Tilburg University

School of Humanities

Master Thesis: MSc Data Science: Business & Governance

Academic Year 2016-2017

Profile Pictures as Predictor for Network Size on Twitter

Author:

T.J.J. Rockx

ANR: 551809

Supervisor & Second Reader:

C.D. Emmery MSc

N.J.E. van Noord MA

Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science of Data Science
at the Tilburg School of Humanities
of Tilburg University

July 2017

Preface

In front of you lies the thesis "Profile Pictures as Predictor for Network Size on Twitter". It has been written as partial fulfillment to complete the master Data Science: Business & Governance at Tilburg School of Humanities. I was engaged in the research and writing of this thesis from January, 2017, until August, 2017.
Together with my supervisor C.D. Emmery MSc, I defined the topic and research question. He provided the dataset and the accompanying tool for labeling. Additionally, he provided the machine to run the expensive computations. My supervisor was responsive to any question and he could help where needed. I would like to express my sincere gratitude to C.D. Emmery MSc for his guidance, support and giving me the opportunity to work like a real data scientist. Furthermore, I want to thank my second reader for his feedback and critical review.

Thomas Rockx

Abstract

Several methods have addressed the prediction of user attributes such as age and gender by profile pictures. Users value the profile picture and they provide more information than what is shown. Network size may be related to how the user expresses oneself visually. In this paper this problem was posed as a classification task. Different convolutional neural network architectures and a traditional classifier were used to explore the optimal method for predicting in which class the user belonged. A dataset from Twitter and a manually labeled subset for age and gender were used. Experimented with different tests including annotations and data augmentation, no support was found for the posed problem.

Contents

Appendices **33**

1 Introduction

Social media have seen a significant increase in use over the past years (Lenhart et al., 2010). Privacy is a growing concern and a growing number of users manage their privacy settings on social network sites and selectively remove friends, comments or photo tags (Madden, 2012). It seems users are cautious about privacy missteps and try to avoid these. Yet, social media may (unknowingly) cause the exposure of personal information beyond the intention of the user (Kosinski et al., 2013). Research by Kosinski et al. (2013) has shown that it is possible to accurately predict personal attributes such as age and gender from an individual's social media profile. The negative implication Kosinski et al. (2013) note, is that this is possible without individual consent and without the user noticing it. The richness of information has led to research efforts that study users' behavior and attributes. The ability to identify these (latent) user attributes successfully, especially age and gender, can have many consequences for targeted advertising and personalized services (Rao et al., 2010).

This latent information also holds true for (profile) images. An example of a marketing firm using images for personalized targeting is Cluep, which offers services on social media that are able to detect among others brands, scenarios, emotions to enhance the targeting options based on pictures posted by users (clu, 2017).

Most social network sites are profile-based and therefore the profile picture forms a central part of a social network (Hancock and Toma, 2009). Subsequently, social media encourage setting such a picture, which cannot be (completely) hidden on most social network sites and thus information may always be exposed via this attribute. Because of the prevalence and importance of the profile image, users seems to attach extra value to the profile picture above other parts of the online self-representation, which refers to how a user wishes to appear and thus what to expose intentionally, and research has been conducted to uncover what drives users choosing a certain picture. For example, research by Kapidzic (2013) has found that narcissism is a significant predictor when setting up a new profile picture. This research also shows that pictures are not randomly chosen, but carefully chosen, as they are considered an important part of the online self-representation (Kapidzic, 2013). Profile pictures are chosen strategically in order to reflect an ideal rather than the user's actual self (Ellison et al., 2006). Kapidzic's work shows that people care how they are perceived by others. It seems profile pictures are used to convey a reflection chosen by the user and as a result expose more information than the intention of the user.

Using Online Information to Make Judgments

The deliberate choice of profile pictures may be explained by how these are perceived. Walther (1992) suggests that users process the social information in "computer-mediated communication", and this special way of communication and exposure to affect the amount of friends the user has network size (i.e. network size) (Walther et al., 2008). In their research, photos posted by friends on the user's profile were shown to influence user's physical attractiveness. Being better-looking than one's friends showed no effect, while having more good-looking friends did. Moreover, another study regarding social attractiveness and extraversion found that users who have too small or too large a number of friends are perceived more negatively than those with an optimal network size (Tong et al., 2008). In the study, the optimal network size was related to the number of friends the raters had. For example, raters with 100 friends judge an individual to be less like them if s/he has more friends than the rater does. Users base their judgments on more than one aspect of a profile and it seems that pictures and network size both play a role.

On social networks, people seem to use information that is presented to them to form a judgment. Profile pictures tend to be chosen after consideration, which implies that it is important how the user is perceived. Furthermore, while network size is a system-generated number, it does affect how a person is perceived. While there is no research that directly links profile images and network size, both seem to influence online behavior. This leads to the hypothesis that there is a connection between profile pictures and network size.

Using Profile Pictures to Predict User Behavior and Characteristics

The content or style of an image is obvious from its visual content. Network size may not be concretely visible in a picture. However, research shows it is possible to predict latent information in an image. A study by Redi et al. (2015) has for example shown that it is possible to use profile pictures to predict the ambiance of a place to which people had checked in online with FourSquare. This was done by matching (stereotypical) characteristics from visitors to places without them actually having taken the picture at that place. Their method extracted information from the images, such as emotions and self-presentation characteristics. An example of a self-presentation characteristic is whether the picture shows a 'natural' face or depicts the user wearing reading glasses. They also focused on the stylistic characteristics of an image. For example, how the face is photographed, as opposed to object recognition, which considers which objects are present. These extracted features were then tested for correlations with different ambiances. This study provides additional evidence that profile pictures can be used as a predictor for latent visual information.

Combining the works of Madden (2012), who show that privacy is a growing concern, Kapidzic (2013) and Ellison et al. (2006), who show that the content of profile images is an intentional choice, with the works of Walther et al. (2008) and Tong et al. (2008), who show that the perception of images generate judgments, users with a certain network size may be identified by the content of the profile image. For example, popular users may attach more value to their profile picture than less-popular users, who may not care about their online self-representation. While there is no difference in privacy management between young and adult users, there are differences between age groups and gender with regard to the "pruning" of the online profile (Madden, 2012). Furthermore, real life differences may be reflected online. To illustrate the differences: younger people tend to have different preference for clothing styles than older people and wearing makeup is more accepted for women. Such (stereotypical) characteristics may apply to the online self-representation.

1.1 Theoretical and Practical Relevance

Given the objective to find patterns or objects that relate to a certain network size, knowledge of what these objects are can provide insights into individual's considerations when choosing a profile picture. The findings could also be useful for individuals who want to expand their network for marketing purposes or simply to increase the reach of their posts. Furthermore, they could potentially explain aspects of an individual's personality; for example, it may be that if a user favors a certain appearance, they find themselves into a certain network with similar people. Moreover, from a marketing firm perspective, it makes (better) personalized targeting possible. Additionally, the findings could reveal that machine learning can be used to predict user attributes that are not directly visible in an image, yet contribute to an online self-representation. While it is difficult to predict someone's network size, especially by looking only at an image, machine learning techniques may see (abstract) things that a person does not which relate to network size. If prediction is found to be possible, the contribution of this thesis is two-sided: it shows that there is a relationship between profile images and network size, and it exemplifies the application of machine learning for latent user attributes.

1.2 Scope of the Research & Research Question

This thesis thus addresses the challenge of the combination of profile pictures and network size. More specifically, the thesis aims to use profile pictures to predict the network size of social media users. As such, this study aims to answer the following research question:

> **To what extent can profile pictures be used to predict network size?**

1.3 Research Method

In contrast to Redi et al.'s method, this study uses state-of-the-art machine learning techniques to analyze images and train a prediction method. Developments in computer vision have led to the development of convolutional neural networks (CNNs), which have been used to study images and effectively apply them for object recognition. Using Twitter as the platform of choice, given a set of profile pictures and a set of network size categories, the classification objective is to maximize the accuracy of predicting these categories. Using CNNs for object recognition could make it possible to find patterns or objects related to network size, and certain patterns or objects might prove to relate to certain network sizes.

1.4 Overview of the Thesis

In order to answer the research question, an overview of the current state of machine learning for object recognition and the content of profile images is discussed is given in Section 2. The method used for the experiments is detailed in Section 3. The dataset is reviewed in Subsection 3.1 and the evaluation method in Subsection 3.3. In Section 4 the results are presented. In Section 5 the results are reflected with regards to the research question. The conclusion is presented in the final section (Section 6).

