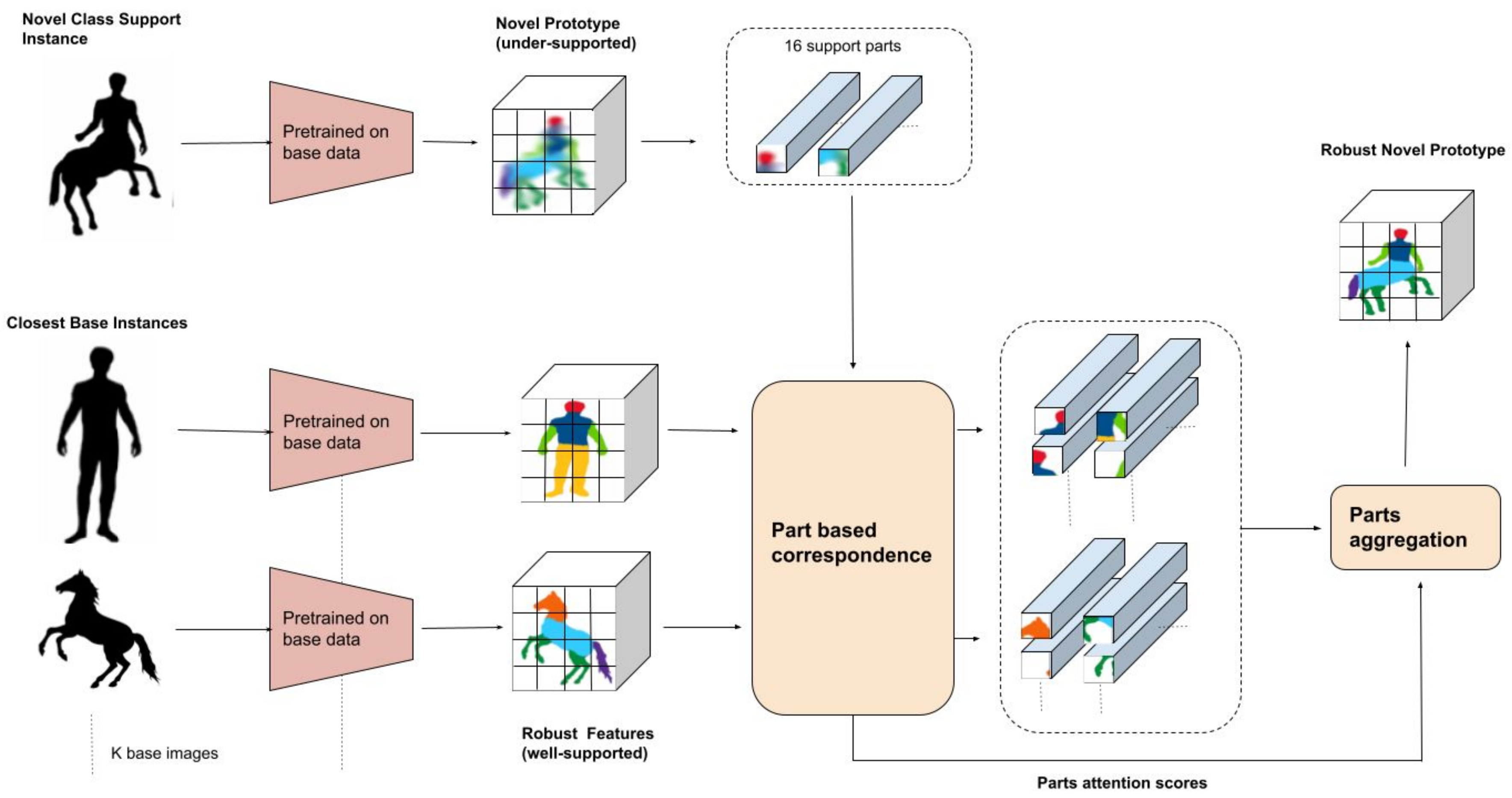
Base Transformers: Attention over base data-points for One Shot Learning

