DogPose – dog pose classification

I.M. Shigabeev[I], James Rodriguez[II], N.Yu. Chernykh[III]

[I] National University of Science and Technology "MISIS", Moscow, Russia
[II] Arizona State University
[III] Pirogov Russian National Research Medical University (RNRMU), Moscow, Russia

Аннотация. In this work, we present a dataset for a pose classification of dogs as well as a sample pipeline for employing this dataset into an AI-powered application that tracks dog activity throughout the day, giving its user information on whether his dogs sleep all day or it stays active even while the dog owner is not home. This application is essential for dog owners to spot the trends of increasing dog passivity.

Ключевые слова: *Computer vision, Pose Estimation, Image Classification, Internet of things, Semi-supervised dataset generation, Data Collection, Artificial Intelligence, Pattern Recognition.*
DOI: 10.14357/20790279210306

Introduction

Dog owners often do not spend as much time with their pets as their pets would have wanted. People in pursuit of their own needs would go to work, would go out to see their friends, spending time with their pets only to feed them and to go for a walk, only occasionally playing with them. While locked in a house, the dogs would become less active, more depressed, and sometimes require professional help. That is why it is essential to track pets' activity during the day to see if they are becoming too lazy and sleep too much.

This problem exists not only for dogs, and there is an emerging market for health monitoring for household animals. A growing number of startups make products that track animal activity. CattleCare - is a company that tracks the posture and milking routine of cows. This product helps to find any health anomalies on a scale of big farms early on. Multiple works in this field cover how to track the pose of different animals, such as DeepLabCut [1]. Tools like DeepLabCut do not require a vast supervised dataset to allow pose estimation on a video, but it still requires some supervision in general. Authors report that they have their proprietary collection of labeled poses of cats, dogs, horses, and numerous other animals.

Additionally, there is existing work on tracking dog activity from dog's POV [2] and classification of a pose of a robotic dog [3]. They all show that tracking of animal activity is a topic growing in popularity.

In this work, to track the dog's activity, we use the classifier of the dog's pose to find if the dog sits, stands, or lies on the floor. Placed on a device with a camera, a solution like the one presented in this paper can track and collect statistics without human interaction on whether a dog has spent all day sleeping or was active and often changed its pose and location.

In this work, to track dog's activity we use classifier of dog's pose to find if the dog sits, or stands, or lies on the floor. Placed on a device with a camera, a solution like the one presented in this paper can track and collect statistics without human interaction on whether a dog has spent all day sleeping or was it active and often changed its pose and location.

1. Dog pose classification

To solve the stated problem, our solution would require to have the following problems solved. They are:

1. Determining if the dog is present on a video frame (detection)
2. Tracking the dog location in the frame (localization)
3. Classifying the pose of a dog (classification)

For detecting and localization of the animal position, object detection neural networks would be employed. Object detection is a solved problem, so we will only take the existing neural networks such as single-shot detectors (SSD). The more complicated task would be a pose classification as it would require a new dataset that did not exist before that has to be labeled manually. The dataset can be small as we would only need to fine-tune the head of the object detection neural network, which is already good at computer vi-

* Работа выполнена при частичной финансовой поддержке РФФИ в рамках научных проектов № 18-29-03070 и № 18-29-03085.

sion tasks. So, it would require only some examples of dogs in different poses. This neural network would get an image cropped after object detection only to contain a dog as an input.

Before 2020 the most popular architecture for solving computer vision problems was ResNet [4] due to its impressive success in most scenarios. However, recent advancements in Attention mechanisms and the Transformer [5] invention have discovered a whole new class of visual architectures. These new architectures can perceive the scene differently compared to convolutional nets and demonstrate different properties. CLIP by OpenAI [6] has shown how to connect the visual and textual domain of a scene, allowing not only to classify the scene but to see which prompt is more likely to describe the picture with words like "a dog" or "a computer," obsoleting the whole idea of exhaustive image classification. These models do not require fine-tuning and are called "zero-shot learners." For architectures like CLIP, this dataset is still required for quality assurance.

2. Dataset collection

Dog pose classification is a challenging task for computer vision due to the inherent nature of how the dataset would be collected and how computers would classify images. No matter how collected, the dataset would show significant inner-class variance: the dogs would be of different breeds and sizes, collected in different conditions and backgrounds. At the same time, very similar images of the same dog can have a different class which is a bit tricky considering how computers solve visual tasks (Neural networks are designed to distinguish shapes). The dataset needs to consist of images of similar dogs in different poses so that the classifier would not consider the difference in dog's color or size, or background as important features for pose classification. That is why it is crucial to filter out as many variables from the dataset as possible. There are some requirements to images made during the dataset collection process to make this classification even possible for computers.