2 Background

Machine learning provides an automated approach to studying images and is therefore currently an important component in the field of computer vision. This section begins by presenting an introduction to machine learning techniques for computer vision, followed by an introduction to age and gender prediction. The importance and information a profile picture can hold is discussed afterwards. This section will be wrapped up with the discussion of network size and how it is related to the profile picture.

2.1 Computer Vision

Computer vision can be explained from two points of view. From a biological viewpoint, computer vision aims to create computational models that mimic the human visual system. From an engineering viewpoint, the aim is to build a system that can autonomously perform some of the tasks that the human visual system performs and surpass human performance on these tasks (Huang, 1996). Huang (1996) described two major difficulties in computer vision:

> (1) How do we distill and represent the vast amount of human knowledge in a computer in such a way that retrieval is easy? (2) How do we carry out (in both hardware and software) the vast amount of computation that is often required in such a way that the task (such as face recognition) can be done in real time?

. The current approaches to these two problems have been found over time with some notable events as Malisiewicz (2015) describes in his blog. He observes that the methods for recognition systems are similar to those of earlier systems, yet higher performing systems are needed.

2.2 Machine Learning and Deep Learning

One of the major improvements that has been noted by Malisiewicz (2015) and that is enabled by more computational power is machine learning, which is a branch of computer science. Its goal is to develop methods (algorithms) for automatic pattern detection in data and to then use these patterns for prediction (Murphy, 2012). Deep learning is a variant of representation learning, which is a broader family of machine learning techniques that extend beyond neural networks. LeCun et al. describes deep learning as "computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction" 2015. Deep learning methods have been applied in different fields, such as translation and computer vision, where the results for a select set of tasks are comparable to or better than human performance (Krizhevsky et al., 2012). Improved performance is achieved by the collection of larger datasets, more powerful models and better techniques that allow to generalize better to new data (Krizhevsky et al., 2012).

Learning thousands of objects from millions of images requires models with increased learning capacity, and this has been recently shown to be most effectively applied by convolutional neural networks (Krizhevsky et al., 2012). Convolutional neural networks, also referred to as "ConvNets", have gained attention due to their success in image and video recognition (Simonyan and Zisserman, 2014). This success is attributed to the availability of large image repositories such as ImageNet (Deng et al., 2009) and increasingly powerful computer systems (Dean et al., 2012). CNNs are typically built with the use of many layers. The convolutional layer extracts features from the image and preserves the relationship between pixels. This is often followed by an activation layer that introduces non-linearity in the model as real-world data is often non-linear. This result is then passed through a pooling layer, which is often called subsampling and it reduces the dimensionality of each feature map while retaining the most important information (LeCun et al., 2015). After these layers, the obtained high-level features are then passed to traditional neural network layers to form the prediction. A thorough explanation of CNNs, and deep learning in general, is discussed in LeCun et al. (2015).

These developments have only been recent as these architectures were computationally unfeasible in the past. Earlier research in computer vision has focused on smaller recognition tasks such as digit classification. LeCun et al. (1995) have surveyed several classification methods for this classification task using the popular MNIST dataset, including CNNs. The authors evaluated on the aspects of accuracy, memory requirement, training- and recognition time. Their CNNs achieved the highest accuracy, while requiring little memory for the weights and recognition time was quick for most CNNs. However, the training times were by far the longest. This lack of computational power restricted (deep) convolutional nets, as noted in Dean et al. (2012). Later, in 1998 LeCun et al. reviewed several methods more extensively for the same classification task using newer machines. In their paper, they concluded that CNNs have error rates close to traditional methods, while running much faster. Their self-crafted LeNet-5 architecture (LeCun et al., 1995) showed good performance in both studies and this architecture will be used in experiments.

More recently,Simonyan and Zisserman from the Visual Geometry Group conducted a thorough evaluation of deep convolutional networks architectures, achieving first and second places in the 2014 ImageNet Challenge. The challenge of that year was a classification and localization task of 1,000 object categories, such as "dog", "soccer ball" and "electric guitar". As training data, 1.2 million labeled images were provided (Russakovsky et al., 2015). Simonyan and Zisserman's architecture had a classification error of 7.5% for their so-called "VGG 16" architecture. To put this into perspective, the human error rate is 5.1% (Karpathy, 2014), but humans take considerably longer to perform this task. The proposed networks also proved to be successful on other datasets and across several tasks, i.e. different classification and regression problems. Their best-performing networks, the VGG 16 and VGG 19, have been implemented in popular frameworks. The full VGG 16 configuration and its applications are discussed in Simonyan and Zisserman (2014).

VGG 16 is one of several available top deep learning CNNs, and the architecture will be used in several experiments to leverage the learned features without the need to rely on available data (Chollet, 2016). Additionally, the LeNet-5 architecture will also be used next to the VGG 16 architecture, because of the performance advocated by LeCun et al. to test for difference in performance. This configuration is considered to be narrow compared to the VGG 16 architecture, because the configuration based on LeNet-5 has a smaller learning capacity, such that the network has to focus on the most important features.

2.2.1 Use of Machine Learning to Predict Age and Gender

The prediction of user attributes by the application of machine learning methods has been focused on the estimation of age and gender, e.g. Rothe et al. (2015); Eidinger et al. (2014) whose approaches are similar to the proposed method in this thesis, showcasing the effectiveness of CNNs in the prediction of user attributes. The observations from their experiments will therefore be discussed here.

Several machine learning competitions have focused on the prediction of age and gender via images. One example is ChaLearn LAP, which is a challenge focused on apparent age estimation from images (ChaLearn, 2017). The winner in 2015, Rothe et al., had an error rate which was significantly lower than human error rate, using the VGG 16 architecture.

A notable difference in the methods of several studies is the alignment of faces (Rothe et al., 2015; Eidinger et al., 2014). Regular images are in-the-world (i.e. having poor alignment), whereas constrained images are filtered to meet a certain quality or condition and thus yield better results. Eidinger et al. (2014) present a face alignment technique to overcome the in-the-wild faces, which detects facial features and it is able to transform the image accordingly. While this thesis does not focus on the prediction of age and gender, prior research on the use of faces in the wild is especially relevant. Eidinger et al.'s research aligns well with the proposed classification task because of the use of noisy data.

2.3 Profile Pictures as a Source of Data

Examples of studies that use profile pictures as the source of data demonstrates that faces in images (partly) reflect personality traits and can show what a user likes (Mehdizadeh, 2010; Kapidzic, 2013), and that photo content (Hu et al., 2014). These studies often do not use an automated procedure with CNNs, or deep learning techniques, for the content analysis or information extraction, but make use of contextual information, surveys or other computer vision techniques. Hu et al. (2014) for instance used a computer vision technique (SIFT) for the object recognition and then calculated the Euclidean distance between obtained features to cluster the images, which revealed eight different categories of images posted on Instagram. SIFT was one of the earlier leading-edge techniques in computer vision and needed hand-crafted features to learn, while newer methods make it possible to automate this (Malisiewicz, 2015).

While there is no study that applies profile pictures in an automated deep learning method, the current state-of-the-art techniques as mentioned may be used for this. Furthermore, images from social media are often used for studies to train techniques (e.g. ChaLearn (2017) which consists partly of images from Flickr) or to evaluate how the techniques would perform on a social media (e.g. Becker and Ortiz (2008) applied face recognition techniques to validate their performance on in the wild images on Facebook).

2.4 Network Size

Given that the current research focuses on the the prediction of network size by the use of profile pictures, network size is therefore the target. Because of the way networks are established, it is worthwhile to consider the characteristics to be expected.

The sum of all an individuals' connections constitute his/her network size. A study by Hill and Dunbar (2003) has shown that the maximum (offline) social network size in Western societies is around 150 people, as measured by Christmas cards. The average network size was around 125 cards. While Twitter, the platform of interest in this study, does not release statistics about follower numbers, KickFactory, a marketing and advertising firm, claims the average amount of followers a person has is 707 (KickFactory, 2016). Their sample size of nearly 96 million active users shows that 93.1% of users have under 1,000 followers, around 6.3% have between 1,000 and 10,000 followers and only 0.6% have over 10,000 followers. The difference between offline and online network size suggests that online networks consist of people with whom the user has a positive connection, which may either be with people from the offline or online network. Furthermore, younger people are more likely to have larger networks (Ong et al., 2011). As age and gender tend to have a moderating effect on network size, it is worthwhile to run an experiment with this information implicitly included as features to test for this effect.

