MetaFormer : A Unified Meta Framework for Fine-Grained Recognition
arXiv 2022.03.05

Jian Kim

Introduction

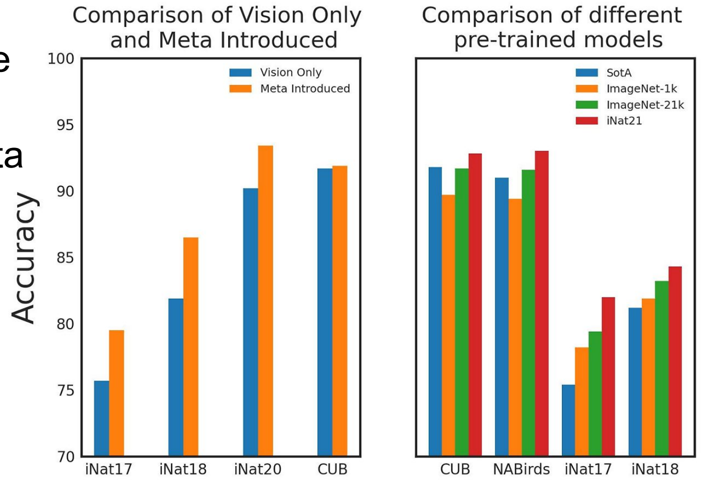
MetaFormer

Fine Grained Visual Classification (FGVC) :distinguishing between subtle differences within the same class
→ The challenge arises due to small inter-class variations and large intra-class variations

Existing methods primarily rely on specific information

→ limitations in distinguishing objects are not universally applicable

Proposal for a Model Structure using Transformer with Image + Meta Information for FGIC