As a base for the dataset, the subset of Dogs from Open Image Dataset [7] was used. In general, the amount of poses that an animal can take is vast. A significant share of pictures from OID (around 20%) could not be classified by pose because some poses can be too rare or unique (e.g., a dog standing on its back feet). Although, most of the dogs in this dataset were seen in the following poses:

1. Standing
2. Sitting
3. Laying on a belly
4. Laying on a back
5. Laying on a side
6. On hands of the owner
7. Flying (during a jump or run)
8. Swimming

Only the three most popular poses – sitting, standing, and lying on a belly were kept to keep it simple. Others were discarded as the number of photos was increasingly small for less popular poses.

Open Image Dataset had as many as 20K pictures of dogs, but only around 2K pictures remained after filtering. Only images of $500 * 500$ pixels or larger were kept. Degenerate (as shown in fig 1.) or unusual images (distorted colors, very blurry or noisy, largely occluded, extreme close-ups, depictions) were removed manually. Pictures that contained multiple objects were discarded as well. Eventually, even a dataset as big as 20000 pictures from OID became small, and the neural networks required far more data, so we came up with our data collection technique.

Fig. 1. Hard to classify images of dogs: multiple objects, perspective distortions, rotation, and occlusion

We scrapped open image sources such as Flickr and Facebook. Flickr images are rarely labeled, and even labeled photos are very noisy, so only 10K images from there were downloaded. On the flip side, Facebook groups of dog owners provided more flexibility and far more pictures of dogs taken in the wild and on consumer phone cameras, closer to a target device than photos taken on professional cameras from Flickr.

The pipeline for the collection of the dataset is quite common. At first, we managed to get as many dog photos as possible and then filter them to conform to requirements. Three-stepped process:

1. Scrap a Facebook group and download all pictures from it.
2. Run an object detection over all pictures
   a. If a picture doesn't contain a dog on the photo, it is automatically discarded
   b. If it has two dogs – it is discarded
   c. Check if there is a margin between the edge of the frame and a dog. If there is not – the dog is not captured fully
   d. Run a classifier of dog's pose (the topic of this paper) and use only pictures with high model certainty.
3. Verify pictures that remained after step 2 with human force. Humans also make sure that every dog is on flat ground.

This technique allowed to speed up the data gathering process as humans only needed to look through roughly 10% of all pictures. This way, after scrapping more than 100K pictures, only 4500 of them ended up in a dataset, and only 15000 of them were examined by humans (excluding OID pictures). The final dataset consists of 4500 images of dogs, 1500 images per class.

There are three classes:

1. Dog stands
2. Dog sits (shown on figure 2)
3. Dog lays on the floor

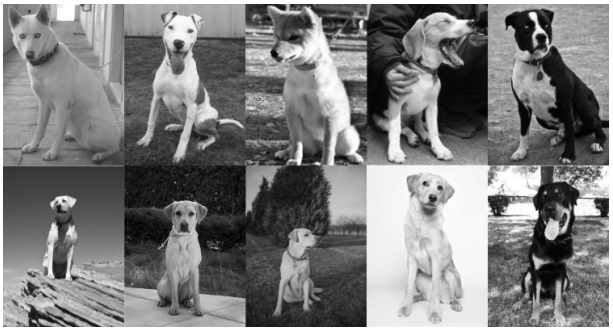

Fig. 2. Example of sitting dogs in the dataset

Performance assessment. We measured performance on our dataset for four neural architectures. Every model in the table except CLIP uses pretrained and frozen convolutional or attentive backbone and 2-3 fully connected layers. There is also a human performance as a reference in Table 1.

Due to limited training data, it was only possible to use pretrained models and measure performance on them.

MobileNet is a small model, with only 10% of the size of a ResNet, but it shows comparable performance on some visual tasks. With an input size of 96, MobileNet demonstrates the best performance overall in the architectures tested.

Visual Transformer [8] from Google also showed good performance despite its small input image size of 72. We pretrained it on the CIFAR-100 dataset.

CLIP required some time to adjust for the task. After picking the correct input prompts, the quality showed up to 96% accuracy in one class. This model has great potential, but even without any adjustments, it shows 76% accuracy. For CLIP, no training was used. As the classification prompts, we used the words "sit," "stand," and "on the floor."

Table 1

Performance of different architectures on the dataset

Architecture	Accuracy
ResNet-50	80%
MobileNetV2	96%
Google's Visual Transformer	87%
OpenAI CLIP	76%
Human	99%

3. The application

As seen in figure 3, the resulting application takes an optical flow (image frames) in real-time from the camera and tracks the dog's position if it is present on a frame. The image is cropped to the bounding box of the dog. A margin of 15 pixels on each side to ensure that the dog cropped as a whole. The information about the dog's position in the frame and its pose is saved into a database. As the camera is stationary, the statistics about location would make sense to the user because it will show that dog walked around the room in which the camera is located.