The size of networks may also be explained by for what purpose people connect with others. Networks on Twitter vary from other social networks as Kwak et al. (2010) have found that 85% of all trending topics are headline news or persistent news. The study found deviations from known characteristics of social networks. A notable deviation is the degree of separation, i.e. the average number of hops between two (random) people. It was found to be quite short, especially for the size of Twitter. This could indicate that the networks consist (partially) of people who share similar affinities, and that people follow others for information purposes.

Different networks sizes seem to be related to individuals and their personal characteristics. To illustrate, research has shown that highly extraverted adolescents have a larger social network size and post more photos (Ong et al., 2011). Personality traits (Mehdizadeh, 2010; Kapidzic, 2013) and user attributes can be predicted by profile pictures (Rothe et al., 2015; Simonyan and Zisserman, 2014; Karpathy, 2014). Both have an effect on network size and this effect suggests that there is a relation between the profile picture and network size.

3 Method

The prediction task consists of a set of unfiltered images to predict a category of network size by using machine learning techniques, including deep CNNs. A dataset from a social network site is need which includes profile pictures and network sizes of a group of users. Given that there are no predetermined architectures shown to work well for the prediction task, an exploratory analysis is conducted to determine an initial best performing method.

3.1 Dataset

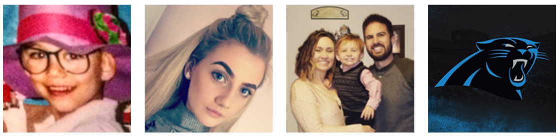
The Twitter data from Emmery et al. (ress) was used for the experiments, which was retrieved using a query that specified to return profiles with messages that self-report gender. As a result, the profiles were mostly English and from active users (Emmery et al., ress). It is a relatively small dataset with 36409 profiles.

The content of profile images on Twitter is diverse and thus the data can be considered noisy. Firstly, profiles can be from persons and bots or pages (i.e. non personal profile for promotion), and the content of the image can be anything. While most profiles are from actual people, their profile pictures do not necessarily contain a face at all; rather a cartoon or a photo from a celebrity.

Table 1: Size Dataset

Unlabeled	**Labeled**	**Total**
30296	6113	36409

Figure 1: **Randomly Sampled Images**



If a person shows himself on a photo, it may be from far away, close up or with more persons. Furthermore, the photo style itself may be different, for example altered using a filter, varying sharpness or diverse lightning conditions. Figure 1 shows the diversity of only four random images.

3.1.1 Annotation

In several of the experiments, age and gender information is added in order to test if it increases performance. Furthermore, it serves as control variable to identify factors relating these features. Twitter does not require users to fill in gender and age and if it is filled in, the information is often set to private. Therefore, manually labeling the profiles for age and gender was necessary. It was only performed on a subset of the entire dataset, because it is time intensive and only needed as control. Table B (found in appendix) describes the information displayed by the annotation tool. Through the use of a tool for annotation, parts of the profile such as description and profile picture were shown in a quick glance and direct links to the feed and media were easily accessible. Annotation of gender and age was based on visible parts of the profile. Additionally, the deciding factors for both labels were recorded. These so-called signals indicate which visible parts were the most important to determine age and gender.

As a result of the labeling, age, gender and signal are obtained. All used information for the descriptive analytics of the dataset and for the experiments is described in table A. The subset was labeled by three different annotators using the images and any visible contextual information, such as tweets and profile description.

3.1.2 Analysis Labeled Dataset

In this section the summarized information of the labeled dataset is provided. The labeled set has 6113 instances. While inferred gender was available for the unlabeled dataset, gender and age data is not acquired that way for this subset. Given this difference in obtaining the information, only the manual labeled subset is used for the statistics. Additionally, Twitter does not publicly

Table 2: Profile pictures count

Gender	**Count**	**Count (%)**	**With PP**	**With PP** (%)
Male	1877	30.7%	1216	64.8%
Female	3686	60.3%	3006	81.6%
Other	65	7.9%	43	66.2%
Bot/Page	485	1.0%	-	-
Total	6113	100%	4265	69.8%

Note: bots and pages do not have profile pictures

Table 3: Mean, Standard Deviation Continuous Variables. The highest followers amount is from the page of the Carolina Panthers

Variable	Mean	Standard Deviation	Min	Max
Age	26.8	10.7	9	77
Followers	2,993	43,940	0	2,004,810
Following	868	2,292	0	77,101

provide statistics about gender division, yet it is valuable to compare findings from this analysis with statistics from other work. If these align, the method may generalize better with data from Twitter.

A study conducted by Duggan et al. in January 2015 showed more men used twitter than women. Later that same year, a study showed that women are more likely to use social networking sites than men (Perrin, 2015). While the difference in usage between men and women is modest in the publication, it is strongly visible in the labeled dataset. Women accounted (60.3%) for twice the size of men (30.7%) of all profiles. The full information is found in table 2. In a filing (SECURITIES and COMMISSION, 2014), Twitter estimated that 8.5% of all accounts are bots. The findings show 7.9% of the accounts are bots, which does not deviate from the estimate. In a more recent study (Varol et al., 2017) the estimates are between 9% and 15%. Such a large share is not evident in our dataset.

The average age is 26.75 and the standard deviation of age is slightly over 10 years. Thus, younger generations are more prevalent in the dataset.

The amount of followers has a high standard deviation, which is mainly due to profiles with many followers. There were 16 profiles in this subset with no followers and the mean followers is

2993. The mean is higher than the claimed average of 707 by KickFactory, which may be explained by the many outliers as shown in figure 3. At 50% percentile, the followers count is 460. At 75% percentile 1.168 followers. Furthermore, the average, standard deviation and maximum of friends are all lower than the followers, which indicates that most user have a "positive" followers to friends ratio.

The spread of followers, however, is different between genders. Figure 3 clearly shows most of the followers are below 2,500, except for bots and pages. Their maximum amount lies around 4,000, if outliers are not taken into account.

Ong et al. (2011) have shown in their research age had a moderating effect on network size as younger people tended to have larger networks. Without any adjustments or leaving out any outliers, figure 6 illustrates that the effect is not apparent in the data.

While determining age and gender of online profiles, signals were documented during the annotation process. For users with profile pictures with their face in it, the image is the most important indicator of gender and age. This confirms that the profile image is an important factor to determine gender and age. Notably, for "others" the descriptions of profiles were the second indicator, while for male and female this were their tweets.

For profiles without labeled profile pictures, including bots and pages, the most important indicators were not the image. It shows the description and tweets are the factors to look for if their profile picture is not showing the user. Image is still in certain cases a suitable indicator, because it can be a cartoon portraying the person or showing text indicating user attributes.

On Twitter the following and followers count is separated. The relation between these two is found to be positively linear, despite using a method to optimize for the best fit. Figure 7 shows this relation, which may be explained partly by how the data is gathered and the appearance of the system-generated number related to the friends and followers count. For instance, users who only