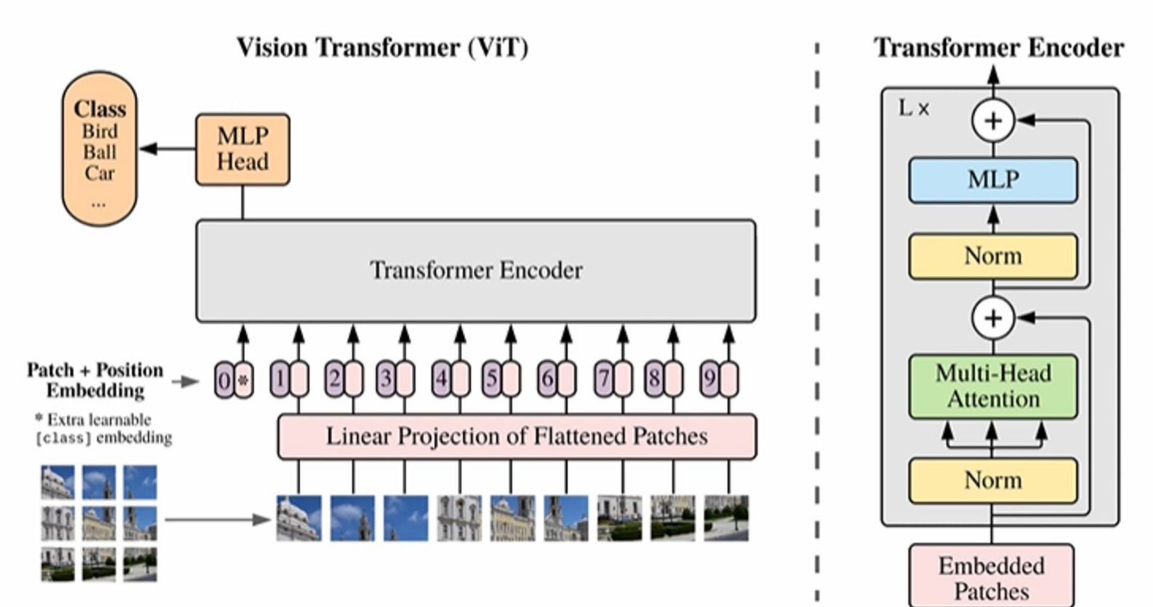
Background

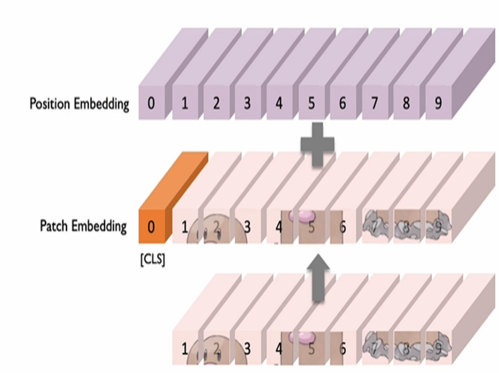
Vision Transformer

Divide the image into patches and use them like tokens in NLP.

linear projection for each patch

Add a CLS embedding, as in BERT, and combine it with position embeddings to use as the input for the encoder





Background

CoAtNet : Convolution and Attention Network

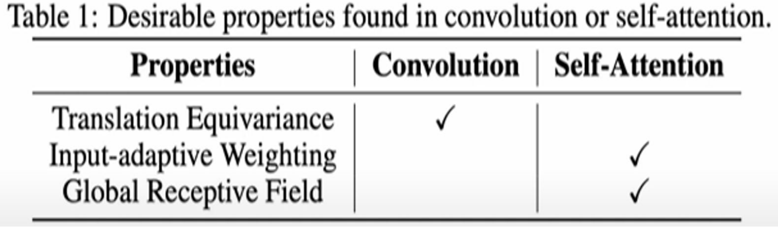


Table 1: Desirable properties found in convolution or self-attention.

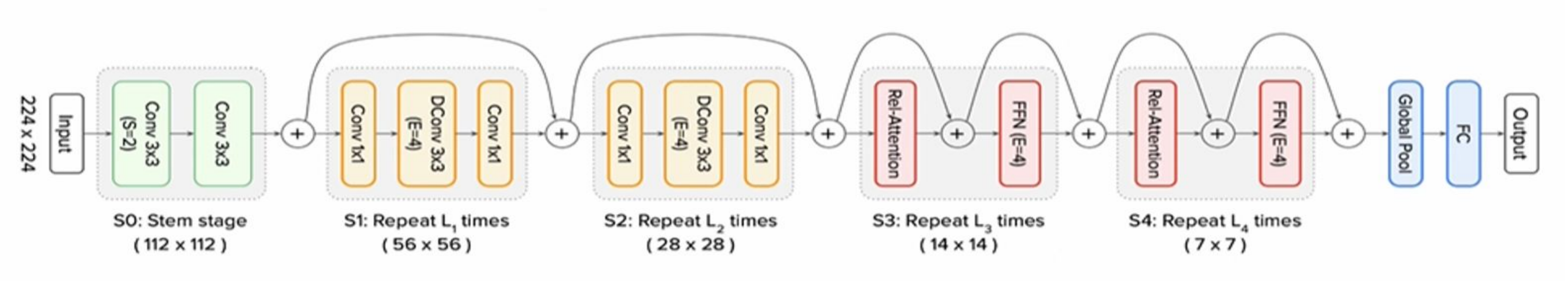
Properties	Convolution	Self-Attention
Translation Equivariance	✓	
Input-adaptive Weighting		✓
Global Receptive Field		✓

Convolution Layers:

- Translation Equivariance :ecognize patterns regardless of their position in the image
- Efficient Local Feature Extraction

Self-Attention Mechanism:

- Input-adaptive Weighting:applies different weights to each input
- Global Receptive Field :considers all parts of the image simultaneously



