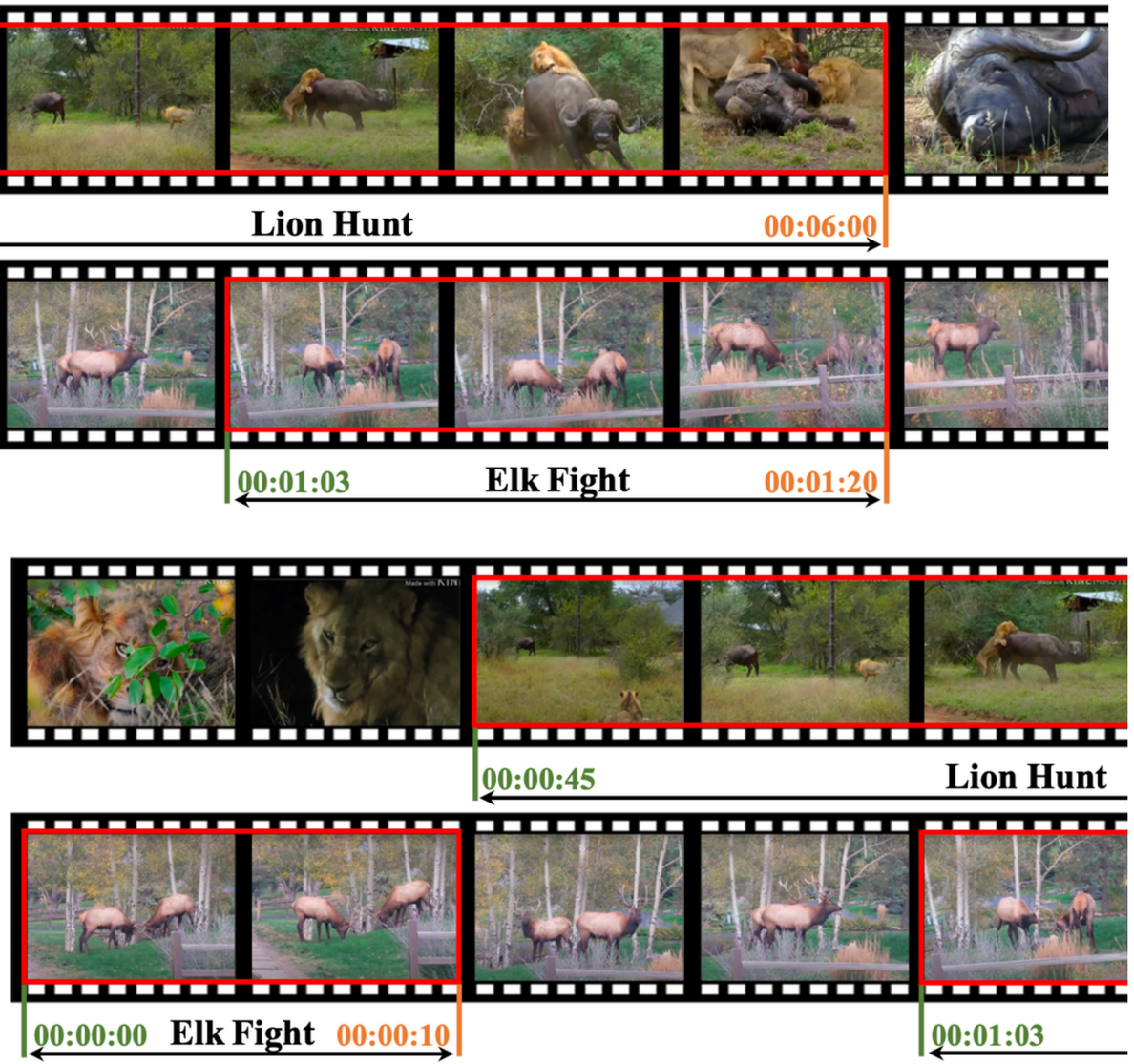
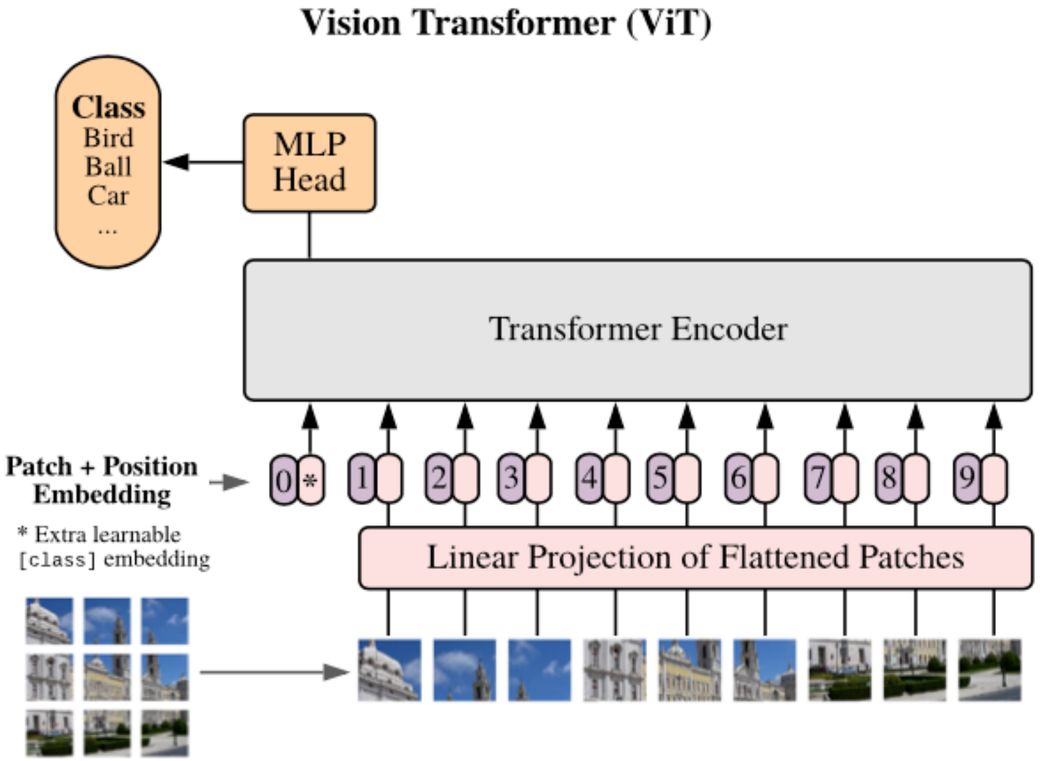


