HOTCHAT3000.COM

MODEL DESCRIPTION

Hotchat3000 implements a prediction server that rates the attractiveness of a person on a scale of 1 to 10 based on their submitted photo. These ratings are predicted by a large machine learning model, which we will describe in detail in this document.

Our task was to find a model and channel its inner critic towards all the selfies we would throw at it. We arrived at CLIP, a large machine learning model trained by OpenAI that matches visual concepts with written captions.

CLIP has been trained on a dataset of 400 millions image-text pairs to predict which image goes with which caption. The model consists of two parts: an image encoder (Residual Network (ResNet) or Vision Transformer (ViT) that encodes images and a text encoder (Transformer) that encodes captions. The image and text encodings are two vectors that contain high-level features describing the image and caption, respectively. CLIP is trained using contrastive learning to predict the correct image-text pairing in a batch of many possible image-text pairs. In other words, CLIP is trained to choose the correct caption for an image among a number of incorrect captions.

The relevant property of CLIP is that it can score arbitrary captions by how well they describe an image, a process the authors call zero-shot transfer to downstream tasks. The trick is to come up with useful captions that describe the downstream task of rating a person, e.g., "she is beautiful". Every caption represents a category of attractiveness (i.e., a label): (1) beautiful / handsome, (2) average looking, and (3) ugly. The model presented with these captions will score each one according to its own understanding of beauty concepts (e.g., "beautiful", "ugly") and how they relate to what's in the picture. We decided to use a small number of captions that clearly describe the attractiveness categories to avoid conceptual overlaps, which would make it harder to reason about model predictions. Once CLIP outputs a vector of probabilities (i.e., "scores") that the person belongs to each category, the last thing left for us to do is to turn the scores into a hotness rating on a scale of 1 to 10.

This conversion is where we inevitably start injecting our own biases. We assigned every person the same starting base rating and depending on how well they score in each category of attractiveness, they can either increase or decrease their rating by some weighted amount. We approximated the initial weights for each category by training a simple linear regression model on subsets of SCUT-FBP5500 and "Hot or Not" datasets. Both datasets are heavily biased towards certain ethnicities, and not at all representative of the broader population. We additionally collected a small album of pictures from Google Search that included a broader selection of people of different ages, races, genders, etc. and used this to further manually adjust our weights and evaluate the performance of the model. We assert that the magnitude of our manual adjustment of the CLIP output is small relative to the biases inherent in the model and training sets, as our primary goal was to exacerbate the relative differences between scores, to produce a relative 10-point spread.

The outputs of LLMs are a reflection of the data they were trained on. Hotchat3000 very deliberately sets out to expose, visualize, and exacerbate these biases.

1. CLIP was released by OpenAI (official article: https://openai.com/research/clip).
   Model weights and source code can be downloaded at
   https://github.com/openai/CLIP.
   Alternatively, you can use the Hugging Face platform
   (https://huggingface.co/docs/transformers/model_doc/clip).
   We opted for the official repository on GitHub.
   Paper: https://arxiv.org/abs/2103.00020

2. The SCUT-FBP5500 dataset for facial beauty prediction was released by Lingyu Liang, Luojun Lin, Lianwen Jin, Duorui Xie, Mengru Li at the Human Computer Intelligent Interaction Lab of South China University of Technology. It can be accessed at
   https://github.com/HCIILAB/SCUT-FBP5500-Database-Release.
   Paper: https://arxiv.org/abs/1801.06345

3. The "Hot or Not" dataset collected from the once popular website hotornot.com was released by Jeff Donahue and Kristen Grauman at the Dept. of Computer Science, University of Texas at Austin. It can be accessed at
   https://vision.cs.utexas.edu/projects/rationales/#data.
   Paper: https://vision.cs.utexas.edu/projects/rationales/rationales.pdf