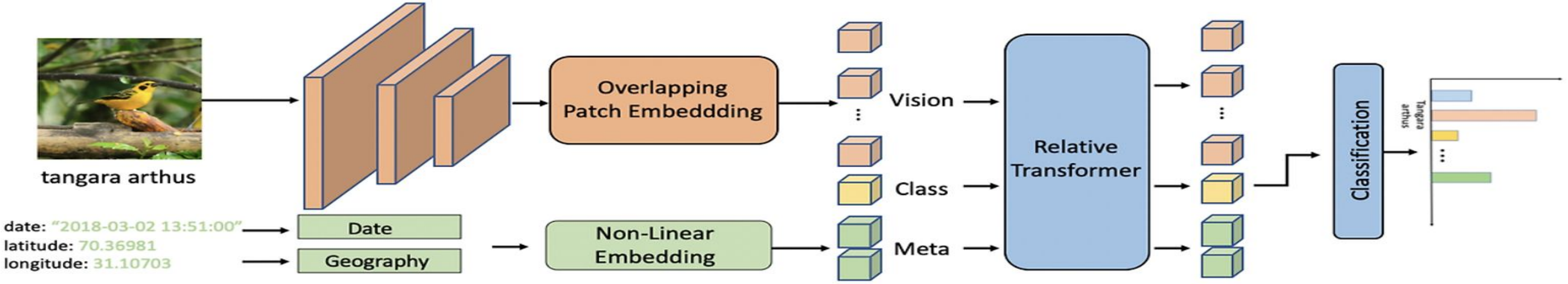
Method

Hybrid Framework

Joint Learning of Image and Attribution (data, geography, text)

Convolution : encode vision information
Transformer layer: fuse vision and meta information

- The image undergoes a convolution stage to generate vision tokens, reflecting semantic information.
- The attribution goes through a non-linear embedding process to generate meta tokens as special tokens.

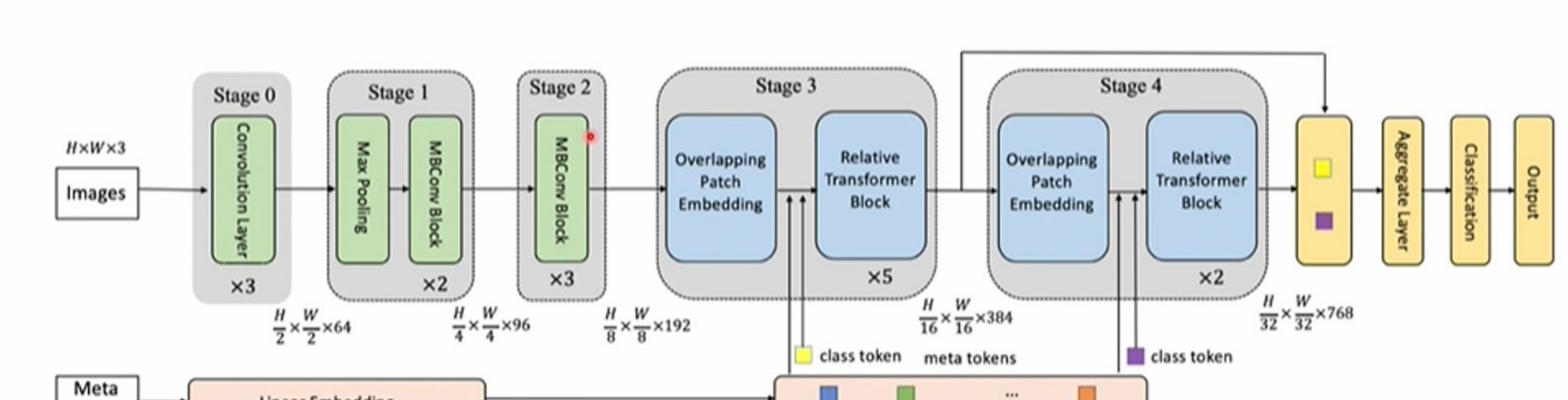


MetaFormer
Model Architecture

The structure consists of four stages excluding Stage 0 (stem stage): C-C-T-T

Relative Transformer Layer: vision token, Meta token and Class token are used for information fusion
Overlapping Patch Embedding: to tokenize the feature map and implement downsampling to reduce computational consumption

aggregate layer : In stages 3 and 4, different class tokens are used, and the final two class token(applies 1D convolution to the concatenated)