Figure 2: Spread of age

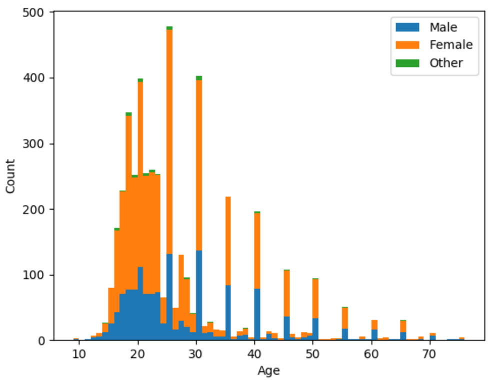


Figure 3: Spread of followers

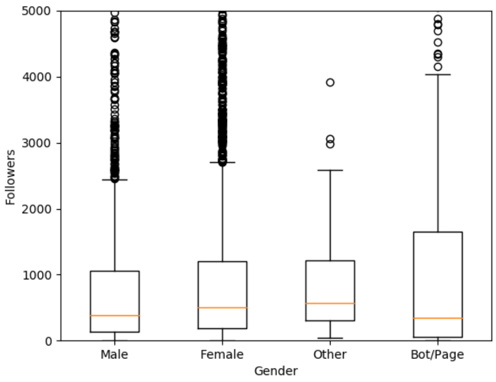


Figure 4: Indicators for gender and age for profile pictures with face

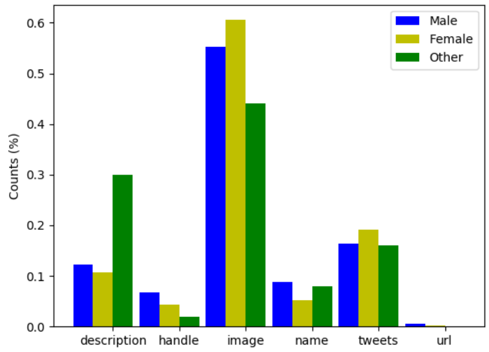


Figure 5: Indicator for gender and age for profile pictures without face

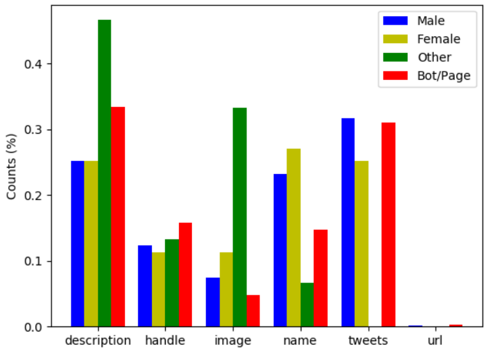


Figure 6: Age plotted against Followers

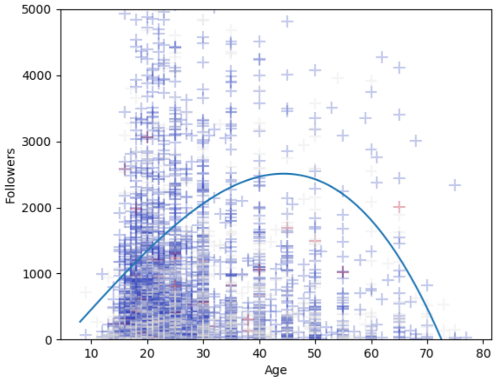
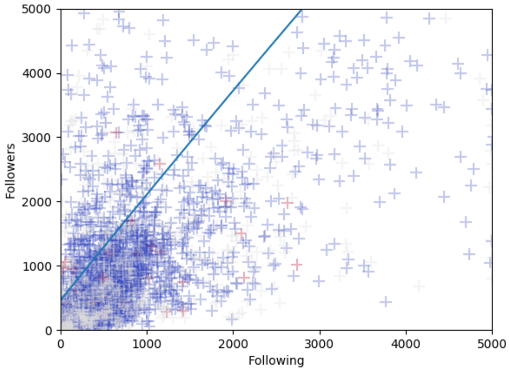


Figure 7: Relationship between friends and followers



follow others for news purposes and not post anything are filtered out by the query heuristic used for collection by Emmery et al. (ress), which results in such users having more friends than followers. On the other hand, having more followers than friends, may suggest to be a indicator for an optimal network size; as noted before, having an optimal network size influences social attractiveness and extraversion (Tong et al., 2008). The linear relationship suggest the optimal network size consist a positive friends to followers ratio and it could influence how a user is perceived if the friends to follower ratio is lower.

In short, the characteristics of my dataset do not align with claimed statistics by having a higher followers count and two times more females than men. Furthermore, the moderating effect of age is not present. The effect of these characteristics on the success on prediction is unknown. Each profile picture is unique and any pattern may arise. The experiments should determine whether gender and age have an influence.

Table 4: Division into 3 classes with the size for the subset and complete dataset before splitting into training, validation and test sets.

Dataset size	6113	36409
Range Classes	[0, 76) [76, 1947) [1947, 2004810]	[0, 70) [70, 1812) [1812, 14215657]
Count Classes (%)	0.15, 0.7, 0.15	0.15, 0.7, 0.15
Count Classes	923, 4274, 916	5501, 25447, 5461

3.2 Target Value

Twitter makes a distinction between the people that follow a user, referred to as "followers", and the people the user follows, referred to as "friends". A mutual connection is not necessary. Following someone is by default a one-way connection. For the prediction task at hand, it has been decided to use only the followers count. While using both the follower and friend counts would provide additional information, the number of followers is the number that is used to measure the degree of the user's appeal to a larger audience. To illustrate: a celebrity will easily have more followers than friends, because s/he is popular and well known. The number of followers therefore indicates the reach of the person. Additionally, having more friends does not lead to a user's post being seen by more people, while having more followers does. From this perspective, the followers count is the measurement for reach and thus network size.

A regression is expected to have a worse performance over classification for the current study's task, because of characteristic patterns of a specific class. Such patterns may not relate to an ordinal order of the classes. Therefore, the followers count, which is a continuous variable, is transformed into several classes with two different bin sizes.

The first transformation of the target value is into 3 bins to reflect small, normal and large network size and the classes represent 15%, 70% and 15% respectively. It is assumed that the majority will have a normal network size and the percentages reflect the rule of thumb of normal distributions to capture the majority within the range of +/- 1 times the standard deviation. The classes are shown in Table 4. Balancing the classes is performed by providing class weights to the algorithms. Randomly leaving out data for the purpose of equalization of the classes, because the dataset size would become smaller, or oversampling, because that makes it more likely that the architectures will overfit, were not preferred.

The second transformation of the target value is into 20 even bins, because that should be able to provide insight into what is predicted by the algorithms. As a result, a different evaluation protocol will be used for these 20 classes.

3.3 Evaluation

For the experiments using the network size binned in 3 classes, evaluation is done by measuring the accuracy on a validation and test set. Experiments in which 20 classes will be predicted, the mean absolute error (MAE) metrics and confusion matrices are also used for evaluation on a validation and test set. While the MAE is generally used in regression tasks, it is suitable for my task. The MAE provides a simple error measurement which essentially penalizes predictions that are further away, and it shows how close predictions are. Therefore, it functions similarly to a ranking method. The confusion matrices are used to identify in which area the most mistakes or correct predictions are.

The validation (15%) and test set (5%) are split from the entire set (see Table 5) and is equal for all experiments. The test set size is arguably small, but due to noisy data and thus unseen instances, a larger training set was preferred. For each of the experiments with the same dataset size (i.e. complete or labeled), the same validation and test set is used to ensure the validity of the results across the different experiments. Furthermore, the accuracy of a majority vote classifier is used to reflect obtained scores against the image classifiers.

Training is stopped before the image classifiers start overfitting by the so-called early stopping method (Morgan and Bourlard, 1990), which functions as a special form of regularization, and it is triggered after the accuracy on the validation set did not improve for 10 epochs. After training stops, the model version from the best performing epoch is chosen.

There are no previous reported results to the posed problem. Furthermore, the dataset is publicly available.

3.4 Input

The profile images from the annotated Twitter profiles are the input in every experiment. Twitter downsamples images to a tiny size or it leaves it in its original uploaded format, which could be any size. The original uploaded images were resized to 224 x 224 pixels, (which is the size used for the trained VGG 16 network) and converted to RGB. Due to the resizing, images could either have been upsampled or downsampled, depending on their original format.

Table 5: Size of training, validation and test sets.

	Complete	**Labeled**
Train	29127	4890
Validation	5461	917
Test	1821	306

Table 6: Age groups and their count.

	9-12	13-19	20-36	37-65	66-77
Count	23	1503	3927	633	27

For one instance of the models used in this experiment, gender and age have been added as extra features. If age was missing, it was filled with the mean. Sequentially, age has been manually binned according to age groups used in a research by Gallagher and Chen (2009). Table 6 shows the groups and their counts. Finally, gender and age have been transformed into an indicator variable by one hot encoding. Gender, while the name is suggestive, can represent male, female, other and bots/pages.

3.5 Preprocessing

For each experiment the data is normalized by subtracting the mean for each image channel. The input for the models using the VGG 16 architecture are normalized by using the means from the trained network in order to resemble the network (Simonyan and Zisserman, 2014; Chollet, 2016). The 'centering' on the images for the experiments using the LeNet architecture is performed by using the mean of the input data and is based on the approach of Chollet (2016).

From Keras (Chollet et al., 2015) the built-in image data generator is used to artificially increases the amount of images and this is expected increase performance as it prevents overfitting (Krizhevsky et al., 2012), especially since the dataset is small. Furthermore, in the blog of Chollet (2016), the shown performance increase led to the choice of including the generator and this is expected to happen in the experiments. The images are randomly augmented by rotating up to 60 degrees, zooming in up to 20% and flipping horizontal. While Figure 8 shows four possible augmentations, every picture is fed twice to the model, which effectively doubles the training data while not using the exact image twice.

3.6 Feature Extraction & Prediction

Now that the target, data and input have been discussed, the different methods for feature extraction and prediction are explained. The proposed experiments follow the general pipeline described

Figure 8: Original and randomly augmented profile pictures.