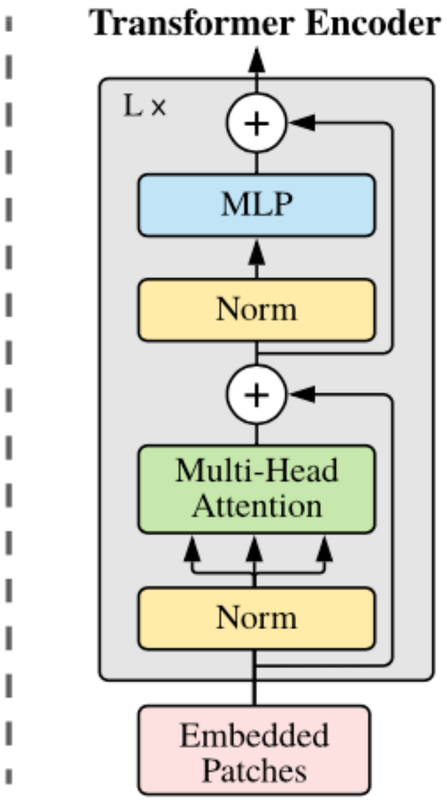
Mammel-Net

- Train an MViT (Multi-scale Vision Transformer) V2 model using the MammalNet dataset.
- The primary goal is to leverage advanced computer vision techniques to develop a robust and accurate model capable of classifying images of mammal species and behaviour.