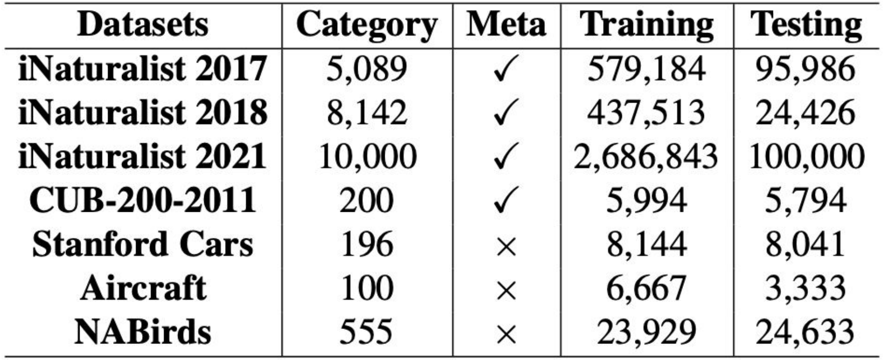

H×W×3
Images
Stage 0
Convolution Layer
×3
Stage 1
Max Pooling
MBConv Block
×2
Stage 2
MBConv Block
×3
Stage 3
Overlapping Patch Embedding
Relative Transformer Block
×5
Stage 4
Overlapping Patch Embedding
Relative Transformer Block
×2
Aggregate Layer
Classification
Output
$\frac{H}{2}\times\frac{W}{2}\times 64$
$\frac{H}{4}\times\frac{W}{4}\times 96$
$\frac{H}{8}\times\frac{W}{8}\times 192$
$\frac{H}{16}\times\frac{W}{16}\times 384$
$\frac{H}{32}\times\frac{W}{32}\times 768$
class token
meta tokens
class token
Meta
...

MetaFormer

Meta Information

- **Date**: Represented as [month, hour] using sine and cosine functions (similar to Transformer Position Embedding).
- **Geography**: Represented as [latitude, longitude] using x, y, z coordinates.
- **Attributes**: Used as a 312-dimensional vector.
- **Text**: Embeddings extracted using BERT.

Only one of the four types of meta information is used for training per benchmark dataset, not all at once.
Date and geography information are concatenated when applied.



Datasets	Category	Meta	Training	Testing
iNaturalist 2017	5,089	✓	579,184	95,986
iNaturalist 2018	8,142	✓	437,513	24,426
iNaturalist 2021	10,000	✓	2,686,843	100,000
CUB-200-2011	200	✓	5,994	5,794
Stanford Cars	196	×	8,144	8,041
Aircraft	100	×	6,667	3,333
NABirds	555	×	23,929	24,633