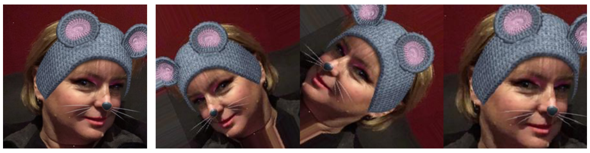


Table 7: Naming scheme of the models. VGG: Representation of the VGG 16 architecture. VGG-P: Prediction of the full VGG 16 architecture. FC: Self-made fully-connected layers. RTF: Retrained FC. LeNet: Self-adjusted LeNet-5 architecture. Note that the full architectures are discussed in Section 3.6.1

Name	**Purpose**
VGG-GBC	Combination of the representation of the VGG 16 architecture, which is then fed to another "traditional" machine learning approach. It serves as a second baseline
VGG-P-SVM	Combination of a prediction by the VGG 16 architecture followed by a "traditional" machine learning approach to serve as a baseline.
VGG-FC	This model trains the "bottleneck" features from the VGG 16 architecture, followed by self-made fully connected layers that are specified for the task at hand.
VGG-RFC	Same architecture as VGG-FC, yet additional layers are retrained. In essence, this model fine-tunes the VGG-FC model to the proposed task.
LeNet	Compared to the models using the VGG 16 architecture, it is a narrow CNN and it exemplifies as a simple model.

in Table 8, while the details of each step and model will be discussed in this section. The models are based on the implementations demonstrated by Chollet (2016), which showed good results with simple and effective implementations. The names for the five models reflect their architecture and these are shown in 7.

Three different aspects are tested to see if they improve performance. The aspects are inclusion of labeled data, larger dataset size and image augmentation. The performed experiments to test this are chosen based on training time. A general overview of all experiments is found in Table 9 and the full pipelines for each experiment split by model is found in the appendix (appendix C).

3.6.1 VGG 16 Architecture

VGG-GBC: This baseline model uses a representation obtained from the VGG 16 model. The representation is the extracted output before the fully connected layers (i.e. the output of a convolutional layer) and flattened from three dimensions to one dimension. The architecture uses the weights from the trained VGG 16 network on the ImageNet dataset and classes. It is then passed through to a Gradient Boosting Classifier (GBC). This classifier uses an ensemble of decision trees to form a better prediction model (Pedregosa et al., 2011), which can handle unbalanced classes and this classifier is to some extent robust against overfitting. In several experiments, the gender

Table 8: General pipeline structure. Note: The steps are an indication and do not reflect individual experiments.

Models:	VGG-P-SVM / VGG-GBC / VGG-FC / VGG-RFC / LeNet
Step 1: Input	Images, gender & age
Step 2A: Preprocessing	Center features
Step 2B: (Optional) Image Augmentation	Rotation & zoom
Step 3: Feature Extraction	Trained VGG 16 architecture / LeNet-5 architecture
Step 4: Prediction	GBC / SVM / Softmax

Table 9: Overview of all tests. Labels and Augmentation refer to the inclusion of the annotated data and use of image augmentation.

Name	**Labels**	**Bins**	**Augmentation**	**Size training data**
VGG-GBC	Yes	3	-	4890
VGG-GBC	-	3	-	4890
VGG-GBC	-	3	-	29127
VGG-P-SVM	-	3	-	4890
VGG-FC	-	3	Yes	29127
VGG-FC	-	3	-	29127
VGG-RFC	-	3	-	29127
LeNet	-	3	-	4890
LeNet	-	3	Yes	29127
LeNet	-	3	-	29127
VGG-GBC	Yes	20	-	4890
VGG-GBC	-	20	-	4890
VGG-P-SVM	Yes	20	-	4890
VGG-P-SVM	-	20	-	4890
VGG-P-SVM	-	20	-	29127
VGG-FC	-	20	-	29127
VGG-RFC	-	20	-	29127
LeNet	-	20	-	4890
LeNet	-	20	-	29127

and age information is included.

A common approach is to use a Logistic Regression or a Linear SVM to obtain initial results as these are simple to understand machine learning methods. However, images have as many features as pixels and the representation from the VGG network is large (i.e. 512x8x8 features) and the training times could grow exponentially with the amount of samples, especially with a SVM (Nalepa, 2017). The GBC generally has shorter training times. Initial tests in a small set showed the training times were over twice as long and achieved equal accuracy when using SVM rather than GBC. Furthermore, if age and gender are included, the data has mixed feature types that are better handled by this classifier.

Initial tests with the GBC showed that the maximum depth of three individual regression estimators at combined with 200 estimators showed good performance and these parameters have subsequently been used for the actual test.

VGG-P-SVM: The learnings of the trained VGG 16 architecture are leveraged by using the predictions of the full model, which are then used as input for a SVM. This also lowers the training times as it has less features for each instance compared to using the representation. The SVM is chosen for this final prediction task because of the generally good results. Gender and age features are added in one experiment.

VGG-FC: Similar to the baseline models, the VGG 16 architecture is used as basis and the representation is extracted. On top of this basis, self-made fully-connected layers are stacked (Chollet, 2016). The whole model is then trained. Table 10 shows the architecture of these layers. The dropout regularization prevents overfitting as it alternately randomly disables neurons during training to learn independent representations (Srivastava et al., 2014). The final output layer is adapted to the same amount of units which correspond to the amount of classes.

VGG-RFC: This model uses the exact same architecture of the VGG-FC, i.e. VGG 16 representation with fully-connected layers stacked on top, which is shown in table 10. The weights are initialized from the VGG-FC model in order to leverage their learned task and not destroy them

Table 10: The fully-connected layers from VGG-FC and VGG-RFC

Input: VGG Architecture
Flatten layer
Dense: 512 units
ReLu activation
Dropout: rate 0.5
Dense: 3 units
Softmax activation

(Chollet, 2016). The last layer of the VGG 16 model and the fully-connected layers are re-trained, which essentially fine-tunes the architecture.

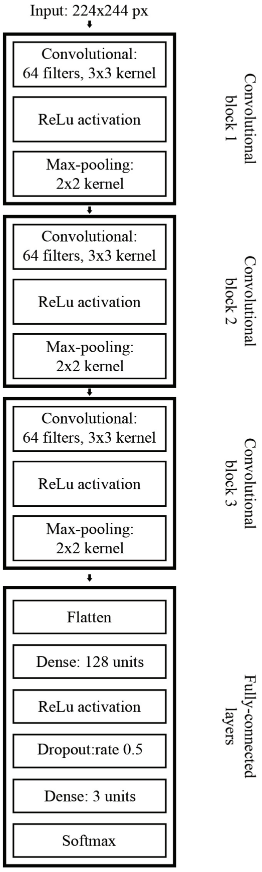
3.6.2 LeNet

The LeNet model is included as a smaller CNN learns less features than the previous methods using the VGG 16 architecture, such that the focus will be on the most significant features found in the data. Thus, a narrow CNN is used with few filters per layer. It is based on findings of LeCun et al. (1998) and their LeNet-5 architecture.

Additions to the architecture from LeCun et al. (1998) are the ReLu activation layers (Nair and Hinton, 2010) and the dropout regularization (Srivastava et al., 2014). These are noted to be efficient methods to increase performance (LeCun et al., 2015). The architecture is shown in table 9. Three blocks and fully-connected layers compose the network. Each block is built up with a convolutional layer, a ReLu activation layer and a max pooling layer. A dropout layer is placed in between the two fully connected layers. The result is passed through a softmax function to produce the final output for each class.

The model is trained from scratch on the dataset. By not depending on trained recognized objects, it should find its own patterns. The weights are initialized by default with the Glorot Uniform initializer (Glorot and Bengio, 2010). To compare the findings with the other models, the same input images are used.

Figure 9: Architecture LeNet



3.7 Implementation

The workflow is written in Python and implemented using several libraries. The VGG architecture and LeNet are implemented using the Keras API (Chollet et al., 2015) running on top of TensorFlow (Abadi et al., 2016). The Gradient Boosting Classifier (GBC) Friedman (2001) and SVM (Cortes, 1995) are implemented by the scikit-learn library (Pedregosa et al., 2011).

Training times of each model deviated quite strongly. Prediction of a single image was around or less than 1s for 1821 images for most experiments. The training and prediction times are reported

Table 11: Training times reported for 29127 images and prediction times for 1821 images. The prediction time to obtain the representation from the VGG 16 architecture is also included.

Model	**Training Time**	**Prediction Time**
VGG-GBC	287m	36s
VGG-P-SVM	35.8m	49s
VGG-FC	2.8m	9s
VGG-RFC	27.6m	9s
LeNet	19.6m	1s
VGG Representation	-	7s

in Table 11. Early stopping decreased training times significantly as the experiments never passed over all epochs. Training and prediction times using scikit-learn were significantly slower.