Pastwork

ViT Architecture





- **Patch embedding.**
- the input image is divided into small, non-overlapping patches.
- converted linear patch arrays to vector

- **Stacked Transformer Encoders**
- The multi-head self-attention mechanism.

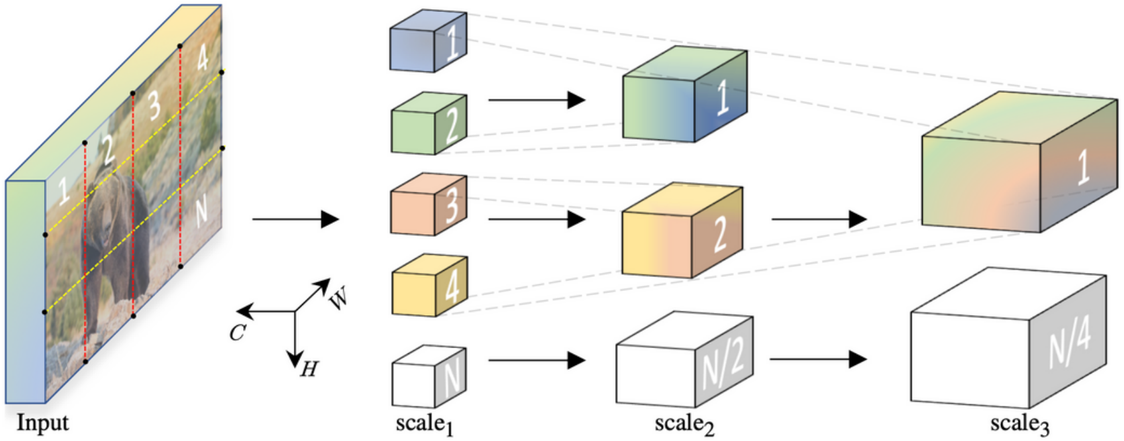
- **Classification Head**
- a neural network layer or a set of layers that take the high-level features extracted by the stacked transformers and map them to specific output classes or categories.

- Patch Embeddings.
- Linear projection layer
- sum patch and positional embedding

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

MViT

Mvit: Multiscale Vision Transformers (link)
Jitendra Malik, Facebook AI Research, UC Berkeley



pros:

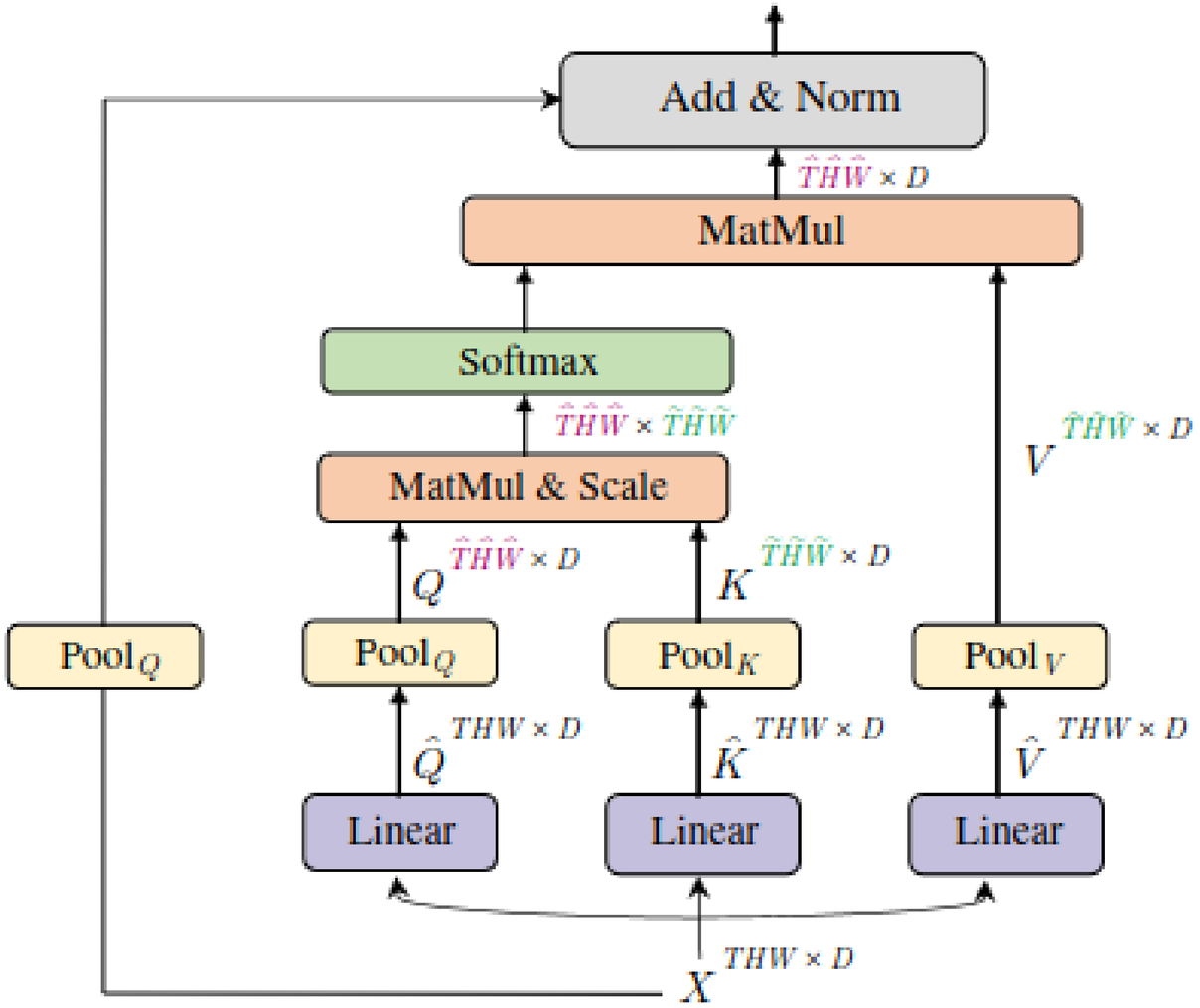
- intention is to connect the seminal idea of multiscale feature hierarchies with the transformer model.
- multiscale transformer arises from the extremely dense nature of visual signals

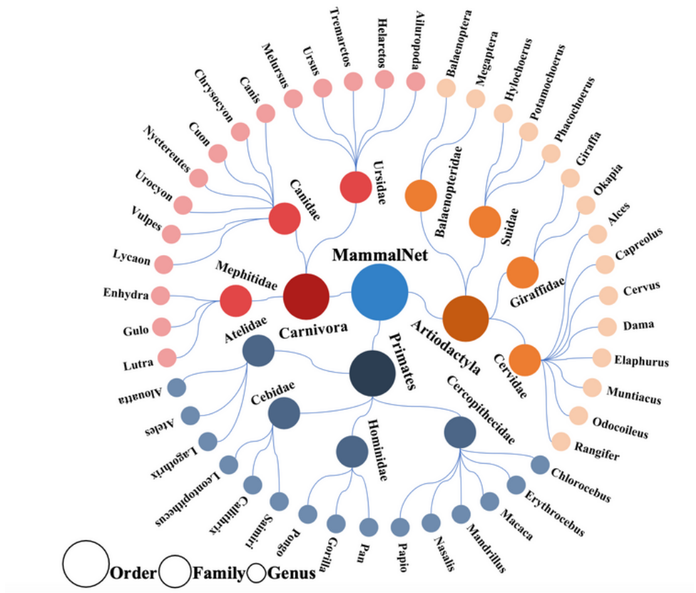
cons:

- Increased computational complexity.
- Limited applications, Lack of pre-training data

- Multiscale Vision Transformers (MViT) outperform previous vision transformers in image classification tasks. MViT shows significant gains over single-scale vision transformers for image recognition. It achieves better performance without the need for large-scale external pre-training on datasets like ImageNet-21K. MViT also outperforms prior work on vision transformers in terms of accuracy.
- It offers an architectural advantage by hierarchically expanding feature complexity while reducing visual resolution. Overall, MViT demonstrates superior performance compared to previous vision transformers in image classification tasks.