} Date & Geography

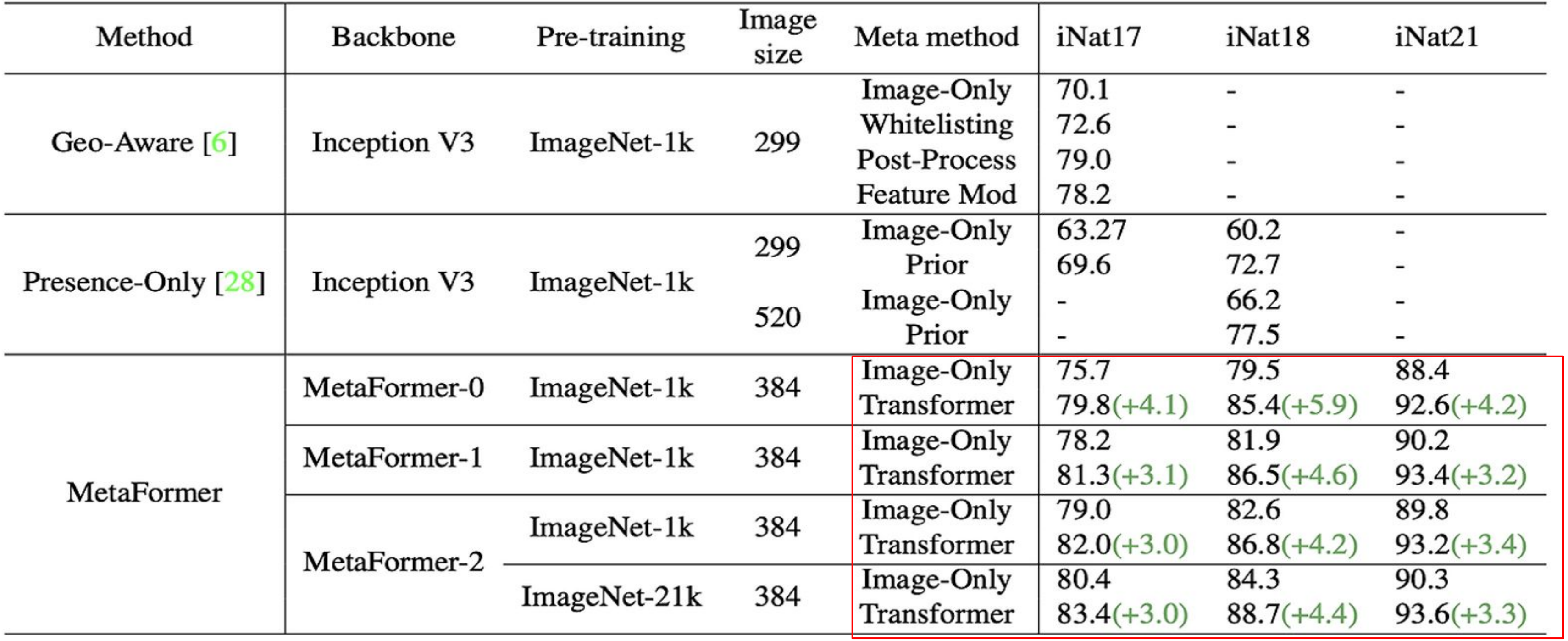
→ Attributes & Text

Experiments

<The Power of Meta Information>

To utilize additional information without separate heads,
Transformer layers were used as the backbone

