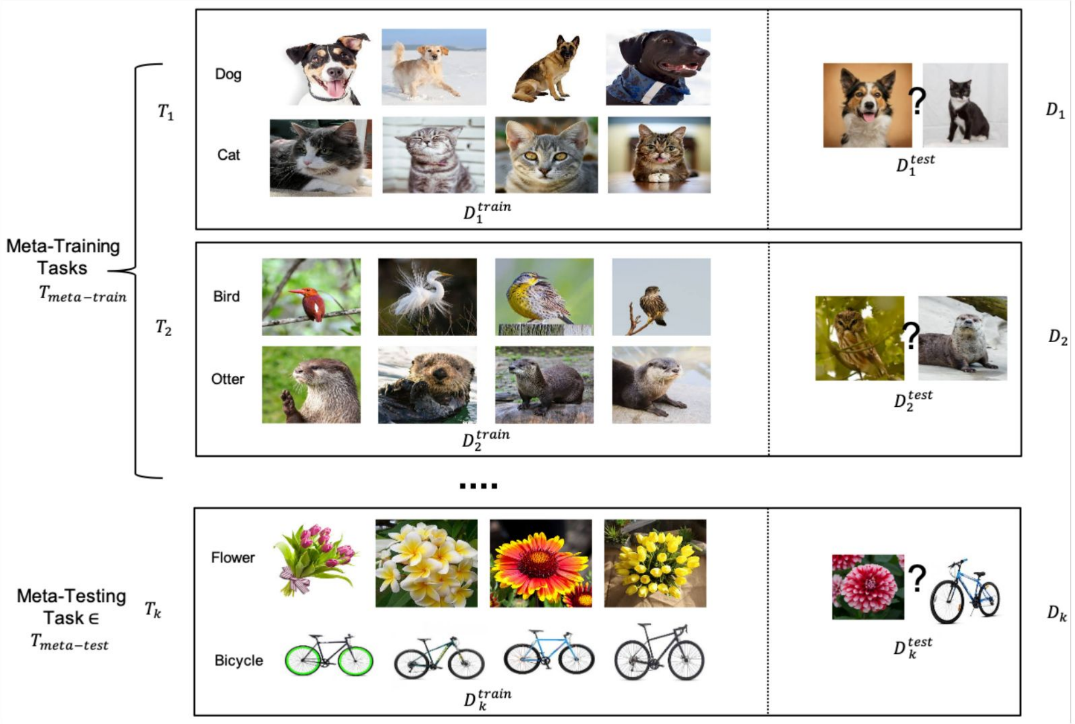
```
["eventDate", "habitat", "countryCode",
"hasCoordinate", "substrate", "latitude",
"longitude", "region", "district",
"metaSubstrate","elevation","landcover","
biogeographicalRegion"]
```

top-1 accuracy ≃ 12%
top-3 accuracy ≃ 20%

Table of Contents

- Introduction & Paper
- The Project pt. 1
- Background
- The Project pt. 2
- Conclusion