Mayug Maniparambil, Kevin McGuinness, Noel E. O'Connor
Dublin City University, Dublin, Ireland
contact: mayugmaniparambil@gmail.com

BMVC 2022
The 33rd British Machine Vision Conference
21st - 24th November 2022, London, UK

Motivation

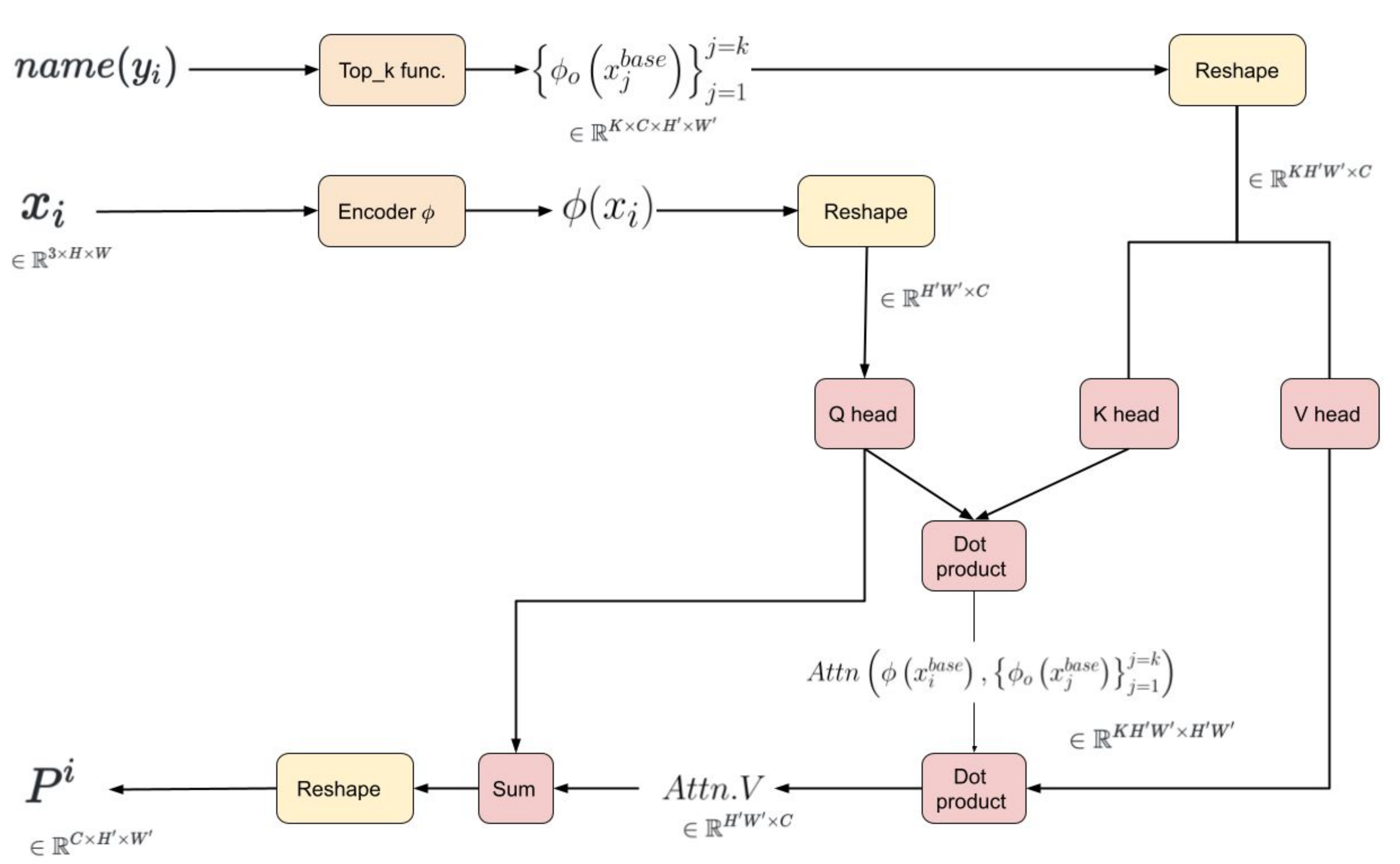
- Current few shot approaches make use of base dataset with many labelled examples per class to train an encoder to get novel class representations.
- Encoders trained on base data provide poor quality test support representations due to distribution shift between base and novel classes.
- BaseTransformers (BT) attends to the most relevant parts of the well-trained base dataset feature space to improve novel class representations.



