Paper Notes:

Link: <https://arxiv.org/pdf/2203.07436v3.pdf>

From: SuperAnimal pretrained pose estimation models for behavioral analysis

* reliable inference of poses currently requires domain knowledge and manual labeling

effort to build supervised models

* models: 2,3,7,8
* Open-Source pose estimation software, such as
  + Others (11,14,7)
  + DeepLabCut (9,10)
* DeeplabCut offers flexibility to train customized pose models of various animals in diverse settings, requires around 100-800 human labeled images to train typical lab animal pose estimator that matches human level accuracy (9,10), due to its transfer learning abilities (9,16)
* Panoptic paradigm -> adaptation across many species, environment, and video sizes
* SuperAnimal combines diverse datasets into two broad unified pose models that covers over 45 species of mammals with 27-39 keypoints.

Keywords:

* keypoint matching algorithm to automatically align out-of-distribution datasets with our models. Then, at inference time, to minimize domain shifts, we developed a spatial-pyramid search method to account for changes in animal size (or use a top-down detector). We also developed a rapid, unsupervised video-adaptation method that uses pseudo-labeling to minimize temporal jitter in videos and allows users to fine-tune videos without any data labeling.
* We developed models based on state-of-the-art convolutional neural networks (CNNs), such as HRNet (18) and DLCRNet (10), and introduce AnimalTokenPose that uses transformers (19–21). We show that the resulting models have excellent zero-shot performance (i.e., with no additional training, tested on new data), and our approach outperforms ImageNet-pretraining on five benchmarks. If users then want to use these new weights for fine-tuning, we show they are 10X more data efficient, and our video adaptation method allows for smooth, refined videos that can be used in behavioral analysis pipelines.
* Figure 1. The DeepLabCut Model Zoo, the SuperAnimal method, and SuperAnimal-TopViewMouse model performance. a: The website can collect data shared by the research community; SuperAnimal models are trained, and can be used for inference on novel images and videos with or without further training. b: The panoptic animal pose estimation approach unifies the vocabulary of pose data across labs, such that each individual dataset is a subset of a super-set keypoint space, independently of its naming. c: For canonical, task-agnostic transfer learning, the encoder learns universal visual features from ImageNet, and a randomly initialized decoder is used to learn the pose from the downstream dataset. For task-aware fine-tuning, both encoder and decoder learn task-related visual-pose features in the pre-training datasets and the decoder is fine-tuned to update pose priors in downstream datasets. Crucially, the network has pose-estimation specific weights. d: Memory-replay combines the strengths of SuperAnimal models’ zero-shot inference, data combination strategy, and leveraging labeled data for fine-tuning (if needed). It achieves better data efficiency. e: Data efficiency of baseline (ImageNet) and various SuperAnimal fine-tuning methods on the DLC-Openfield OOD dataset. Grey shadow represents minimum, maximum and blue dash is the mean for zero-shot performance across three shuffles. Large, connected dots represent mean results across three shuffles and smaller dots represent results for individual shuffles. f: Using memory replay avoids catastrophic forgetting; here all keypoints are predicted. g: Top: SuperAnimal-TopViewMouse qualitative results on the within distribution test images (IID). They were randomly selected based on visibility of the keypoints within the figure (but not on performance). Full keypoint color and mapping is available in Extended Data Figure S1). h: Top: Transformers also perform well on OOD tasks. Bottom: visualization of model performance on OOD images using DLCRNet.