EuroCC@Turkey

https://eurocc.truba.gov.tr/

This document is prepared by EuroCC@Turkey for EuroCC under GA NO 951732

CASE STUDY REPORT

Image Content Moderation Project

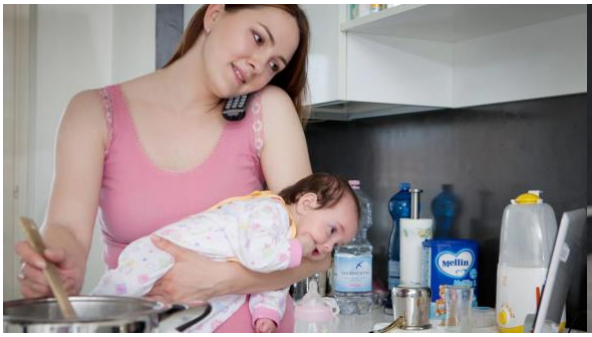
NCC Partner	*TRUBA*
Company*	*SİGMA ARGE BİLİŞİM TEKNOLOJİLERİ TİCARET (Machinetutors) –* https://www.machinetutors.com/
Expert	*Machinetutors: Samet Hiçsönmez - samethicsonmez@machinetutors.com* *TRUBA: Onur Temizsoylu - onur.temizsoylu@tubitak.gov.tr* *TRUBA: Kerem Kayabay – kerem.kayabay@tubitak.gov.tr*
Start & End Date	*17.05.2021 - TBD*

***Company** accepts that the Case Study Report is shared with the EuroCC Project and the community through the EuroCC@Turkey awareness creation activities and platforms.

1. Problem Identification

In this project, we tackle the problem of real-time image-based content moderation.

As part of a web-based content filtering system, being developed by one of our global clients, the developed model will analyze images and detect the presence of certain content types to be moderated. The current list of content types includes {offensive images, clothing type detection, underwear & swimwear, presence of real and or synthetic humans, age and gender of present humans}. For each image to be analyzed, an output value for each tag will be presented. Two example use cases for the described image categorization problem are given below.



Offensive	Contains Adult	Offensive	Contains Adult
Contains Female	Contains Baby	Contains Female	Contains Baby
Contains Male	Contains Child	Contains Male	Contains Child
Low_neckline	Underwear	Low_neckline	Underwear
short_skirt	bare_chest	short_skirt	bare_chest
Contains Humans	Synthetic Humans	Contains Humans	Synthetic Humans

The system will be trained with a custom dataset crawled from web images with proper licenses and annotated by our internal tools. In terms of performance evaluation, it is critical that the system is balanced towards user experience such that browsing experience. Since the outputs of the categories described above will be used to block images while browsing the internet, having false negatives will give the appearance of not being able to create a safe browsing experience. On the other hand, having too many false positives will impede browsing experience. Thus, categorization thresholds for each category will be determined individually to customize user experience according to user preferences.

2. First Suggestion

As an initial strategy, we will be implementing two models and using them in a pipeline to achieve highly accurate image categorization. The first model will be an image categorization model based on deep learning utilizing popular image classification backends. The second

model will be an object detection model for detecting age and gender from body bounding boxes.