MVIT Architecture





Datasets	Dataset Properties								Tasks		
	Publicly Available?	Taxonomy-guided Animal Annotation?	No. of Videos	No. of Actions	No. of Behaviors	No. of Animal Categories	No. of Mammal Categories	Total Duration	Animal Classification	Action/Behavior Recognition	Action/Behavior Detection
Wild Felines [20]	×	×	2,700	3	-	3	3	-	✓	✓	×
Wildlife Actions [29]	×	×	10,600	7	-	32	11	-	✓	✓	×
Animal Kingdom [36]	✓	×	4,301	140	-	850	-	50 (h)	×	✓	×
MammalNet (ours)	✓	✓	18,346	-	12	173	173	539 (h)	✓	✓	✓

MammelNet

- MammalNet is a large-scale video benchmark for mammal recognition and behavior understanding.
- The multi-head self-attention mechanism.The goal of MammalNet is to enable the study of animal and behavior recognition, investigate challenging compositional scenarios,.
- The paper presents several experiments that demonstrate the usefulness of MammalNet for studying animal and behavior recognition
- Data size : 148GB + 365GB
- 539 hrs dataset

- The **purpose** of MammalNet is to provide a large-scale video benchmark for recognizing mammals and understanding their behavior.
- It aims to address the limitations of existing animal behavior datasets by curating a diverse and representative dataset that covers a wide range of mammal species and behaviors. MammalNet enables the study of animal and behavior recognition, both separately and jointly.
- Additionally, MammalNet includes behavior detection by localizing when a behavior occurs in a video. It serves as a valuable resource for the computer vision community to develop and evaluate solutions for mammal recognition and behavior understanding.

- Behavior detection in MammalNet refers to the task of localizing when a specific behavior occurs within an untrimmed video.
- It involves identifying the temporal boundaries of the behavior within the video sequence. MammalNet provides annotations that allow for the detection and localization of behaviors, enabling researchers to study and analyze specific behaviors exhibited by mammals.

Disadvantage

- Bias towards captive animals or habituated animals: Many videos in the dataset are shot in zoos, farms, and homes, which may result in different behaviors compared to wild or non-habituated animals.

- Lack of classification according to established biological taxonomies: Existing video datasets do not classify animals based on established biological taxonomies, which limits the usefulness of these datasets for large-scale behavioral studies.

- Limited environmental diversity: Some previous datasets have a small number of videos with insufficient environmental diversity, which can affect the generalization of models trained on these datasets.