Meta-Learning and Few-shot classification

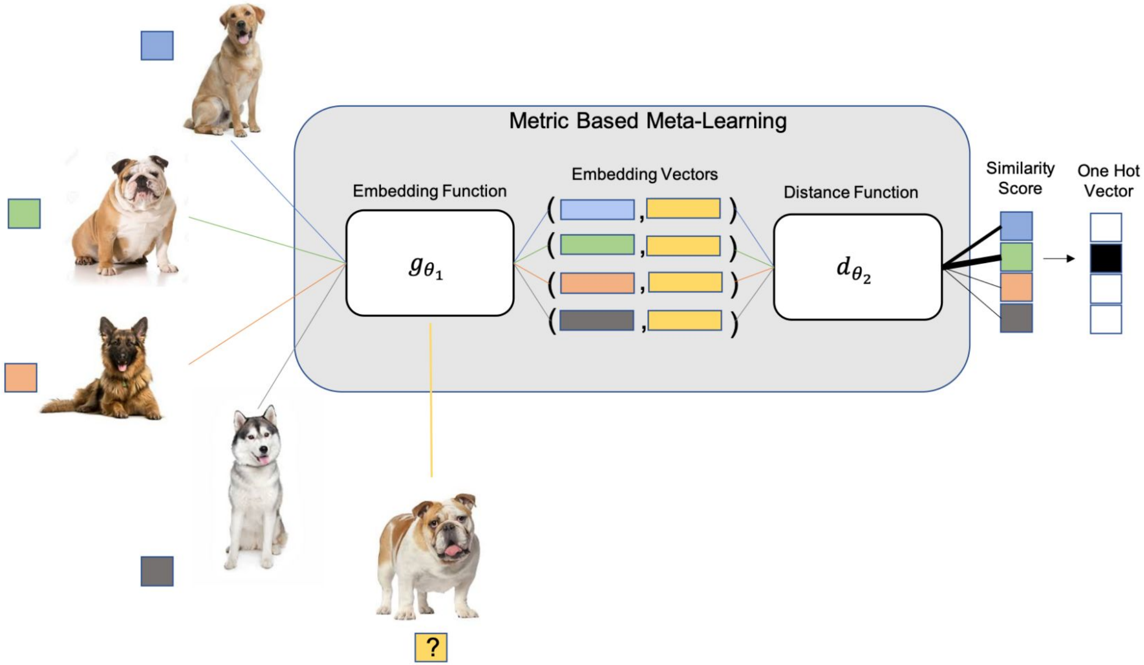


Few-shot classification aims to learn a model that can quickly adapt to a novel classification task given only few observations.

The few-shot classification task can be defined as a standard M-way-K-shot task, where M is the number of classes and K is the number of examples per class present in D^{train}.

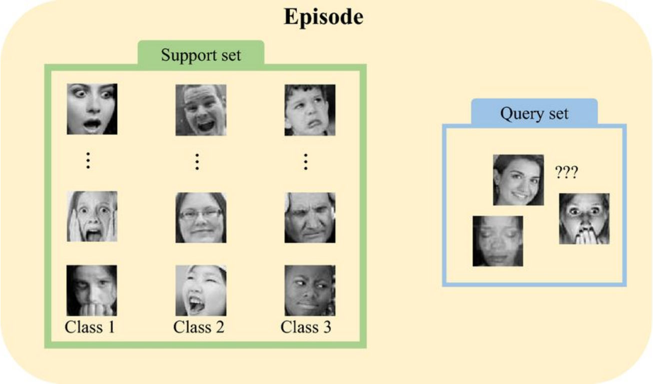
Metric-based Meta-Learning

Metric Learning is the task of learning a distance function over data samples.



Few-shot Episodic Training

- Training proceeds by randomly sampling M-way-K-shot episodes from the training set. Each episode has a support set and a query set
- The average error computed on query sets across multiple training few-shot episodes is used to update the parameters of the embedding function and the distance function (if any)
- Finally, new M-way-K-shot episodes are sampled from the testing set to evaluate the performance of the network



Prototypical Networks

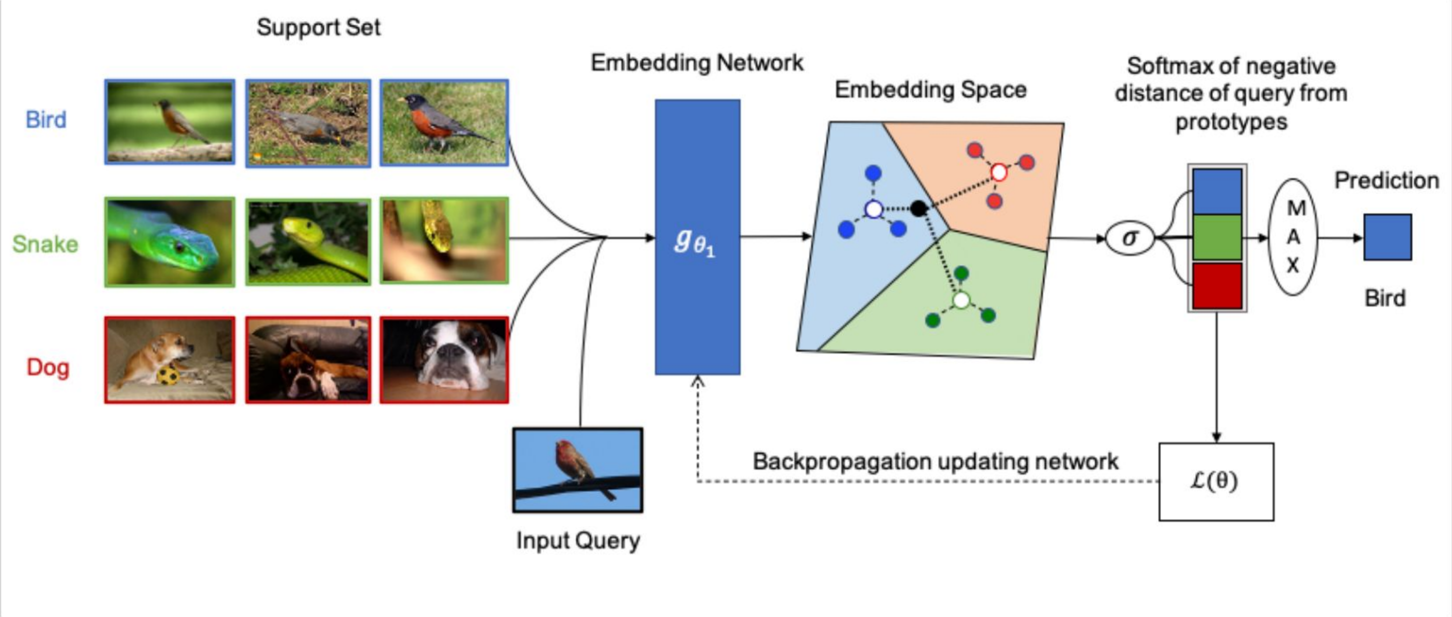


Table of Contents

- Introduction & Paper
- The Project pt. 1
- Background
- The Project pt. 2
- Conclusion