4 Results

For the classification task accuracy is measured and compared for all experiments. In order to answer which architecture and pipeline works the best, the differences between pipelines will be discussed. During initial experimental tests classification worked noticeably better than a regression as expected. Therefore, solely classifications were performed.

4.1 Results of Experiments Using 3 Bins

Detailed information of each experiment is found in the appendix (table 19). Table 12 presents the accuracy on validation and test data. First the outcome of each experiment is discussed, followed by highlights in different models and general findings.

VGG-GBC: The accuracy of the experiment run on the labeled subset, was below the majority classifier. Using age and gender lowered the accuracy slightly on the validation set by 0.1%. On the complete set, the model showed an improvement over the majority vote and it reached the highest accuracy of all experiments.

VGG-P-SVM: Accuracy on both the validation and test was far below the majority classifier and all other classifiers. Therefore, more tests were not conducted.

VGG-FC: The bottleneck features also showed no lower error rate. The accuracy of 69.8% on validation and 69.4% accuracy on the test set. It is equal to the percentage of the majority class.

VGG-RFC: During the training the accuracy improved over the training set compared to the VGG-FC. It is an indicator that the network is learning or overfitting. However, the learned features had no value as the validation and test accuracy remained exactly the same as the reported

Table 12: Accuracy of the experiments using the target value binned in three categories. Labeled indicates inclusion of age and gender features. Augm indicates use of image augmentation.

Model Name	**Validation**	**Test**	**Size Training Data**
Majority	**0.696**	**0.712**	4890
VGG-GBC (Labeled)	0.694	0.706	4890
VGG-GBC	0.695	0.706	4890
VGG-P-SVM	0.430	0.050	4890
LeNet	**0.696**	**0.712**	4890
Majority	0.698	0.684	29127
VGG-GBC	**0.710**	**0.698**	29127
VGG-FC (Augm)	0.698	0.684	29127
VGG-FC	0.698	0.684	29127
VGG-RFC	0.698	0.684	29127
LeNet (Augm)	0.698	0.684	29127
LeNet	0.698	0.684	29127

accuracies of the bottleneck features. Notably, the training times of this model were the longest.

LeNet: This architecture is the most different from the others as it based on the LeNet-5 architecture opposed to the VGG 16 architecture. For the complete set, the accuracy score on the validation and accuracy without augmentation was slightly higher than the reported 69.8%, thus the performance was alike to the other CNNs and the majority class. On the subset, the accuracy was identical to the majority vote and it was the method with the highest accuracy.

Early Stopping & Checkpoint: Every experiment, except for the two baselines, had the early stopping callback and it was activated after 12 epochs for each experiment. After the reaching maximum accuracy on the validation set, often the accuracy on training data would increase slightly, while this did not reflect on the validation set.

Dataset Size: When using the Gradient Boosting Classifier, training times increased with the addition of data. The accuracy was close for experiments using the GBC on the subset (6113 images) and the complete set (36409 images) in favor of the larger set which performed better. The GBC had the highest accuracy compared to the other architectures on the larger dataset on both the validation and test set. Experiments run on the subset were not higher than the majority vote, while on the larger dataset the GBC managed to achieve the best score and the only score above the majority class.

Gender & Age: The added user attributes was only possible to perform on the small dataset using the baseline model. Compared to the dataset without labels, the performance was slightly

Table 13: MAE of the experiments. Note that the lowest MAE by the LeNet model is because of the guess on one particular class.

	MAE		**Accuracy**		**Size**
Model name	**Validation**	**Test**	**Validation**	**Test**	**Training Data**
VGG-GBC (Labeled)	6.4	6.4	0.051	0.059	4890
VGG-GBC	6.3	6.6	0.052	0.052	4890
VGG-P-SVM (Labeled)	6.7	7.5	0.072	0.049	4890
VGG-P-SVM	6.6	7.0	0.062	0.056	4890
LeNet	4.9	4.9	0.058	0.042	4890
VGG-P-SVM	6.3	6.3	0.077	0.075	29127
VGG-FC	9.5	9.6	0.050	0.050	29127
VGG-RFC	8.6	8.8	0.056	0.052	29127
LeNet	7.2	7.1	0.049	0.052	29127

worse on the validation set and equal on the test set. Due to a insignificant difference of 0.1%, the inclusion of age and gender was not found to be beneficial.

Data Augmentation: The use of data augmentation did not increase the accuracy significantly. During exploratory tests, I noticed the accuracy on training data was at its maximum earlier on when using data augmentation. However, it did not result in better overall performance.

No method significantly outperformed another method. The accuracy is equal or below to the majority class in all experiments, except one. The quickest method was the training of the VGG-FC. Using CNNs in general proved to be quicker. The training time of 4890 images took 47 minutes using VGG-GBC, which was longer than the 27.6 minutes the VGG-RFC needed for 12 epochs over 36409 images.

4.1.1 Results of Experiments Using 20 Bins

The results of these experiments are found in Table 20, alongside the relevant confusion matrices from validation sets. The scores of the experiments are found in Table 13. It should be noted beforehand that the lowest MAE does not show the best working method. Image augmentation was not used in these experiments.

VGG-GBC: The confusion matrix of this showed the classifier made predictions on all classes. The inclusion of the gender and age data did show a small improvement on the validation set, but a small decrease on the test set.

VGG-P-SVM: This model uses the full VGG 16 architecture and the MAE was slightly higher than the VGG-GBC model, while the accuracy was the highest compared to all other models. The labeled data lowered the MAE for both the validation and test set. Using the entire dataset showed similar performance as the subset.

VGG-FC: This architecture was only run on the large dataset. The confusion matrix for both the validation and test set revealed the predictions were on one class. The training of this model stopped early after 16 epochs.

VGG-RFC: Using the weights of the VGG-FC model and retraining the architecture, a little improvement in both the MAE and accuracy. However, the confusion matrix showed another class was predicted. Unlike the VGG-FC, the experiment stopped after 12 epochs.

LeNet: As with the VGG-FC and VGG-RFC models, the confusion matrices show only one class was predicted. The predicted class was the one of the two middle ones, namely class 10. It results in the lowest MAE as this is the most optimal choice to "optimize" for the metric.

Dataset Size: The two dataset sizes were only tested with the VGG-P-SVM experiments. It showed no increase in performance. The training time increased manifold from 3 minute to 30 minutes, which shows the SVM may not be suitable for large datasets.

Gender & Age: The labeled information was included for the experiments using VGG-GBC and VGG-P-SVM. It failed to increase the performance of the model.

No method stood up with equal sizes. The confusion matrices of the VGG-FC, VGG-RFC and LeNet showed the CNNs did not prove to be reliable in this setup. Again, the inclusion of age and gender made not difference. As noted before, the training times were shorter for the CNNs, but that is irrelevant since the performance was not equal or better.

5 Discussion

Different classifications methods were performed on a real-world dataset obtained from Twitter with the adoption of architectures previously demonstrated to work well in similar tasks. The different experiments are designed to counter overfitting with three methods; dropout, early stopping and data augmentation. Additionally, augmentation was performed to artificially increase the number of training instances by rotating them. The resulting experiments were compared with each other and when relevant, they were reflected on a majority vote. Moreover, the inclusion of age and gender information did not show any improvement.

It can be concluded from this preliminary research that profile pictures cannot be used to predict network size with the dataset and experiments I propose. Firstly, images were unfiltered and thus reflected real-world conditions. Previous studies focused on constrained images, which provides assurance of faces or objects. By including any picture, the task was not focused on finding characteristics of a face or object, but patterns relating to network size. CNNs have been proven to have a good performance with real-world images for prediction of other user attributes than network size, but given the models performance no pattern was found.

Secondly, the size of the dataset was small. The problem of overfitting was in a minimal way apparent in one of the VGG-RFC experiment. While there was no performance increase, experiments with extra features could only be run with a subset, limiting the size of training even more.

Despite finding no results indicating that profile pictures can be used for the prediction of network size, the scope of this research did not fully explore more methods. Future research in this area could focus on constrained images. A larger dataset could enable filtering of images and at the same time have more images for training, which is beneficial for training purposes. State-of-the-art deep learning architectures enable the analysis of photos on many aspects, such as aesthetics and object recognition. A thorough analysis is not performed on the content of the images (e.g. group or solo photo, whole body or face) and it might provide even more directions for the most appropriate method.

