Versuchsdesign

2 Modelle auf Basis von Google NIMA trainiert. Diese funktionieren mit Transfer Learning über CNN (Convolutional Neural Networks).

Modell trainieren

Trainierdaten: AVA dataset [Murray et al. 2012] (ästhetisches Modell) und TID2013 dataset [Ponomarenko et al. 2013] (technisches Modell). Letzteres ist mit optisch verzerrten Bildern befüllt.

Trainierprozess: MobileNet Architektur mit ImageNet[1] Gewichten für das Modell verwenden und den letzten dichten layer in MobileNet durch einen dichten layer der auf 10 Klassen output liefert (Bewertungen 1-10).

Error Function: EML – Earth Mover's Loss → Misklassifikation einer echten 10 als eine 4 ist schlimmer als eine Fehlklassifikation einer echten 5 als 4. →Besonderheit von NIMA

Mean Score: Es ist davon auszugehen, dass der Mean Score beim AVA Datensatz einfach das arithmetische Mittel aller Bewertungen ist. Dieser wird somit auch leicht berechnet. Der TID2013 Datensatz ist für mein Experiment eher irrelevant. In der Abbildung sieht man wie die Bewertungen auf der DPChallenge Webseite erscheint.

Statistics

Place: 1 out of 319
Avg (all users): 7.3648
Avg (commenters): 8.7600
Avg (participants): 7.3101
Avg (non-participants): 7.4327
Views since voting: 12026
Votes: 233
Comments: 61
Favorites: 49 (view)

Um EML auf NIMA anzuwenden, benötigt man eine Verteilung für die Bewertungen über alle 10 Klassen für jedes Bild. Im AVA Datensatz existieren diese bereits. Im TID2013 Datensatz mussten diese aus dem Mean Score (diese wird im Datensatz gegeben) inferiert werden.
2 Trainierphasen:

1. Erstmals nur den letzten dichten layer trainieren, und das mit einer hohen Lernrate.
2. Nach dieser „Burn-In" Periode trainiert man alle Gewichte des CNN mit einer niedrigen Lernrate

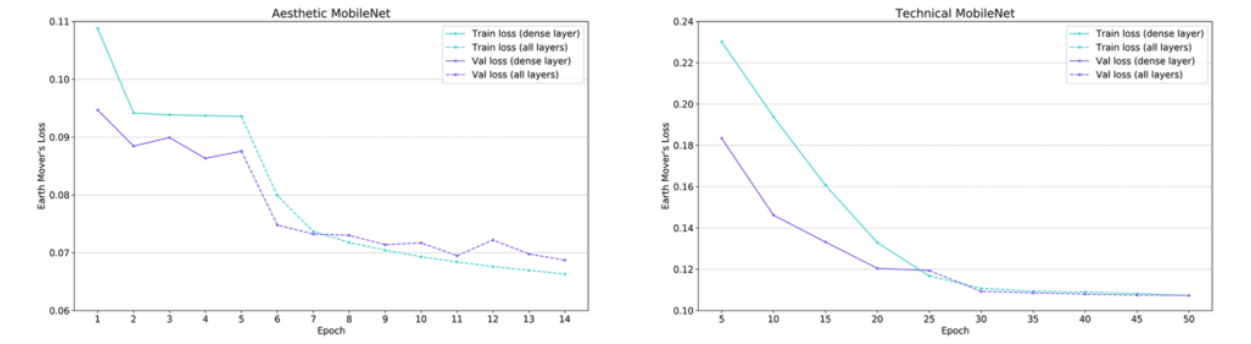

Abbildung 1: Ästhetisches Modell links, technisches Modell rechts

[1] ImageNet Gewichte sind für Objekterkennung optimiert

Hier stoßen die Macher auf das Problem, dass das Modell auf das AVA Datenset trainiert ist, obwohl dieses auch Bilder von Essen und anderen nicht relevanten Bildern beinhaltet. Zudem bestand auch das Problem, dass die Vorhersagen des Modells dann nicht den ganzen Spielraum von 1-10 ausschöpfen. Siehe folgende Abbildungen 2 und 3.