The FungiTastic–FS dataset

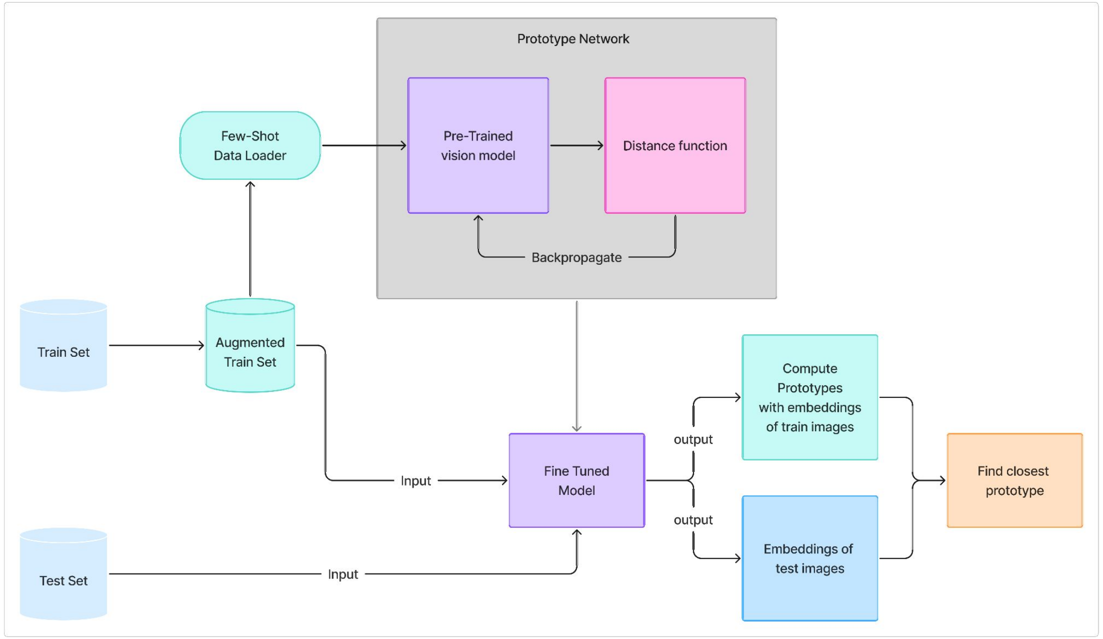
- The dataset is divided by the authors in a train, a test and validation set. Because of the Kaggle challenge, the test set provided does not come with labels. Thus, we used the given validation set as out test set.

- Each class in the dataset can have as low as one sample image per class. To build our few-task in a M-way K-shot setting we would need at least K images per class plus some query images.

- Dataset augmentation:
  - First we tried with simple transformations like rotate, flip, …
  - Then AutoAugment from pytorch