- In above example, undersupported prototype of novel class centaur can be constructed by taking the head, torso of a human and the body and legs of horse base classes which are individually well supported in the feature space of a base-trained encoder.

Method

- BT uses cross attention between 2d feature space of support instance and closest base instances.
- Closest base instances are uniformly sampled from the closest base classes queried using semantic similarity between the support class label and base class labels.



Querying base classes using semantic similarity

- For mini-ImageNet, LCH similarity on wordnet graph is used.
- For tiered-ImageNet, similarity between BERT embeddings of class and hypernym descriptions is used.
- For CUB, we use cosine similarity between already available category level attributes

Results

- Tables below presents evaluations on mini-ImageNet, tiered-ImageNet and CUB datasets. Tiered-ImageNet results reported for Resnet-12 only.
- Evaluation is over 10,000 randomly sampled test episodes. Best results in bold.

Mini-ImageNet

Setups	1-shot		5-shot	
Backbone	Conv4-64	Res12	Conv4-64	Res12
ProtoNets[28]	49.42±0.78	60.37±0.83	68.20±0.66	78.02±0.57
SimpleShot[34]	49.69±0.19	62.85±0.20	66.92±0.17	80.02±0.14
CAN[11]	-	63.85±0.48	-	79.44±0.34
FEAT[40]	55.15±0.20	66.78±0.20	71.61±0.16	82.05±0.14
DeepEMD[42]	-	65.91±0.82		82.41±0.56
IEPT[43]	56.26±0.45	67.05±0.44	**73.91**±0.34	82.90±0.30
MELR[7]	55.35±0.43	67.40±0.43	72.27±0.35	83.40±0.28
InfoPatch[17]	-	67.67±0.45	-	82.44±0.31
DMF[37]	-	67.76±0.46	-	82.71±0.31
META-QDA[44]	56.41±0.80	65.12±0.66	72.64±0.62	80.98±0.75
PAL[18]	-	69.37±0.64	-	**84.40**±0.44
BaseTransformer	**59.37**±0.19	**70.88**±0.17	73.40±0.18	82.37±0.19

