Insights from Personal Photos

Jeff Gonda, Kevin Feng, Kexin Huang
Youhee Choi, Jackie Kim, Yuntao Wang

Why this project?

Our social media profiles often portray our lives quite accurately, and the information we decide to publicly post reveals traits of our personalities. Our team wanted to use the predictive power of data to get even more quantitative and specific -- *is it possible to narrow down a person's Myers-Briggs personality type through their profile picture?* Most people already have a basic understanding of what an **Extroverted** person's profile is expected to look like, as opposed to an **Introverted** person's, but what about a **Judging** person's versus an **Perceptive** person's? Or **Sensing** vs. **Intuitive**? **Thinking** vs. **Feeling**? We were interested in seeing which tags most frequently arose from each Myers-Briggs group in order to have a better visual understanding of the subtleties between each personality type.

Problem statement

We want to develop several machine learning models to see which is best suited for our data input of a **profile photo's tags** and most capable of accurately predicting the corresponding person's **4-letter Myers-Briggs type.**

Our Journey

In our presentation, we divided our project "journey" into **4 different phases.** In this report, we will further explain the challenges behind completing phase.

1. **Collect Data**
    a. Tools: Facebook API, Python Webscraping, Amazon AWS S3 Storage
    b. Problems & Solutions:
        i. **Noisy Data:** Our initial project scope relied on Twitter profile photos and Myers-Briggs hashtags, but our initial investigation showed that most of the photos were "troll" and/or didn't have people's' faces in them. For better quality data, we used Facebook and Myers-Briggs group members. We also removed duplicate data points of people who had joined multiple groups.
        ii. **Large File Size:** We pulled 2,000 photos for each MBTI, which took some computational time and was over 4GB total. We used zip files, USBs, Google Drive, and AWS S3 to transfer and store data.

2. **Tag Photos**
    a. Tools: Amazon Rekognition API, Python scripting, OpenCV
    b. Problems & Solutions:
        i. **Available Resources/API:** After receiving suggestions from our mentor, we tested out Clarifai, Google Cloud Vision, and Amazon Rekognition APIs and weighed each one's pros and cons (ie. price point, computational cost, facial analysis, scene detection, sentiment analysis). We ended up choosing Amazon Rekognition for the best balance of tradeoffs, and used OpenCV to identify faces.
        ii. **Computational Effort:** Amazon API took more time/computational power than expected, so we tagged 1,000 (out of 2k) photos from each of the 16 groups.
3. **Train Models**
    a. Tools: NLTK Wordnet, SKL Decision Trees/SVM/Logistic Regression, Keras, TensorFlow
    b. Problems & Solutions:
        i. **Matrix Sparsity:** In total, there were 1761 features but only 916 of them appeared with significant frequency (more than 5 times). Because the tags we were getting from the API were so specific, this would lead to overfitting. We grouped the tags that showed up a significant number of times into broader categories using NLTK Wordnet similarity scores to increase the frequency of the broader categories. This increased our test accuracy rate from 53% to 55%.
4. **Apply Insights**
    a. Tools: Matplotlib, Google Slides

Solution system

Our project is unique in its exploration purpose. With further improvements to accuracy, a prediction model like this could help develop our understanding of social media archetypes and may be applied to businesses for marketing, targeted ads, or user research.

Our solution system included **KNN, Logistic Regression, SVM, Random Forest** models and a **CNN**. The hypertuned Random Forest model gave us the best results, achieving >55% test accuracy. We trained these models using ~12,000 Facebook photos and Amazon Rekognition API, as detailed in the "Journey" section of this report. From these, we could visualize a tag correlation matrix and see the most distinctive characteristics of each MBTI letter, as described in our presentation.

Possible future improvements

- Incorporating **multiple profile photos** of each person for a more comprehensive perception of personality. At this time, we were limited by Facebook security measures that prohibited us from viewing multiple photos of people that have tighter privacy settings and no mutual friends as us.
- Investigate **other social media sources**, such as Instagram, for photos that have less to do with direct facial features that would still give personality-indicative information.
- Being able to pull **data from statuses/likes/groups** and seeing if adding non-image data found on social media profiles would give better tags. Our mentor suggested that hobbies and interests may be more indicative of personality type than profile photos.

Team contributions

- **Jackie:** low-tech demo, report
- **Jeff:** AWS S3 setup, Amazon API tagging
- **Kevin:** Facebook API photos, feature grouping, SKL models, graphs
- **Kexin:** SKL models, CNN
- **Youhee:** low-tech demo, final presentation, mentor communication
- **Yuntao:** SKL models, CNN

Mentor collaboration

- **Peter Cnudde:** One of our team members met Peter during the mentor mixer, and he suggested this project idea. There was some miscommunication, and we were assigned another mentor from Yahoo (Gerry), but we touched base with Peter towards the end of the project. He gave us very specific approach directions, suggesting that we use a pre-learned neural net and take the last cut of the layer to preprocess the images.
- **Gerry Pesavento:** Some miscommunication on project direction when we contacted him, but was very helpful in guiding us through our initial project stages. Suggested trying out some image content APIs that were critical to our data analysis.