adding spatio-temporal information, improvements of typically 3-6%

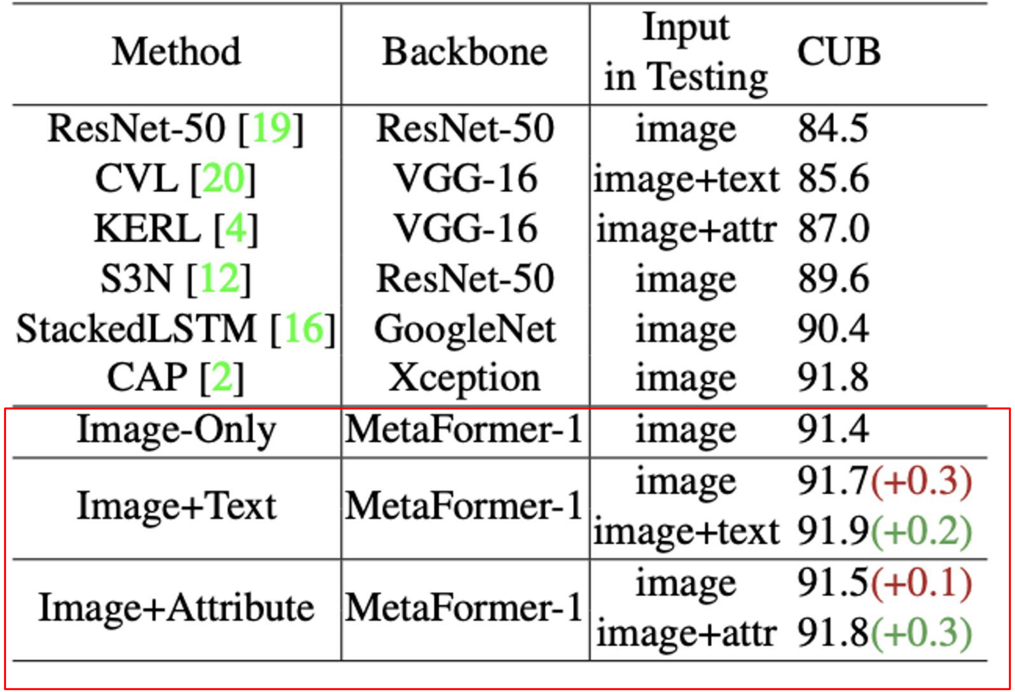


Method	Backbone	Pre-training	Image size	Meta method	iNat17	iNat18	iNat21
Geo-Aware [6]	Inception V3	ImageNet-1k	299	Image-Only	70.1	-	-
				Whitelisting	72.6	-	-
				Post-Process	79.0	-	-
				Feature Mod	78.2	-	-
Presence-Only [28]	Inception V3	ImageNet-1k	299	Image-Only	63.27	60.2	-
				Prior	69.6	72.7	-
			520	Image-Only	-	66.2	-
				Prior	-	77.5	-
MetaFormer	MetaFormer-0	ImageNet-1k	384	Image-Only	75.7	79.5	88.4
				Transformer	79.8(+4.1)	85.4(+5.9)	92.6(+4.2)
	MetaFormer-1	ImageNet-1k	384	Image-Only	78.2	81.9	90.2
				Transformer	81.3(+3.1)	86.5(+4.6)	93.4(+3.2)
	MetaFormer-2	ImageNet-1k	384	Image-Only	79.0	82.6	89.8
				Transformer	82.0(+3.0)	86.8(+4.2)	93.2(+3.4)
		ImageNet-21k	384	Image-Only	80.4	84.3	90.3
				Transformer	83.4(+3.0)	88.7(+4.4)	93.6(+3.3)

Experiments

incorporate various types of additional information

the proposed method effectively utilizes meta information for fine-grained recognition