6 Conclusion

In an attempt to predict the network size based on profile pictures. The proposed methods using a traditional classifier and convolutional neural networks based on either the (trained) VGG 16 or LeNet-5 architecture. The VGG 16 architecture was used to obtain the representation or used to train the bottleneck features and subsequently fine-tune the network. A dataset from Twitter was used with the required information and manually labeled a subset for age and gender. Different experiments were performed to test for the use age and gender, training size and data augmentation. The research was posed as a classification problem. None of the proposed experiments had significant performance over any other. Only one experiment achieved a higher accuracy than a majority vote classifier, thus no evidence was found to support the hypothesis that the latent user attribute network size could be predicted.

References

(2017). *Cluep Inc.* `https://cluep.com/`.

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Becker, B. C. and Ortiz, E. G. (2008). Evaluation of face recognition techniques for application to facebook. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE.

ChaLearn (2017). *ChaLearn Looking at People.* `http://chalearnlap.cvc.uab.es/` [Accessed: 12-may-2017].

Chollet, F. (2016). *Building powerful image classification models using very little data [Blog post].* `https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html`.

Chollet, F. et al. (2015). Keras. `https://github.com/fchollet/keras`.

Cortes, C. (1995). Support-vector network. *Machine learning*, 20:1–25.

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Senior, A., Tucker, P., Yang, K., Le, Q. V., et al. (2012). Large scale distributed deep networks. In *Advances in neural information processing systems*, pages 1223–1231.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE.

Duggan, M., Ellison, N. B., Lampe, C., Lenhart, A., and Madden, M. (2015). *Demographics of Key Social Networking Platforms.* `http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/#twitter`.

Eidinger, E., Enbar, R., and Hassner, T. (2014). Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12):2170–2179.

Ellison, N., Heino, R., and Gibbs, J. (2006). Managing impressions online: Self-presentation processes in the online dating environment. *Journal of Computer-Mediated Communication*, 11(2):415–441.

Emmery, C., Chrupała, G., and Daelemans, W. (In Press). Simple queries as distant labels for predicting gender on twitter. *WNUT 2017.*

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Gallagher, A. C. and Chen, T. (2009). Understanding images of groups of people. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 256–263. IEEE.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256.

Hancock, J. T. and Toma, C. L. (2009). Putting your best face forward: The accuracy of online dating photographs. *Journal of Communication*, 59(2):367–386.

Hill, R. A. and Dunbar, R. I. (2003). Social network size in humans. *Human nature*, 14(1):53–72.

Hu, Y., Manikonda, L., Kambhampati, S., et al. (2014). What we instagram: A first analysis of instagram photo content and user types. In *Icwsm*.

Huang, T. (1996). Computer vision: Evolution and promise.

Kapidzic, S. (2013). Narcissism as a predictor of motivations behind facebook profile picture selection. *Cyberpsychology, Behavior, and Social Networking*, 16(1):14–19.

Karpathy, A. (2014). *What I learned from competing against a ConvNet on ImageNet*. `http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/` [Accessed: 12-may-2017].

KickFactory (2016). *The Average Twitter User Now has 707 Followers*. `https://kickfactory.com/blog/average-twitter-followers-updated-2016/` [Accessed: 12-may-2017].

Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

LeCun, Y., Jackel, L., Bottou, L., Cortes, C., Denker, J. S., Drucker, H., Guyon, I., Muller, U., Sackinger, E., Simard, P., et al. (1995). Learning algorithms for classification: A comparison on handwritten digit recognition. *Neural networks: the statistical mechanics perspective*, 261:276.

Lenhart, A., Purcell, K., Smith, A., and Zickuhr, K. (2010). Social media & mobile internet use among teens and young adults. millennials. *Pew internet & American life project.*

Madden, M. (2012). Privacy management on social media sites. *Pew Internet Report*, pages 1–20.

Malisiewicz, T. (2015). Tombone's computer vision blog.

Mehdizadeh, S. (2010). Self-presentation 2.0: Narcissism and self-esteem on facebook. *Cyberpsychology, behavior, and social networking*, 13(4):357–364.

Morgan, N. and Bourlard, H. (1990). Generalization and parameter estimation in feedforward nets: Some experiments. In *Advances in neural information processing systems*, pages 630–637.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective.*

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Nalepa, J. (2017). When size matters: selection of training sets for support vector machines.

Ong, E. Y., Ang, R. P., Ho, J. C., Lim, J. C., Goh, D. H., Lee, C. S., and Chua, A. Y. (2011). Narcissism, extraversion and adolescents' self-presentation on facebook. *Personality and individual differences*, 50(2):180–185.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Perrin, A. (2015). Social media usage. *Pew Research Center.*

Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. (2010). Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM.

Redi, M., Quercia, D., Graham, L. T., and Gosling, S. D. (2015). Like partying? your face says it all. predicting the ambiance of places with profile pictures. *arXiv preprint arXiv:1505.07522*.

Rothe, R., Timofte, R., and Van Gool, L. (2015). Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 10–15.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

SECURITIES, U. S. and COMMISSION, E. (2014). *FORM 10-Q*. `https://www.sec.gov/Archives/edgar/data/1418091/000156459014003474/twtr-10q_20140630.htm`.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Tong, S. T., Van Der Heide, B., Langwell, L., and Walther, J. B. (2008). Too much of a good thing? the relationship between number of friends and interpersonal impressions on facebook. *Journal of Computer-Mediated Communication*, 13(3):531–549.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., and Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. *arXiv preprint arXiv:1703.03107*.

Walther, J. B. (1992). Interpersonal effects in computer-mediated interaction: A relational perspective. *Communication research*, 19(1):52–90.

Walther, J. B., Van Der Heide, B., Kim, S.-Y., Westerman, D., and Tong, S. T. (2008). The role of friends' appearance and behavior on evaluations of individuals on facebook: Are we known by the company we keep? *Human communication research*, 34(1):28–49.

Appendices

A Description Dataset

Field	**Description**	**Data Type**	**Obtained**
Profile Image	The profile image of a person. The images can be anything; from photos to cartoons.	RGB Image	Twitter
Face	Checkbox to indicate whether the face of the person is shown on the picture, even if multiple people are in the picture.	Binary true/false	Annotated
Bot	Checkbox to indicate whether the profile is a bot of page.	Binary true/false	Annotated
Gender	The assumed gender the person has.	Letter indicators for male, female, other	Annotated
Age	The assumed age the person has. While it is often hard to determine exact age, ranges of ages are not used.	Integer - continuous	Annotated
Friends Count	Twitters field to indicate the amount of people the person is following.	Integer - continuous	Twitter
Followers Count	Twitters field to indicate the amount of followers the person has.	Integer - continuous	Twitter
Signal	A description of the elements that determined the annotation. Possibilities are: image, name, description, handle, tweets, URL, other. These refer to parts of a Twitter profile.	Hot encoded words	Annotated

B Description Annotation Tool

Field	**Explanation**	**Fill Options**
ID	User ID.	-
Image	Small representation of the profile image.	-
Media feed (url)	Media feed.	-
Tweet feed (url)	Tweet feed (no retweets).	-
Handle	Twitter handle, also link to profile.	-
Name	Full name of the person.	-
Face	Check if the picture shows the face of the person that owns the profile - even if there are multiple people in the picture (not an entire crowd!). Do not check if the image is e.g. a celebrity, a cartoon picture, or anything else that is not a face.	Checked / Unchecked.
Bot	Check if this person might be a bot or page (i.e. not an actual person).	Checked / Unchecked.
Description	Person's profile description.	-
Gender	The assumed gender the person has	m, f, o, -
Age	The age you assume the person has. Left blank if not a person.	integer
Signal	Short description of the elements that determined the annotation. If e.g. the image and name of the person were indicative of the gender and age, signal is 'image name'.	image name description handle tweets url other

C Pipelines

Table 14: Pipeline for obtaining the VGG 16 Representation

Step 1: Input	Image
Step 2: Preprocess	Subtract mean

Table 15: Pipelines Baseline

	1	2	3
Step 1: Input	VGG Representation 4890 images	VGG Representation 4890 images + gender & age	VGG Representation 29127 images
Step 2: Preprocessing	Gradient Boosting Classifier (depth: 3, estimators: 200)		

Table 16: Pipelines VGG-FC

	1	2
Step 1: Input	29127 images for training	
Step 2: Preprocessing	Subtract mean VGG 16	
Step 3: Image Augmentation	None	Rotation & zoom, 2 rounds
Step 4: Feature Extraction	Trained VGG 16 architecture	
Step 5: Prediction	Softmax	

Table 17: Pipeline VGG-RFC

Step 1: Input	29127 images for training
Step 2: Preprocessing	Subtract mean VGG 16
Image Augmentation	None
Step 3: Feature Extraction	Trained VGG 16 architecture
Step 4: Prediction	Softmax

Table 18: Pipelines for LeNet

	1	2	3
Step 1: Input	4890 images for training	29127 images for training	
Step 2: Preprocessing	Subtract mean		
Step 3: Image Augmentation	None		Rotation & zoom, 2 rounds
Step 4: Feature Extraction	3 Convolutional blocks		
Step 5: Prediction	Softmax		

D Results

Table 19: Results from all experiments using 3 classes.