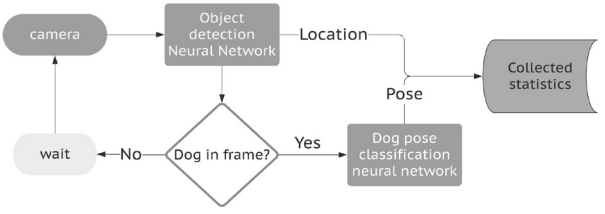

Fig. 3. Flowchart of the application's logic

Because the target device is Raspberry Pi 4, the computing limitations are very tight. The heaviest model is the one for object detection. It is hard to object detection on Raspberry pi at all, but with the tiny and quantized models, it is possible to achieve acceptable quality on up to 5 frames per second (FPS). The common choice is to use MobileNet Single Shot Detector (SSD) trained on COCO dataset.

The classification Neural network uses the same backbone as the one for object detection, so their weights can be shared. MobileNet alone is used for classification can work as fast as 8 FPS on a Raspberry Pi 4. With shared weights between object detection and classification neural networks, it is possible to achieve a 3-4 FPS performance.

From the user perspective, it is not necessary to store data even that often. So, to avoid overheating of Raspberry in the long run (more than 2 hours), it is enough to run the pipeline only once a second. The collected statistics can be stored in a cloud to ease access to this data.

Conclusion

This research is an application that can track the dog's activity throughout the day on an auxiliary device and provide a user with statistics on how long

his dog sleeps and walk daily. The Subproduct of this research is a dataset that provides a challenge for the computer vision neural networks. Our baseline approach for dog pose classification solves the task with 96% accuracy.

This product has intended to be a step into more sophisticated tasks like tracking all dog's limbs and detecting dogs' signals like tail movement.

References

1. *Mathis A., Mamidanna P., Cury K.M.* et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. // Nat Neurosci 21, pp.1281–1289 (2018).
2. *Y. Iwashita, A. Takamine, R. Kurazume and M.S. Ryoo.* First-Person Animal Activity Recognition from Egocentric Videos // International Conference on Pattern Recognition (ICPR), 2014, pp.4310-4315.
3. *Körner M., Denzler J.* JAR-Aibo: A Multi-view Dataset for Evaluation of Model-Free Action Recognition Systems. // New Trends in Image Analysis and Processing – ICIAP 2013, 2013, vol 8158, pp.527-535, https://doi.org/10.1007/978-3-642-41190-8_57.
4. *He, Kaiming & Zhang, Xiangyu & Ren, Shaoqing & Sun, Jian.* Deep Residual Learning for Image Recognition. // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.770-778.
5. *Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan & Kaiser, Lukasz & Polosukhin Illia.* Attention Is All You Need. // Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017, vol.30
6. *Radford, Alec & Kim, Jong & Hallacy, Chris & Ramesh, Aditya & Goh, Gabriel & Agarwal, Sandhini & Sastry, Girish & Askell, Amanda & Mishkin, Pamela & Clark, Jack & Krueger, Gretchen & Sutskever, Ilya.* Learning Transferable Visual Models From Natural Language Supervision // 2021, arXiv:2103.00020.
7. *Kuznetsova, Alina & Rom, Hassan & Alldrin, Neil & Uijlings, Jasper & Krasin, Ivan & Pont-Tuset, Jordi & Kamali, Shahab & Popov, Stefan & Malloci, Matteo & Kolesnikov, Alexander & Duerig, Tom & Ferrari, Vittorio*. The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale // International Journal of Computer Vision, 2020. vol 128.
8. *Dosovitskiy, Alexey & Beyer, Lucas & Kolesnikov, Alexander & Weissenborn, Dirk & Zhai, Xiaohua & Unterthiner, Thomas & Dehghani, Mostafa & Minderer, Matthias & Heigold, Georg & Gelly, Sylvain & Uszkoreit, Jakob & Houlsby, Neil*. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale // International Conference on Learning Representations (ICLR), 2020.

Shigabeev Ilya Maratovich. Research associate of "ID R&D" Inc, Serafimovicha, 2. B.Sc. Graduated from NUST MISiS in 2017. 2 published articles. Topics of interest: machine learning, computer vision, audio and speech processing. E-mail: shigabeev@edu.misis.ru

James Rodriguez. Master's student at Arizona State University. Bachelor of Science in Computer Science at Sonoma State University. Fremont, California, United States. Topics of interest: machine learning, computer vision, server management. E-mail: james.mc.rodriguez@gmail.com

Chernykh Nadezhda Yurevna. Researcher at Academician Yu. E. Veltishchev Research Clinical Institute of Pediatrics, N. I. Pirogov Russian National Research Medical University, Ministry of Health of the Russian Federation, Moscow, Russia. Topics of interest: medicine, cardiology. E-mail: Chernykh-nauka@mail.ru