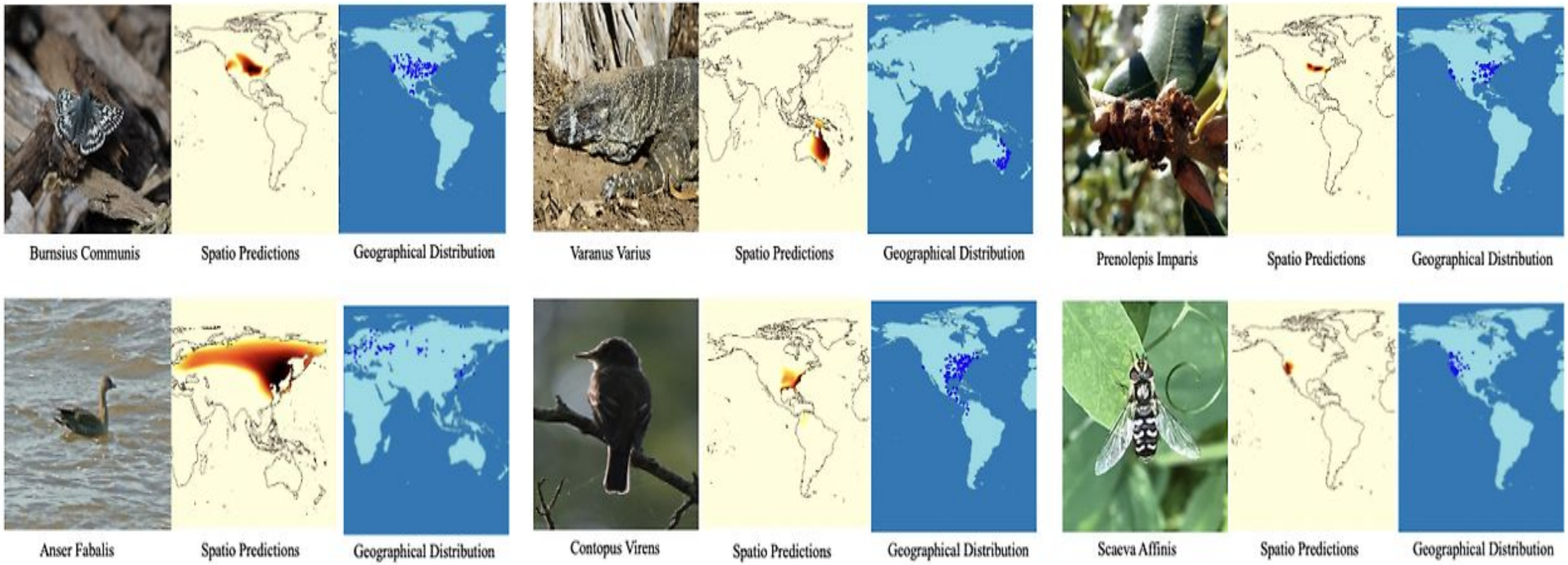
Method	Backbone	Input in Testing	CUB
ResNet-50 [19]	ResNet-50	image	84.5
CVL [20]	VGG-16	image+text	85.6
KERL [4]	VGG-16	image+attr	87.0
S3N [12]	ResNet-50	image	89.6
StackedLSTM [16]	GoogleNet	image	90.4
CAP [2]	Xception	image	91.8
Image-Only	MetaFormer-1	image	91.4
Image+Text	MetaFormer-1	image	91.7(+0.3)
		image+text	91.9(+0.2)
Image+Attribute	MetaFormer-1	image	91.5(+0.1)
		image+attr	91.8(+0.3)

<The Visualization of Meta Information>

Spatial Prediction of Various Species in iNaturalist 2021

The displayed points represent the current geographical distribution of the species.

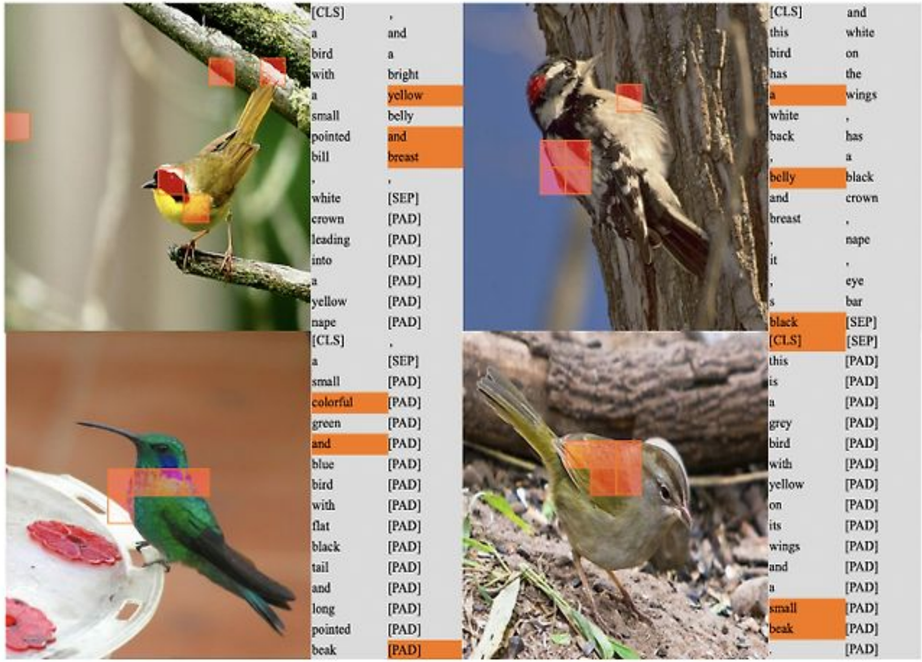
use this geographical distribution as a prior to assist in fine-grained classification