Animal Kingdom Dataset

Dataset	Publicly available?	Diverse types of animals						No. of species	Task 1: Video Grounding		Task 2: Action Recognition		Task 3: Pose Estimation	Types of scene									Weather			
		Mammals	Reptiles	Birds	Amphibians	Fishes	Insects		No. of annotated long videos	No. of statements	No. of annotated video clips	No. of annotated action classes	No. of labelled images	Night scene	Low light	Complex background	Mountain	Forest	Grassland	Desert	Ocean	Underwater	Windy	Foggy	Rain	Snow
Broiler Chicken [14]	×	×	×	✓	×	×	×	NA	×	×	NA	6	556	NA	NA	NA	×	×	×	×	×	×	NA	NA	NA	NA
Fish Action [53]	×	×	×	×	×	✓	×	NA	×	×	95	5	×	×	✓	✓	×	×	×	×	✓	✓	×	×	×	×
Salmon Feeding [39]	×	×	×	×	×	✓	×	1	×	×	76	2	×	×	✓	✓	×	×	×	×	✓	✓	×	×	✓	×
Wild Felines [17]	×	✓	×	×	×	×	×	3	×	×	2,700	3	×	✓	✓	✓	×	✓	✓	×	×	×	NA	NA	NA	NA
Pig Tail-biting [38]	×	✓	×	×	×	×	×	1	×	×	4,396	2	×	×	✓	×	×	×	×	×	×	×	×	×	×	×
Wildlife Action [35]	×	✓	✓	✓	✓	✓	✓	106	×	×	10,600	7	×	✓	✓	✓	×	✓	✓	×	✓	✓	NA	NA	NA	NA
Animal Pose [8]	✓	✓	×	×	×	×	×	5	×	×	×	×	4,666	✓	✓	✓	✓	✓	✓	×	×	×	×	✓	×	✓
Horse-30 [40]	✓	✓	×	×	×	×	×	3	×	×	×	×	8,144	×	×	✓	×	×	✓	×	×	×	×	×	×	×
AP-10K [79]	✓	✓	×	×	×	×	×	54	×	×	×	×	10,015	✓	✓	✓	✓	✓	✓	×	✓	✓	×	✓	×	✓
Macaque Pose [30]	✓	✓	×	×	×	×	×	NA	×	×	×	×	13,083	✓	✓	✓	✓	✓	×	×	×	×	×	✓	×	✓
Dogs [3]	✓	✓	×	×	×	×	×	1	×	×	13	4	2,200	×	✓	×	×	×	×	×	×	×	×	×	×	×
Animal Kingdom (Ours)	✓	✓	✓	✓	✓	✓	✓	**850**	**4,301 (50h)**	**18,744**	**30,100 (50h)**	**140**	**33,099**	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

MammalNet

Datasets	Dataset Properties								Tasks		
	Publicly Available?	Taxonomy-guided Animal Annotation?	No. of Videos	No. of Actions	No. of Behaviors	No. of Animal Categories	No. of Mammal Categories	Total Duration	Animal Classification	Action/Behavior Recognition	Action/Behavior Detection
Wild Felines [20]	×	×	2,700	3	-	3	3	-	✓	✓	×
Wildlife Actions [29]	×	×	10,600	7	-	32	11	-	✓	✓	×
Animal Kingdom [36]	✓	×	4,301	140	-	850	-	50 (h)	×	✓	×
MammalNet (ours)	✓	✓	18,346	-	12	173	173	539 (h)	✓	✓	✓

RESULTS from the Paper

Animal and Behavior

Classification

Animal Kingdom Dataset

Table 2. Results of action recognition

Method	mAP			
	overall	head	middle	tail
Baseline (Cross Entropy Loss)				
I3D [10]	16.48	46.39	20.68	12.28
SlowFast [16]	20.46	54.52	27.68	15.07
X3D [15]	25.25	60.33	36.19	18.83
Focal Loss [37]				
I3D [10]	26.49	64.72	40.18	19.07
SlowFast [16]	24.74	60.72	34.59	18.51
X3D [15]	28.85	64.44	39.72	22.41
LDAM-DRW [9]				
I3D [10]	22.40	53.26	27.73	17.82
SlowFast [16]	22.65	50.02	29.23	17.61
X3D [15]	30.54	62.46	39.48	24.96
EQL [66]				
I3D [10]	24.85	60.63	35.36	18.47
SlowFast [16]	24.41	59.70	34.99	18.07
X3D [15]	30.55	63.33	38.62	25.09

MammalNet

Baselines	Animal	Behavior	Joint
SlowFast [19]	35.4	34.2	17.4
C3D [42]	35.0	33.5	17.1
I3D [11]	35.2	34.3	17.9
MViT V2 [30]	**35.6**	**36.8**	**18.0**
SlowFast*	43.0	39.4	22.8
C3D*	44.4	40.3	24.6
I3D*	43.4	41.2	24.0
MViT V2*	**52.6**	**46.6**	**30.6**