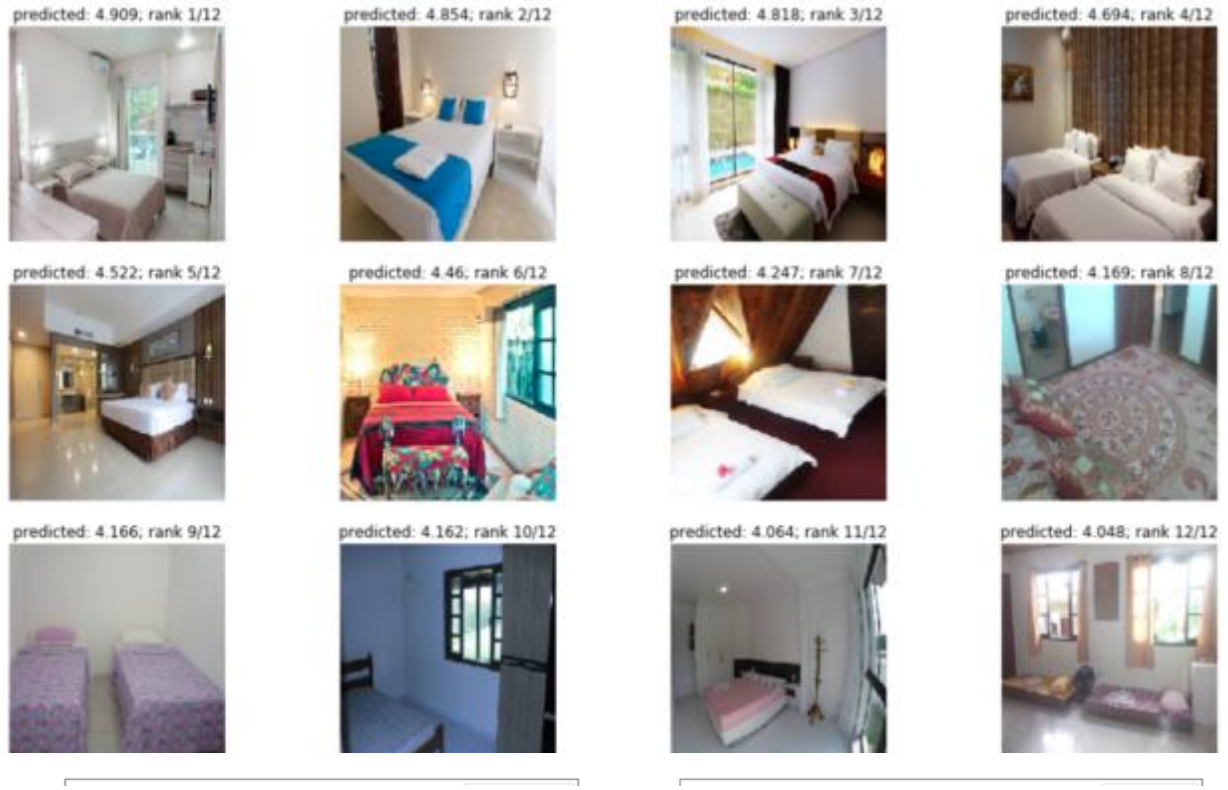

Abbildung 2: Der gesamte Spielraum wird nicht ganz ausgeschöopft, woher weiß man welcher Score nun für ein Bild „gut" ist?

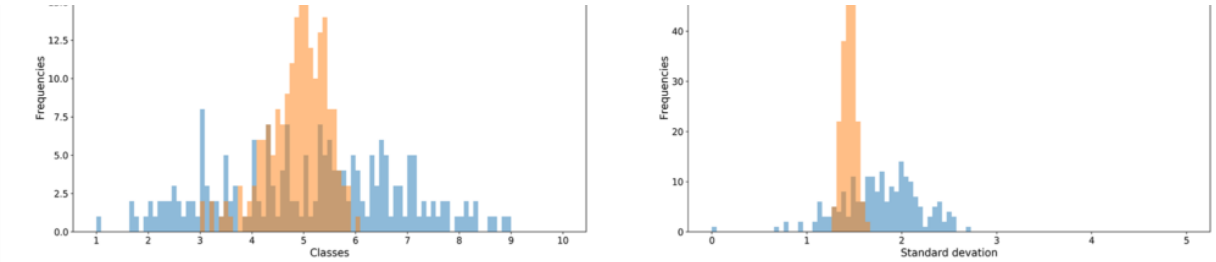

Abbildung 3: Beide Plots beziehen sich auf das ästhetische Modell in der ersten Iteration
links: Verteilung der Häufigkeiten der (blau) echten Bewertungen und der (orange) vorhergesagten Bewertungen
rechts: Verteilung der Häufigkeiten der (blau) echten Standardabweichungen und der (orange) vorhergesagten Standardabweichung

Somit haben sie sich entschieden, 1.000 eigene Hotelbilder (n = 10 Personen pro Bild) selbst zu bewerten und zu labeln. Sinnvolle Ergebnisse kamen bereits nach den ersten hunderten Bildern zustande. Dann haben sie 800 davon benutzt, um das ästhetische Modell weiter zu trainieren und die restlichen 200 Bilder als Testbilder. Somit hat das Modell dann noch mehr Bewertungsspielraum ausgeschöpft und Bilderbewertungen *besser* (im Sinne von menschlicher Ästhetik und Erwartungen für Hotelbilder) vorherzusagen (also ist zum Beispiel ein Badezimmerbild als Cover-Bild schlechter bewertet als ein Schlafzimmerbild).