Tiered-ImageNet

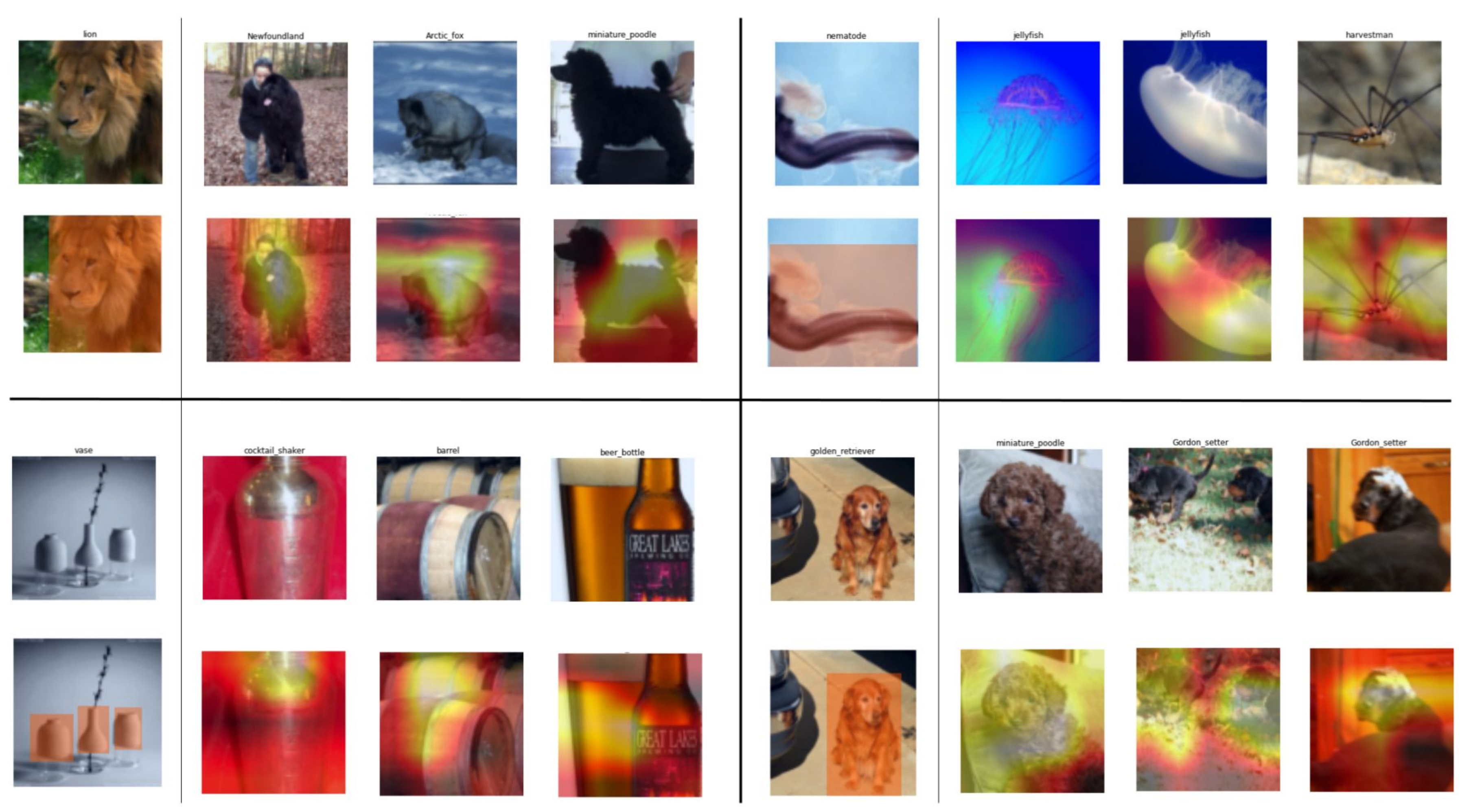
Setups	1-shot	5-shot
ProtoNets[28]	65.65	83.40
SimpleShot[34]	69.75	85.31
FEAT[40]	70.80	84.79
CAN[11]	69.89	84.23
DeepEMD[42]	71.16	86.03
IEPT[43]	72.24	86.73
MELR[7]	72.14	**87.01**
InfoPatch[17]	71.51	85.44
DMF[37]	71.89	85.96
META-QDA[44]	69.97	85.51
PAL[18]	72.25	86.95
BaseTransformer	**72.46**	84.96

CUB

Setups	1-shot		5-shot	
Backbone	Conv4-64	Res12	Conv4-64	Res12
ProtoNets[28]	64.42	-	81.82	-
FEAT[40]	68.87	-	82.90	-
DeepEMD[42]	-	75.65	-	88.69
IEPT[43]	69.97	-	84.33	-
MELR[7]	70.26	-	**85.01**	-
BaseTransformer	**72.15**	**82.27**	82.12	**90.64**

Attention maps visualized over base data

- Attention maps learnt by the BT are visualized below.
- For each support image(left), BT has has learnt to attend to semantically similar regions of base instances.
- In bottom-right quadrant, for golden retriever, BT attends to two instances of gordon setter without being explicitly trained to identify multiple gordon setters.



This project was conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183.

Paper, weights and code at github.com/mayug/BaseTransformers

HOST INSTITUTIONS
University College Dublin
An Coláiste Ollscoile, Baile Átha Cliath
FUNDED BY:
Science Foundation Ireland For what's next