MODEL CARD

MediaPipe BlazeFace

MODEL DETAILS

A pair of lightweight models (224KB and 308KB in size) to detect one or multiple faces within an image captured by a smartphone camera, the first primarily targeting front (selfie) camera images and the second targeting those from the back-facing camera. Run super-real-time (~285 FPS for the front-facing and ~40FPS for the back-facing camera model) on a Pixel 2 CPU single-core via XNNPack TFLite backend.

For each detected face, returns:

- Facial bounding box coordinates
- 6 rough facial keypoint coordinates:
  - Left eye (from the observer's point of view)
  - Right eye
  - Nose tip
  - Mouth
  - Left eye tragion
  - Right eye tragion
- Detection confidence score

MODEL SPECIFICATIONS

Model Type
Convolutional Neural Network

Model Architecture
Convolutional Neural Network: SSD-like with a custom encoder.

Input(s)
RGB image (possibly a video frame) resized to 128x128 (front) or 256✕256 (back) pixels, represented as a 128x128x3 or 256x256x3 array of float values in the range [-1.0, 1.0].

Output(s)
A special encoding of SSD anchors for the detected faces. See SsdAnchorsCalculator and TfLiteTensorsToDetectionsCalculator configuration options in MediaPipe graphs for the front-facing and the back-facing camera models for details.

AUTHORS

Who created this model?
Valentin Bazarevsky, Google

Who provided the model card?
Yury Kartynnik, Google

DATE

October 17, 2019 (front-facing camera model)
September 05, 2019 (back-facing camera model)

CITATION

How can users cite your model?
V. Bazarevsky et al. BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs. CVPR Workshop on Computer Vision for Augmented and Virtual Reality, Long Beach, CA, USA, 2019.

DOCUMENTATION

Paper: https://arxiv.org/abs/2006.10204

LICENSED UNDER

Apache License, Version 2.0

Intended Uses

APPLICATION

Detecting prominently displayed faces within images or videos captured by a smartphone camera.

DOMAIN AND USERS

Mobile AR (augmented reality) applications

OUT-OF-SCOPE APPLICATIONS

- Counting the number of people in a crowd
- Detecting faces looking away from the camera, significantly inclined from the vertical orientation, or individuals' back of the head
- Detecting people too far away from the camera (e.g. further than 2 meters for the front-facing and 5 meters for the back-facing camera model)
- Any form of surveillance or identity recognition is explicitly out of scope and not enabled by this technology

Limitations

PRESENCE OF ATTRIBUTES

Produces only up to a given limit (e.g. 10) of detections even if more people are present.

TRADE-OFFS

The model is optimized for real-time performance on a wide variety of mobile devices, but is sensitive to face position, scale and orientation in the input image.

ENVIRONMENT

In presence of degrading environment light, noise, motion or face overlapping conditions one can expect degradation of quality and increase of "jittering" (although we cover such cases during training with real-world samples and augmentations).

Ethical Considerations

HUMAN LIFE

The model is not intended for human life-critical decisions. The primary intended application is entertainment and assistive technologies.

PRIVACY

This model was trained and evaluated on consented images of people using a mobile AR application captured with smartphone cameras in various "in the wild" conditions.

BIAS

The front-facing camera model is faster but only detects the faces that are relatively large. For the scenarios characterized by higher face size variance, please consider using the back-facing camera model, which is slightly more computationally demanding.

Training Factors and Subgroups

INSTRUMENTATION

- All dataset images were captured on a diverse set of front- and back-facing smartphone cameras.
- All images were captured in a real-world environment with different light, noise and motion conditions via an AR (Augmented Reality) application.

ATTRIBUTES

- Face roll angle should be not more than 45 degrees.
- Face bounding box sides should be at least X% of the corresponding image sides, where X is 20 for the front-facing and 5 for the back-facing camera model.
- At least 70% of the face bounding box should lie inside the input image.

