DeepAttraction: using transfer learning to predict photo popularity

Ivan Bogatyy, Stanford University Michael Tom, Radiate Inc.

Stanford University

Presentation: https://youtu.be/JUvazzty7Pw

Introduction

We explore the problem of predicting photo attractiveness (measured by ratio of "likes" to "dislikes" on a social app).

Inspired by A. Karpathy's experiments on selfies, we use a convolutional neural network architecture and a transfer learning setup to fine-tune a network initially trained on a much larger CV dataset. Key contributions:

- Improved predictions: FaceNet, MSE loss, robust proprietary labels
- Inception vs FaceNet comparison

Problem Statement

Given a photo and no other information, the goal is to predict its ratio of left and right "swipes" (as proxy for attractiveness).
We use MSE loss weighted by total number of swipes for both training and evaluation.
As a simpler yet interesting experiment, we also attempt to predict gender (using log-loss for training and accuracy for evaluation).

Dataset

Data graciously provided by Radiate Inc., a social app for music festivals with Tinder-like swiping mechanics.
We only used the main profile photo and swipes that happened after the receiving user last updated their main photo (ensuring labels are relevant to corresponding photos).
Statistics: **76k** users with valid photos, **42:58** gender ratio, **40m** total swipes, **11m** swipes post-filtering.
To provide some context on the label distribution, below are average LIKES/(LIKES+DISLIKES) ratios grouped by genders of the source and target users.

	→ F	→ M
F →	**0.33**	**0.24**
M →	**0.75**	**0.31**

Methods

Since no previous baselines exist for the problem, our goal was to do a first attempt at exploration. Concrete questions:

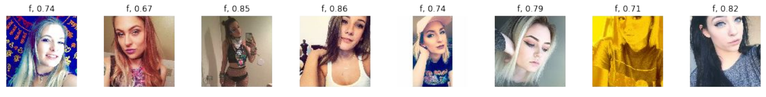
- How satisfactory are model predictions subjectively?
- Using existing SotA CV research (Inception, FaceNet), how far ahead of the baselines can we get without task-specific engineering?
- Given the data contains human subjects, would more task-specific FaceNet perform better than generic Inception? (note: A. Karpathy's experiments found no benefit)

Our models were designed as follows. We load a pre-trained model, either Inception v3 or FaceNet, and build one, two or three FC layers on top of its penultimate layer. We then fine-tune the model on Radiate labels.
FaceNet however needs a separate network to segment faces and provide bounding boxes before it can run, and some photos produced zero bounding boxes or more than one. We threw out all such examples (~50% of data), essentially having two experimental datasets (filtered one marked as *).

In addition to ratio labels, we repeat all the same experiments attempting to predict gender.

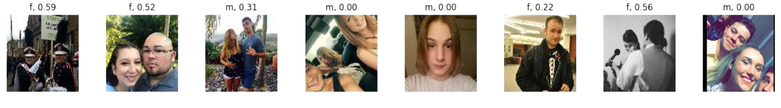
Subjective results

Highest predictions by FaceNet on test with true labels:



Lowest predictions by FaceNet on test with true labels:

Highest gender prediction log-loss by Inception on test.
Mostly group photos or erroneous labels.



Evaluation

Model	Gender acc on Test
Baseline	58% (most frequent gender)
Inception	86.7%
Inception*	87.9%
FaceNet*	93.8%

Model	Ratio MSE on Test
Baseline	0.187 (best const prediction)
Inception	0.0225
Inception*	0.0217
FaceNet*	0.0135

* indicates photos where segmenter NN found exactly 1 face.

Note that while "starred" dataset is indeed simpler for Inception too, FaceNet clearly outperforms Inception in a comparable setting (where it had been provided a bounding box).

Conclusion

Our results show that the problem of predicting photo attractiveness is clearly amenable to transfer learning computer vision methods.
MSE loss is reduced ten-fold by training a ConvNet model (compared to a baseline predicting a single best constant), and predictions robustly match subjective judgements.
Further, contrary to previous findings, we find that using a neural network pre-trained on a more relevant task yields significant improvement, even for same architecture (FaceNet = Inception pretrained on LFT+YTF instead of ImageNet).

Gender prediction can be performed with accuracy ranging from 86.7% in the most general setup to 93.8% in the case where segmenter NN thinks there is only one face in the photo (note that headroom is below 100%).