MoveNet.MultiPose

Model Details

A convolutional neural network model that runs on RGB images and predicts human joint locations of people in the image frame. The main differentiator between this MoveNet.MultiPose and its precedent, MoveNet.SinglePose model, is that this model is able to **detect multiple people in the image frame at the same time while still achieving real-time speed** . The model was designed to be run in the browser using TensorFlow.js or on devices using TensorFlow Lite, targeting **movement/fitness activities**. The released variant "*Lightning*" can run at >30FPS on most modern laptops and detects up to 6 people simultaneously while achieving good performance.

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. The model uses depth multiplier **1.5**.

Inputs

A frame of video or an image, represented as an int32 (for TF.js) or uint8 (for TF Lite) tensor of dynamic shape: 1x*HxWx3*, where *H* and *W* need to be a **multiple of 32** and can be determined at run time. A recommended way to prepare the input image tensor is to resize the image such that its larger side is equal to 256 pixels while keeping the image's original aspect ratio. Note that the size of the input image controls the tradeoff between speed vs. accuracy so choose the value that best suits your application. The channel order is RGB with values in [0, 255].

Outputs

A float32 tensor of shape [1, 6, 56].

- The first dimension is the batch dimension, which is always equal to 1.
- The second dimension corresponds to the maximum number of instance detections. The model can detect up to **6** people in the image frame simultaneously.
- The third dimension represents the predicted bounding box/keypoint locations and scores. The first **17 * 3** elements are the keypoint locations and scores in the format: [*y_0, x_0, s_0, y_1, x_1, s_1, …, y_16, x_16, s_16*], where *y_i, x_i, s_i* are the yx-coordinates (normalized to image frame, e.g. range in [0.0, 1.0]) and confidence scores of the *i*-th joint correspondingly. The order of the 17 keypoint joints is: [*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

remaining **5** elements [ymin, xmin, ymax, xmax, score] represent the region of the bounding box (in normalized coordinates) and the confidence score of the instance.

Authors

(Equal contributions)

Francois Belletti, Google

Yu-Hui Chen, Google

Ard Oerlemans, Google

Ronny Votel, Google

Intended Use

Primary Intended Uses

- Optimized to be run in the browser environment using TF.js with WebGL support or on-device with TF Lite.
- Tuned to be robust on **detecting fitness/fast movement with motion blur** poses.
- Most suitable for detecting the pose of multiple people who are **3ft ~ 6ft** away from a device's webcam that captures the video stream.
- Detect up to 6 people and their poses in the image frame.
- The model predicts **17 human keypoints** of the full body even when they are occluded. For keypoints that are outside of the image frame, the model will emit low confidence scores. A confidence threshold 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 multi-person pose detection** on **standard consumer devices** (e.g. laptops, tablets, cell phones). For the users whose application assumes only a single person in the image, we recommend checking out our MoveNet.SinglePose model.

Out-of-scope Use Cases

- 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 validation set contains 2.7k images (images, annotations) with humans in the 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. This dataset contains 1.9k images and 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 29k 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 contains 120k images.
- **Synthetic Training Set**: 21k procedurally generated synthetic images to compensate under-represented poses in the real image dataset.

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

- **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. A perfect

prediction will have OKS=1.0 while OKS=0.0 indicates that all keypoints are mis-detected or the detected locations are far away from the groundtruth locations. The OKS metric is computed for every groundtruth instance and is averaged within each group.

- **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.

COCO Val2017 Person Image Set

Gender	**% dataset**	**Average OKS**
Male	60.2	0.47
Female	39.8	0.42

Age	**% dataset**	**Average OKS**
Young	64.7	0.47
Middle-age	24.3	0.40
Old	11.0	0.39

Skin Tone	**% dataset**	**Average OKS**
Darker	25.9	0.40
Medium	1.9	0.69
Lighter	72.2	0.43

Active Person Image Set

Gender	**% dataset**	**Average OKS**
Male	49.6	0.84
Female	50.4	0.84

Age	**% dataset**	**Average OKS**
Young	83.8	0.84
Middle-age	14.0	0.83
Old	2.2	0.78

Skin Tone	% dataset	Average OKS
Darker	15.0	0.90
Medium	2.4	0.93
Lighter	82.6	0.87

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. The input shape of the model is set to be 256x256.

Hardware Description	GPU	Inference Time
ThinkPad v7 gLinux laptop	*Intel UHD 620 Graphics*	40ms
Macbook pro 2019 13”	*Intel Iris Plus Graphics 655 1.5 GB*	37ms
Macbook pro 2019 15”	*Radeon Pro 555X 4 GB*	24 ms
Lenovo P520 2018	*GeForce RTX 2080Ti*	19 ms