ENVIRONMENTS

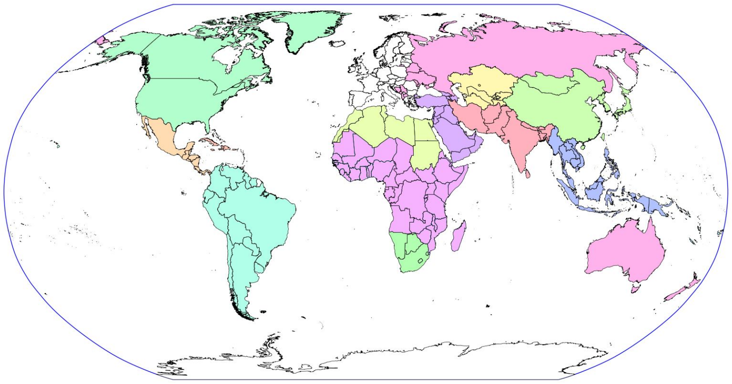
The model is trained on images with various lighting, noise and motion conditions and with diverse augmentations. However, its quality can degrade in extreme conditions. This may lead to increased "jittering" (inter-frame prediction noise).

GROUPS

To perform fairness evaluation we group user samples into 17 evenly represented geographic subregions (based on United Nations geoscheme with mergers):

Australia and New Zealand
Melanesia, Micronesia, Polynesia
Europe*
Central Asia
Eastern Asia
Southeastern Asia
Southern Asia
Western Asia
Caribbean
Central America
South America
Northern America
Northern Africa
Eastern Africa
Middle Africa
Southern Africa
Western Africa

*EU is excluded,
no from there is used



Evaluation metrics

Model Performance Measures

Average Precision, AP
Area under the interpolated precision-recall curve, obtained by plotting (recall@X, precision@X) points for different values of the decisive confidence threshold X.

True positives @X
Correct face predictions where there are faces (when thresholded by confidence >= X).
False positives @X
Incorrect face predictions where there are no faces (when thresholded by confidence >= X).
False negatives @X
Missed faces (when thresholded by confidence >= X).

Precision @X
True positive rate among all face predictions (when thresholded by confidence >= X).
Recall @X
True positive rate among all ground truth faces (when thresholded by confidence >= X).
Precision and recall are represented by point estimates as well as posterior probability distribution characteristics using the model following [1]. The characteristics used are: 95% credible interval, median, and mode.
[1] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation." European Conference on Information Retrieval. Springer, Berlin, Heidelberg, 2005.

Median IOD MAE
Median [over the data points] Interocular distance-normalized Mean [over the keypoints of one face] Absolute Error (MAE) on keypoint coordinates.
Interocular distance (IOD) is estimated as the distance between the eye centers computed as midpoints of segments connecting eye corners; MAE is represented as the percentage of the IOD.

Median IOD Jitter
Median Interocular distance-normalized "jittering" estimate.
MAE between backprojected results on slightly shifted images.
Evaluated similarly to Median IOD MAE, but instead of comparing the predictions of the model against the human annotations, they are related to the predictions of the model given a slightly shifted image (with an appropriate opposite shift of the results). Used to quantify the robustness of the model to e.g. camera movements.

Evaluation results

Geographical Evaluation Results

DATA

- **Front-facing camera model**
  - **Dataset I:** Face bounding box sides >= **20%** of respective image sides. Contains **720 samples: 680 images** evenly distributed across **17 geographical subregions** (see the specification in Section "Training Factors and Subgroups"), **40** images **per region**, **plus 40 images not containing faces** (the same set of no-face images is used in each region).
  - **Dataset II:** Face bounding box sides >= **15%** of respective image sides. Contains **1060 samples: 1020 images** evenly distributed across **17 geographical subregions** (see the specification in Section "Training Factors and Subgroups"), **60** images **per region**, **plus 40 images not containing faces** (the same set of no-face images is used in each region).
- **Back-facing camera model**
  - Contains **1350 samples: 1275 images** evenly distributed across **17 geographical subregions** (see the specification in Section "Training Factors and Subgroups"), **75** images **per region**, **plus 75 images not containing faces** (the same set of no-face images is used in each region).

All samples are picked from the same source as training samples and are characterized as smartphone front- and back-facing camera photos taken in real-world environments (see specification in "Training Factors and Subgroups - Instrumentation").

EVALUATION RESULTS

Detailed evaluation for the models across 17 geographical subregions is presented in the following external spreadsheets:

- Front-facing camera model
- Back-facing camera model

Definitions

BOUNDING BOX

A bounding box is an axis-aligned rectangle containing the object of interest (a face in our case).

KEYPOINTS

"Keypoints" or "landmarks" are prominent facial locations. The models represent them with (x, y) coordinates.

AUGMENTED REALITY (AR)

A technology that superimposes a computer-generated image on a user's view of the real world, thus providing a composite view.

Based on Model Cards for Model Reporting, In Proceedings of FAT* Conference (FAT*2019). ACM, New York, NY, USA