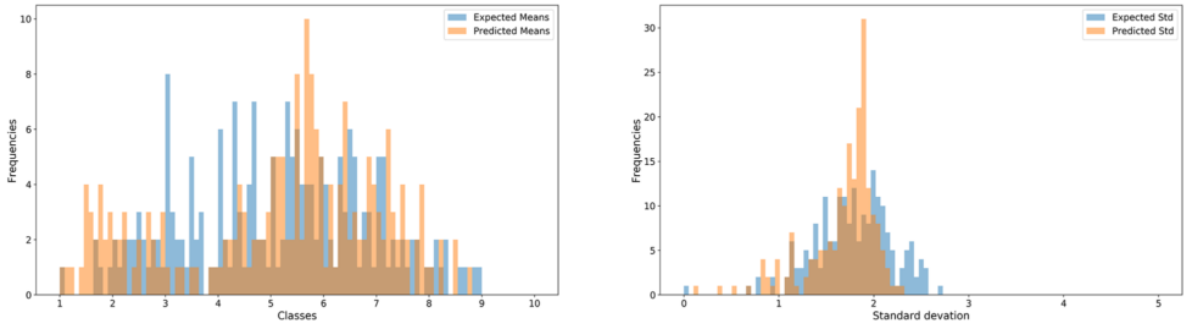

Abbildung 4: Der ganze Bewertungsspielraum wird nun ausgeschöpft

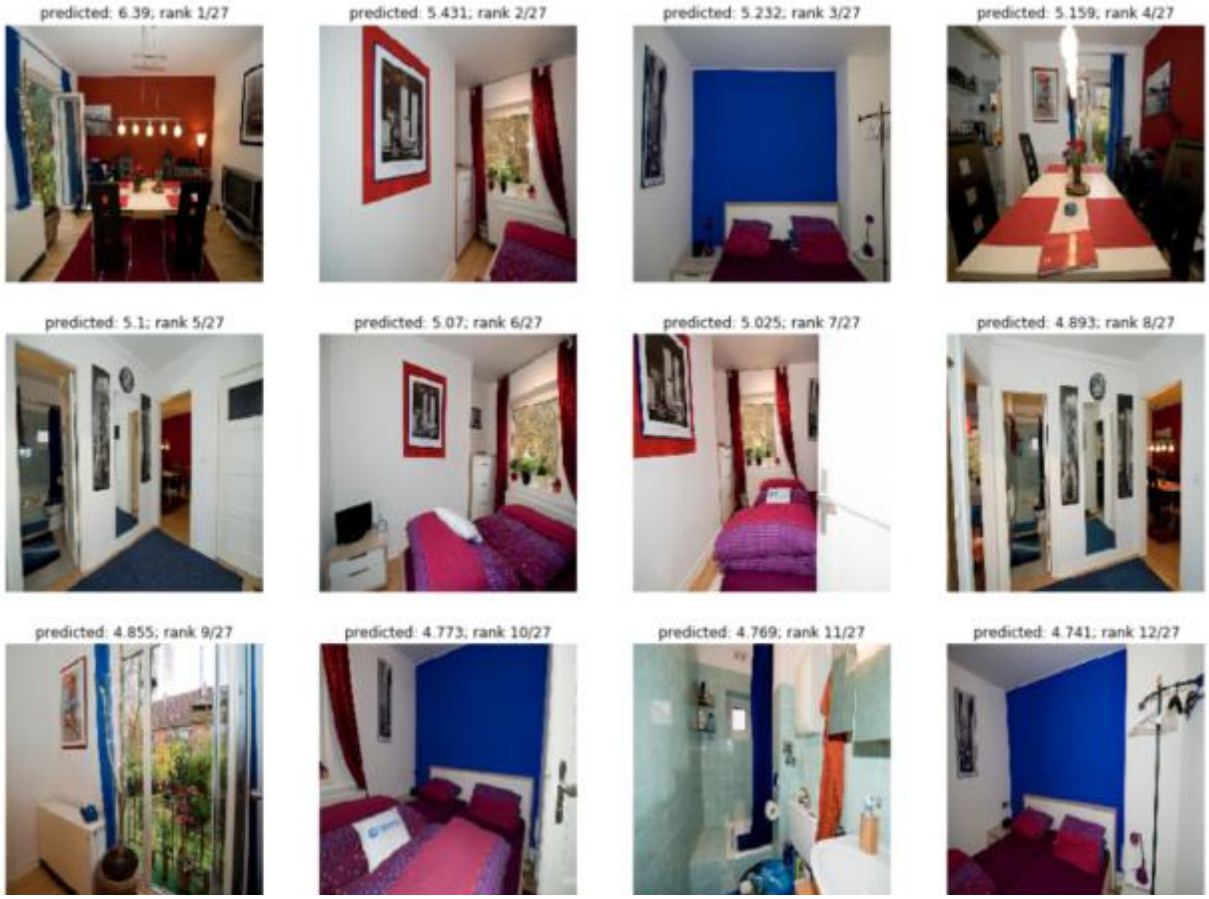

Abbildung 5: Die letztendlichen Vorhersagen.

Quellen:

https://github.com/idealo/image-quality-assessment
https://developer.nvidia.com/blog/deep-learning-hotel-aesthetics-photos/
https://en.wikipedia.org/wiki/Earth_mover%27s_distance
Christopher Lennan