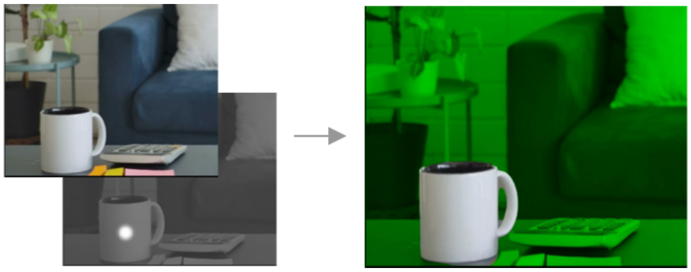
MODEL CARD

MediaPipe MagicTouch

MODEL DETAILS

A model (6MB size) to segment objects / animals / humans on the photos based on point, located within the segmentation mask of the object. Runs in real-time (~30 FPS) on a Pixel 7 via TFLite GPU backend.

Returns a segmentation mask for objects of interest.



Left: Input frame, prior map. Right: Output object mask.

MODEL SPECIFICATIONS

Model Type
Convolutional Neural Network

Model Architecture
Convolutional Neural Network: MobileNetV3-like with customized decoder blocks for real-time performance.

Input(s)
A frame of video or an image, represented as a 512 x 512 x 4 tensor. Channels order: red, green, blue, prior map with values in [0.0, 1.0]. Prior map is encoded as one-hot encoded pixels where the point-of-interest is located

Output(s)
512 x 512 x 2 tensor with masks for background (channel 0) and person (channel 1) where values are in range [MIN_FLOAT, MAX_FLOAT] and user has to apply softmax across both channels to yield foreground probability in [0.0, 1.0].

AUTHORS

Who created this model?
Valentin Bazarevsky, Google
Ben Hahn, Google

DATE
Mar 13, 2023

LICENSED UNDER

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

Intended Uses

APPLICATION

Object segmentation from photos in interactive applications.

DOMAIN AND USERS

- Photo editing
- Object annotation

OUT-OF-SCOPE APPLICATIONS

- Small objects / people too far away from the camera (e.g. further than 14 feet / 4 meters).
- Human segmentation that is not consented to by the human being segmented is out of scope and not enabled by this technology.
- Any form of surveillance or identity recognition is explicitly out of scope and not enabled by this technology.

Limitations

PRESENCE OF ATTRIBUTES

This model may segment multiple objects present in the scene particularly if they are located very closely and may represent once common essence.
Some thin features of objects, animals or humans such as fingers might occasionally be missed in the mask.

TRADE-OFFS

The model is optimized for real-time performance in the web browser and on a wide variety of mobile devices, and may not provide pixel perfect masks.

ENVIRONMENT

When degrading the environment light, adding noise, or fast motions, or including large occluders, one can expect degradation of quality of the predicted mask.

Ethical Considerations

HUMAN LIFE

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

PRIVACY

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

Training Factors and Subgroups

INSTRUMENTATION

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

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.

Evaluation metrics

Model Performance Measures

IoU, Intersection over Union

We evaluate the performance of our model by computing the ratio of the intersection of the predicted mask with the ground truth mask, and their union for the person class. Typical errors occur along the boundary of the true segmentation mask and may move it by a few pixels or lose thin features.

Evaluation results

Skin Tone and Gender Evaluation Results

DATASET

Contains 1870 images, captured in a real-world environment with different light, noise and motion conditions via an AR (Augmented Reality) application, which were annotated with perceived gender (male and female) and skin tone (from 1 to 10) based on the monk scale.

FAIRNESS CRITERIA

We consider a model to be performing poorly for a particular group if any region is further away than 3 stdev from the average of the model's performance across regions

EVALUATION RESULTS

Detailed evaluation across genders and skin tones is presented in the tables below.

FAIRNESS METRICS & BASELINE

We asked 7 annotators to re-annotate the validation dataset, yielding a person IoU of **98.74%**
This is a high inter-annotator agreement, suggesting that the IoU metric is a strong indicator of the person's segmentation mask.

FAIRNESS RESULTS

Examine whether subgroups are within three standard deviation away from the metrics of the entire dataset:

Evaluations of validation datasets result in average performance of 87.7% with a range of [85.3%, 89.4%] across all skin tones. The maximum discrepancy between worst and best performing categories is 4.1%.

Evaluations across gender yield an average performance of 87.7% with a range of [87.3%, 88.0%]. The maximum discrepancy is 0.7%.

Observed discrepancy across different genders and skin tones is less than one defined in our fairness criteria. We therefore consider the model performing well across groups.

Skintone	Test subset items and %		IOU	STDEV
Tone 1, 2	798	42.67%	86.4	16.0
Tone 3	394	21.07%	88.1	14.0
Tone 4	129	6.9%	89.2	12.9
Tone 5	80	4.28%	**89.4**	11.2
Tone 6	164	8.77%	87.8	17.1
Tone 7	227	12.14%	87.8	13.9
Tones 8, 9, 10	78	4.17 %	**85.3**	15.3
Average			87.7	14.3
Range			4.1	5.9

Skin tone evaluation, IOU

Gender	Test subset items and %		IOU	STDEV
Male	751	41%	**87.3**	14.7
Female	1081	59%	**88.0**	14.3
Average			87.7	14.5
Range			0.7	0.4

Gender evaluation, IOU

Definitions

AUGMENTED REALITY (AR)

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

INTERSECTION OVER UNION

A measure of similarity. In the segmentation case, the ratio between the area of intersection of two masks and the area covered by their union.