RESULTS

Animal Behavior Detection

- This model is trained on Untrimmed video.
- Aiming to detect the time period of action for trimming the videos

Baselines	mAP					
	0.50	0.60	0.70	0.80	0.90	Avg.
CoLA [52]	26.02	22.70	18.98	13.46	3.05	15.81
TAGS [35]	23.09	20.97	19.09	16.98	**12.56**	17.63
ActionFormer [53]	**28.48**	**26.14**	**23.17**	**18.69**	10.48	**20.07**

- **Our Results:**
- We trained the ActionFormer model on half of the untrimmed MammalNet dataset with tIoU of 0.5 and attained a result of **19.16**

- AP(tIoU): Precision(tIoU) * Recall(tIoU)

- Temporal Intersection over Union (tIoU) is like a measure of how much overlap is needed for a prediction to be considered correct. It ranges from 0.5 (50% overlap) to 0.9 (90% overlap).

```
Start testing model LocPointTransformer ...
Test: [00010/00212]     Time 1.89 (1.89)
Test: [00020/00212]     Time 0.12 (1.00)
Test: [00030/00212]     Time 0.13 (0.71)
Test: [00040/00212]     Time 0.15 (0.57)
Test: [00050/00212]     Time 0.15 (0.49)
Test: [00060/00212]     Time 0.11 (0.42)
Test: [00070/00212]     Time 0.12 (0.38)
Test: [00080/00212]     Time 0.15 (0.35)
Test: [00090/00212]     Time 0.13 (0.33)
Test: [00100/00212]     Time 0.11 (0.31)
Test: [00110/00212]     Time 0.17 (0.29)
Test: [00120/00212]     Time 0.14 (0.28)
Test: [00130/00212]     Time 0.14 (0.27)
Test: [00140/00212]     Time 0.15 (0.26)
Test: [00150/00212]     Time 0.12 (0.25)
Test: [00160/00212]     Time 0.17 (0.25)
Test: [00170/00212]     Time 0.21 (0.25)
Test: [00180/00212]     Time 0.12 (0.24)
Test: [00190/00212]     Time 0.19 (0.24)
Test: [00200/00212]     Time 0.16 (0.23)
Test: [00210/00212]     Time 0.13 (0.23)
[RESULTS] Action detection results on thumos14.

|tIoU = 0.30: mAP = 82.20 (%) Recall@1x = 84.35 (%) Recall@5x = 96.36 (%)
|tIoU = 0.40: mAP = 78.05 (%) Recall@1x = 80.35 (%) Recall@5x = 95.11 (%)
|tIoU = 0.50: mAP = 71.29 (%) Recall@1x = 74.67 (%) Recall@5x = 92.12 (%)
|tIoU = 0.60: mAP = 59.20 (%) Recall@1x = 64.83 (%) Recall@5x = 83.89 (%)
|tIoU = 0.70: mAP = 43.53 (%) Recall@1x = 52.99 (%) Recall@5x = 71.25 (%)
Average mAP: 66.85 (%)
All done! Total time: 61.04 sec
```

Refrences

MammalNet: A Large-scale Video Benchmark for Mammal Recognition and Behavior Understanding

Actionformer: Localizing moments of actions with transformers.

Animal Kingdom: A Large and Diverse Dataset for Animal Behavior Understanding

MVit_v2: Improved multiscale vision transformers for classification and detection (link)

- Jitendra Malik, Facebook AI Research, UC Berkeley

Mvit: Multiscale Vision Transformers (link)

- Jitendra Malik, Facebook AI Research, UC Berkeley

Vit: Action Recognition? A New Model and the Kinetics Dataset (link)

- Joa˜o Carreira, Andrew Zisserman, DeepMind, University of Oxford

https://github.com/Vision-CAIR/MammalNet

https://github.com/google-research/vision_transformer