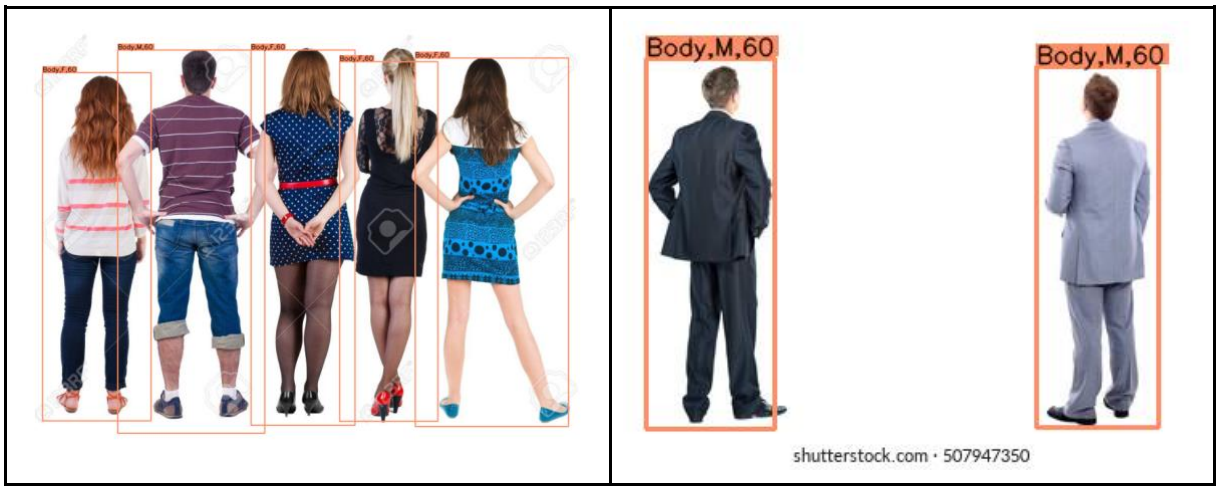
We will be training a multi class categorization model with efficientnet backends to detect the presence of significant content in images. Since the dataset will be imbalanced with each image containing a few positive and a lot of negative tags special methods such as label smoothing and using loss functions capable of handling class imbalance will be preferred. The resulting model will be exported as an onnx / tensorrt file for efficient deployment.

Current age&gender detection methods are “two stage”, i.e. in the first stage there is a face detection model which localizes all visible faces in the given image, then in the second stage age&gender detection is performed on each detected face box. This results in longer run times when the image contains more faces. To overcome the limitations of existing models, we propose a unified body, face, age and gender detection method. This model will be able to localize visible people and faces, and perform age&gender detection at the same time. This will reduce run time dramatically compared to two stage variants.

3. Solution Stage – I

While developing the NSFW content moderation model, we have collected over 500k images with over 5 million annotations. The model predicts three levels of offensiveness. In addition to that it predicts 6 different clothing attributes and whether the image contains synthetic human or real human picture. Providing clothing attributes helps the model to generalize better and prevent it inferring from superstitious features such as human skin, bed etc. The model has achieved over %88 percent in f1 score average over all the labels (11). Improvement in the model has come to a saturation point around this score. To make further improvement, we have been doing adversarial attacks to the model in order to spot weaknesses of the model. We have been targeting data collection based on the findings in the error analysis process.

In order to train our body, age and gender model, we first collected a large amount of data since there are no public datasets available which contain body bounding boxes with associated age and gender labels. Our current dataset contains more than 50.000 images and 170.000 annotations. Then, we developed a one stage body, age and detection model which performs better than state-of-the-art on selected publicly available facial age & gender datasets such as Adience [1] and AgeDB [2]. Our model also runs under 10ms irrespective of the number of people in the given image. Below we provide some example images where current face-based age & gender methods fail to perform any detection. In these figures, box captions follow the ''Body, Gender, Max_Age'' template.



[1] https://talhassner.github.io/home/projects/Adience/Adience-data.html#agegender
[2] https://ibug.doc.ic.ac.uk/resources/agedb/

4. Solution Stage – II

After the both models reached a certain performance level on the collected datasets, we focused on the false positive and false negative results of the models. We collect dedicated datasets which contain only false positive or negative images and re-trained the models with these datasets. These trainings further improve the NSFW model with 3.5% and Age&Gender model with 3% accuracy. NSFW model has achieved over 91.5% percent in F1 score average over all the labels (11). Age&Gender model reached an accuracy of 69% for age and 92% gender detection on Adience [1] dataset, and 73% age and 94% gender detection on AgeDB [2] dataset. At this stage, we are doing targeted data collection and fine-tuning the models to further improve the results.

5. Results and Achievements

Targeted data collection and model fine-tunings further improved the NSFW model with 0.5% and Age&Gender model with 1% accuracy. Finally, NSFW model has achieved over 92% percent in F1 score average over all the labels (11). Age&Gender model achieved 70% age and 92.5% gender detection accuracies on Adience [1] dataset, 73% age and 94% gender detection accuracies on AgeDB [2] dataset. Both NSFW and Age&Gender models are deployed to production and actively in use by the customers of our client. And the feedback from the customers is positive almost all the time. With the TensorRT acceleration, NSFW model runs around 200 to 300 FPS, and Age&Gender model runs around 100 to 200 FPS. Also, both these methods perform on par or better than the other commercial products.