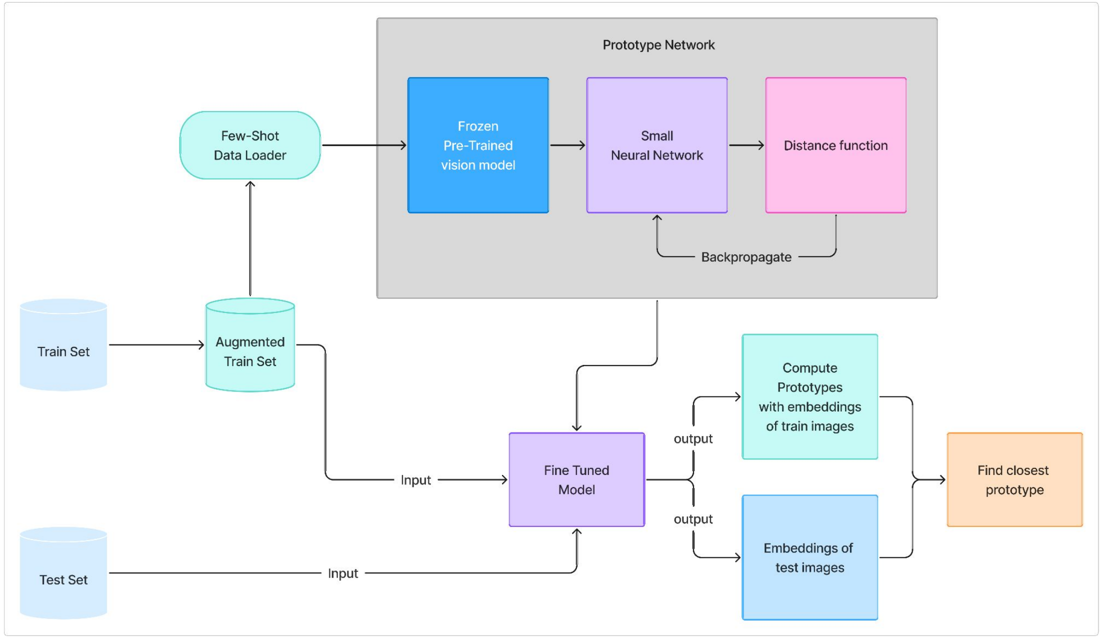
Our implementation

- Train a prototype network with different backbones for 5-way 1-shot and 5-shot:
  - ResNet-18
  - DinoV2
  - BioClip
- Only ResNet-18 seemed to properly train
- Around 5% accuracy

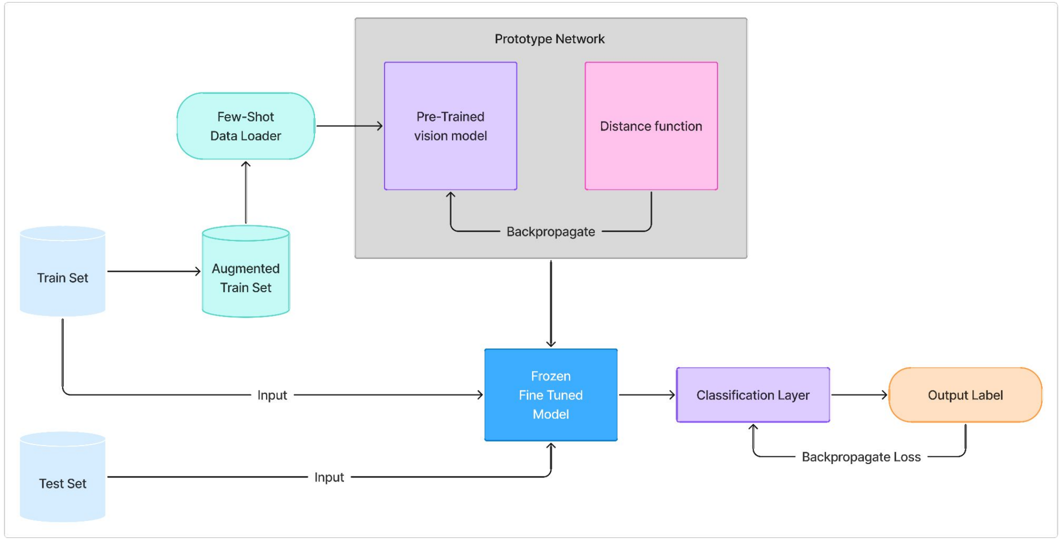


Our implementation

Even freezing the pre trained backbone models and training a small network did not improve the accuracy.



Our implementation



To make the model adapt to the classification task after the few-shot training, we tried to:

- take the model trained with few-shot and freeze it
- add a classification layer made of one linear layer
- train this classifier head for the final classification task

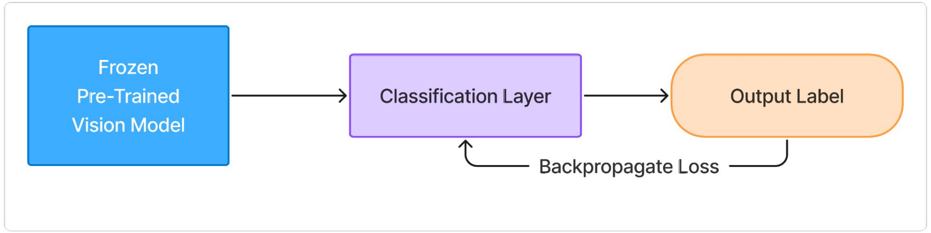
The performance did not show any improvement and stayed at around 5% accuracy

Best approach

We found out the best approach was the simplest one:

- Use just pre trained frozen model to compute embeddings of images
- Add a classification head
- Train for usual multi-class classification task

BioClip was the one that performed better reaching an accuracy of around 20.5%



We also tried adding a small network as the head for classification but we did not notice any improvement

Table of Contents

- Introduction & Paper
- The Project pt. 1
- Background
- The Project pt. 2
- Conclusion

Future Work

- Using more advanced data augmentation techniques like cGANs
- Trying new few-shot techniques like MAML or Siamese Networks
- Taking advantage of the provided metadata
- Leverage image captions with a multi-modal model