MoveNet.SinglePose

Model Details

A convolutional neural network model that runs on RGB images and predicts human joint locations of a single person. The model is designed to be run in the browser using Tensorflow.js or on devices using TF Lite in real-time, targeting **movement/fitness activities**. Two variants are presented:

- **MoveNet.SinglePose.Lightning**: A lower capacity model that can run >50FPS on most modern laptops while achieving good performance.

- **MoveNet.SinglePose.Thunder**: A higher capacity model that performs better prediction quality while still achieving real-time (>30FPS) speed. Naturally, *thunder will lag behind the lightning, but it will pack more of a punch.*

Model Specifications

Model Architecture

MobileNetV2 image feature extractor with Feature Pyramid Network decoder (to stride of 4) followed by CenterNet prediction heads with custom post-processing logic. **Lightning** uses depth multiplier **1.0** while **Thunder** uses depth multiplier **1.75**.

Inputs

A frame of video or an image, represented as an int32 tensor of shape: 192x192x3(**Lightning**) / 256x256x3(**Thunder**). Channels order: RGB with values in [0, 255].

Outputs

A float32 tensor of shape [1, 1, 17, 3].

- The first two channels of the last dimension represents the yx coordinates (normalized to image frame, i.e. range in [0.0, 1.0]) of the 17 keypoints (in the order of: [*nose, left eye, right eye, left ear, right ear, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle, right ankle*]).

- The third channel of the last dimension represents the prediction confidence scores of each keypoint, also in the range [0.0, 1.0].

Authors

(Equal contributions)

Francois Beletti, Google
Ard Oerlemans, Google

Yu-Hui Chen, Google
Ronny Votel, Google

Intended Use

Primary Intended Uses

- Optimized to be run in the browser environment using Tensorflow.js with WebGL support or on-device with TF Lite.
- Tuned to be robust on **detecting fitness/fast movement with difficult poses and/or motion blur**.
- Most suitable for detecting the pose of a single person who is **3ft ~ 6ft** away from a device's webcam that captures the video stream.
- Focus on detecting the pose of the person who is closest to the image center and ignore the other people who are in the image frame (i.e. background people rejection).
- The model predicts **17 human keypoints** of the full body even when they are occluded. For the keypoints which are outside of the image frame, the model will emit low confidence scores. A confidence threshold (recommended default: 0.3) can be used to filter out unconfident predictions.

Primary Intended Users

- People who build applications (e.g. fitness/physical movement, AR entertainment) that require very **fast inference** and **good quality single-person pose detection** (with background people rejection) on **standard consumer devices** (e.g. laptops, tablets, cell phones).

Out-of-scope Use Cases

- This model is not intended for detecting poses of multiple people in the image.
- Any form of surveillance or identity recognition is explicitly out of scope and not enabled by this technology.
- The model does not store/use/send any information in the input images at inference time.

Evaluation Data

- **COCO Keypoint Dataset Validation Set 2017:** In-the-wild images with diverse scenes, instance sizes, and occlusions. The original validation set contains 5k images (images, annotations) in total. The images which contain a single person are retained to be the evaluation set of this model, in total of 919 images. The dataset is chosen to evaluate the model performance in the general in-the-wild scenario.
- **Active Dataset Evaluation Set**: Images sampled from YouTube fitness, yoga, and dance videos which captures people movements. It contains diverse poses and motion with more motion blur and self-occlusions. The set contains 1161 images with a single person in the frame. This dataset is chosen to evaluate the model performance on the targeted domain, i.e. fitness/human motion.

Training Data

- **COCO Keypoint Dataset Training Set 2017:** In-the-wild images with diverse scenes, instance sizes, and occlusions. The original training set contains 64k images (images, annotations). The images with three or more people were filtered out, resulting in a 28k final training set.
- **Active Dataset Training Set**: Images sampled from YouTube fitness videos which captures people exercising (e.g. HIIT, weight-lifting, etc.), stretching, or dancing. It contains diverse poses and motion with more motion blur and self-occlusions. The set of images with a single person contains 23.5k images.

Factors

Groups

To perform fairness evaluation, we analyze the model performance under different person attributes and categories:

- **Gender**: Male/Female
- **Age**: Young/Middle-age/Old
- **Skin tone**: Medium/Darker/Lighter

Instrumentation

The training dataset images were captured in a real-world environment with different light, noise, and motion. Therefore, the model is robust to the input video streams that are captured through common devices' webcams.

Environments

The model is trained on images with various lighting, noise, motion conditions and with diverse augmentations.

Metrics

- **Keypoint mean average precision (mAP) with Object Keypoint Similarity (OKS)**: this is the standard metric used to evaluate the quality of the predictions of a keypoint model in the COCO competition.
- **Inference Time:** the time spent to run the model inference for a single image measured in milliseconds.

Quantitative Analyses

Prediction Quality

The following tables show the evaluation result for different attributes/categories. Both models perform fairly (< 5% performance differences between categories) on our targeted Active Single Person Image Set.

COCO Val2017 Single Person Image Set

Gender	**% dataset**	**Keypoint mAP (Lightning)**	**Keypoint mAP (Thunder)**
Male	63.1	67.4	78.7
Female	36.9	65.4	76.6

Age	**% dataset**	**Keypoint mAP (Lightning)**	**Keypoint mAP (Thunder)**
Young	72.2	65.6	76.6
Middle-age	17.1	68.0	78.0
Old	10.7	72.1	81.5

Skin Tone	**% dataset**	**Keypoint mAP (Lightning)**	**Keypoint mAP (Thunder)**
Darker	26.8	60.5	74.4
Medium	4.0	61.2	73.7
Lighter	69.2	74.4	82.9

Active Single Person Image Set

Gender	**% dataset**	**Keypoint mAP (Lightning)**	**Keypoint mAP (Thunder)**
Male	46.0	90.2	93.7
Female	54.0	87.8	92.3

Age	**% dataset**	**Keypoint mAP (Lightning)**	**Keypoint mAP (Thunder)**
Young	87.6	89.1	93.3
Middle-age	10.5	89.3	91.5
Old	1.9	85.7	90.0

Skin Tone	**% dataset**	**Keypoint mAP (Lightning)**	**Keypoint mAP (Thunder)**
Darker	15.4	89.1	93.1
Medium	2.5	92.2	93.3
Lighter	82.1	92.9	95.4

Speed Benchmark

The model was benchmarked using the TensorFlow.js benchmark tool in the Chrome browser on a few systems with either integrated or dedicated GPU. During the test, we enabled the "**pack depthwise**" option and took the **median** of the inference time over **500** runs.

Hardware Description	**GPU**	**Inference Time (Lightning)**	**Inference Time (Thunder)**
ThinkPad v7 gLinux laptop	*Intel UHD 620 Graphics*	39.0 ms	64.0 ms
Macbook pro 2019 13"	*Intel Iris Plus Graphics 655 1.5 GB*	17.8 ms	33.8 ms
Macbook pro 2019 15"	*Intel UHD 630 Graphics*	18.7 ms	44.6 ms
Macbook pro 2019 15"	*Radeon Pro 555X 4 GB*	16.4 ms	22.1 ms
Lenovo P520 2018	*GeForce RTX 2080Ti*	10.5 ms	15.0 ms