<The Visualization of Meta Information>

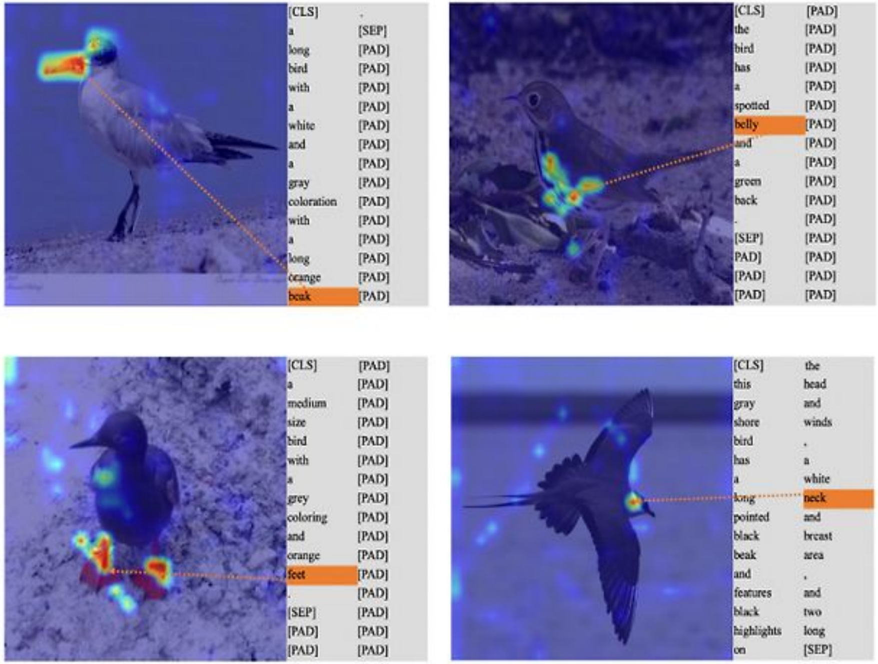
the top-k similarities between word tokens and class tokens

The class token is ultimately used to predict the species category


[CLS] a bird with a small pointed bill , white crown leading into a yellow nape
, and a bright yellow belly and breast . [SEP] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]
[CLS] this bird has a white back , belly and breast . it 's black
and white on the wings , has a black crown , nape , eye bar [SEP]
[CLS] a small colorful green and blue bird with flat black tail and long pointed beak
. [SEP] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]
[CLS] this is a grey bird with yellow on its wings and a small beak .
[SEP] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]

the image attention map corresponding to the word tokens

The words representing the attributes of the species have high similarity with the corresponding image tokens


[CLS] a long bird with a white and a gray coloration with a long orange beak
. [SEP] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]
[CLS] the bird has a spotted belly and a green back . [SEP] PAD] [PAD] [PAD]
[PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]
[CLS] a medium size bird with a grey coloring and orange feet . [SEP] [PAD] [PAD]
[PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]
[CLS] this gray shore bird has a long pointed black beak and features black highlights on
the head and winds , a white neck and breast area , and two long [SEP]

Conclusion

- Propose a model structure that uses image + meta information for fine-grained visual classification (FGVC).
- Conduct experiments to verify the impact of meta information.
- Achieve state-of-the-art (SOTA) performance on various FGVC benchmarks.
  → Meta information is considered essential for fine-grained recognition tasks, and MetaFormer provides a method to utilize various additional information.