Name	VGG-GBC			VGG-FC		VGG-RFC	LeNet			VGG Representation
Input	Image, labels	Image								
Preprocess	Subtract mean									
Augmentation				Rotation, Zoom		Rotation, Zoom				
Augmentation Rounds			2			2				
Size Training Data	4890	4890	29127	29127	29127	29127	29127	29127	4890	
Size Validation Data	917	917	5461	5461	5461	5461	5461	5461	917	
Size Test Data	306	306	1821	1821	1821	1821	1821	1821	306	
Epoch Time	-	-	-	29s	14s	138s	356s		51s	-
Epochs	-	-	-	12 (50)	12 (50)	12 (50)	12 (50)	12 (50)	12 (50)	-
Training Time	45m	47m	287m							
Early Stop	-	-	-	-	Yes	Yes	Yes		Yes	-
Prediction Time	-	-	-	-	1s/1821	9s/1821	1s/1821		1s/1821	143s/36409
Accuracy Validation	0.6934	0.6945	0.710	0.698	0.698	0.6978	0.698	0.696	-	
Accuracy Test	0.703	0.706	0.698	0.684	0.684	0.684	0.684	0.712	-	

Table 20: Results for all experiments using 20 even classes.

Model Name	VGG-GBC		VGG-P-SVM			VGG-FC	VGG-RFC	LeNet	
Input	Image, labels	Image	Image, labels	Image					
Preprocessing	Subtract mean								
Size Training Data	4890	4890	4890	4890	29127	29127	29127	4890	29127
Size Validation Data	917	917	917	917	5461	5461	5461	917	5461
Size Test Data	306	306	306	306	1821	1821	1821	306	1821
Epoch Time						14s	180s	11s	380s
Epochs						16 (50)	12 (50)	16 (50)	12 (50)
Training Time	5.5m	5.4m	41s	48s	31.8m				
Early Stop						Yes	Yes	Yes	Yes
Prediction Time	5s/1821	5s/1821	7s/1821	7.2s/1821	49s/1821	9s/1821	9s/1821	1s/1821	1s/1821
Accuracy Validation	0.051	0.052	0.072	0.062	0.077	0.050	0.056	0.058	0.049
Accuracy Test	0.059	0.052	0.049	0.056	0.075	0.050	0.052	0.042	0.052
MAE Validation	6.4	6.3	6.7	6.6	6.3	9.5	8.6	4.9	7.2
MAE Test	6.4	6.6	7.5	7.0	6.3	9.6	8.8	4.9	7.1
Predict 1 Class	-	-	-	-	-	Yes	Yes	Yes	Yes

E Confusion Matrices

Table 21: Confusion matrix - VGG-P-SVM - validation - subset

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	4	15	7	0	0	1	4	0	2	1	1	0	2	0	6	0	0	0	1	0
2	5	11	3	0	0	2	3	0	1	1	2	0	1	2	6	0	0	0	0	2
3	5	19	10	1	4	1	1	1	0	2	0	3	5	0	3	1	1	0	3	1
4	3	10	3	1	1	4	1	0	3	4	1	2	0	1	0	0	1	0	0	3
5	5	11	7	0	2	5	1	0	2	0	0	2	2	1	6	0	0	0	2	1
6	3	9	9	0	2	4	2	0	0	2	2	2	3	1	4	1	1	0	1	0
7	0	9	2	1	0	7	4	2	1	1	2	3	1	4	3	1	0	2	0	0
8	6	12	9	0	0	4	1	0	0	1	0	3	3	3	4	0	0	0	1	0
9	5	9	5	0	4	10	4	0	2	0	3	4	2	1	0	0	1	1	0	1
10	1	11	5	1	0	3	4	2	2	2	4	0	3	3	5	0	3	1	2	1
11	2	11	3	0	3	4	2	0	4	1	1	3	2	0	3	2	0	0	1	1
12	1	12	1	2	3	1	3	0	2	2	1	6	4	2	6	1	0	1	0	1
13	3	5	1	1	3	5	0	0	3	2	3	1	2	0	5	0	2	1	0	0
14	7	17	5	0	2	3	2	0	3	1	2	2	4	6	2	0	1	1	0	2
15	2	8	3	0	0	6	3	0	2	5	1	0	4	0	1	0	2	2	1	0
16	1	8	3	1	0	4	3	0	0	3	1	2	4	1	4	0	0	1	0	0
17	3	11	3	3	2	4	3	0	2	3	2	2	2	2	4	0	1	0	0	0
18	5	6	2	2	1	5	2	0	2	4	2	3	3	2	1	0	0	0	1	2
19	4	7	3	1	1	7	3	0	2	5	1	0	4	1	4	0	0	1	0	0
20	3	13	5	0	0	4	0	2	1	6	0	3	1	3	6	1	0	0	0	0

Table 22: Confusion matrix - VGG-GBC - validation - image, labels - subset

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	4	2	6	1	0	1	0	2	2	3	2	1	3	2	5	2	3	2	2	1
2	5	3	2	1	1	2	2	1	3	1	3	1	2	1	2	2	1	2	2	2
3	8	5	4	6	4	2	1	0	1	1	3	8	3	1	0	2	3	1	5	3
4	2	5	2	3	2	1	2	2	1	4	0	3	0	2	0	2	1	2	2	2
5	2	1	1	3	7	2	3	1	3	2	5	1	3	1	1	2	2	2	2	3
6	3	1	5	4	2	0	4	0	0	4	1	6	4	2	1	2	1	4	2	0
7	5	4	1	2	0	1	1	5	0	2	5	4	4	0	3	2	1	2	0	1
8	6	5	1	4	0	1	3	6	0	5	1	3	4	1	0	0	4	1	2	0
9	4	3	1	4	6	2	3	4	0	2	1	0	2	1	4	5	2	2	3	3
10	3	5	2	0	1	1	5	4	1	2	3	4	2	4	4	2	3	4	1	2
11	1	4	2	0	2	2	0	5	1	2	1	3	3	2	4	1	3	1	5	1
12	0	2	0	0	6	1	4	2	1	4	4	3	3	5	3	4	1	1	3	2
13	1	1	0	1	6	2	3	0	1	3	3	1	3	0	5	1	2	1	3	0
14	7	3	3	3	2	4	3	1	1	2	5	5	3	1	3	4	2	2	3	3
15	0	1	2	4	2	1	3	2	0	4	2	4	4	3	1	2	3	1	1	0
16	1	4	0	1	3	1	1	4	1	5	1	0	6	2	1	0	2	1	1	1
17	1	2	3	3	3	1	3	1	2	4	2	3	2	1	4	3	4	1	1	3
18	2	1	0	4	2	2	4	3	1	3	2	2	4	2	2	2	3	2	1	1
19	2	3	2	2	3	3	5	2	1	1	2	3	3	4	2	2	2	1	0	1
20	2	4	4	1	1	2	0	2	2	4	2	1	6	5	1	2	1	3	3	2

Table 23: Confusion Matrix - VGG-P-SVM - validation - image, labels - subset

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	10	1	1	3	1	3	0	0	0	7	2	0	1	0	3	1	1	0	10	0
2	12	1	1	2	0	2	0	0	0	6	1	0	0	0	1	1	0	0	11	1
3	9	2	2	5	0	2	1	3	0	5	1	3	0	1	7	5	0	0	15	0
4	4	0	1	7	0	1	0	1	0	5	0	0	0	1	4	1	0	0	12	1
5	3	0	0	6	0	4	0	0	0	6	1	1	0	0	6	2	0	0	18	0
6	3	0	0	5	0	6	0	0	0	8	1	0	1	0	7	1	3	0	11	0
7	1	0	1	2	0	1	0	0	0	9	0	3	0	0	5	3	2	2	14	0
8	6	0	1	7	1	2	0	1	0	7	0	1	0	2	4	0	0	1	14	0
9	5	0	1	4	0	3	0	0	0	10	2	0	0	1	4	4	0	0	18	0
10	4	0	1	5	0	1	2	2	0	6	1	0	1	1	4	1	0	0	22	2
11	4	0	0	4	0	2	0	1	0	9	0	1	0	0	5	1	2	0	14	0
12	5	0	0	2	1	1	0	0	0	13	1	2	0	1	3	1	1	0	18	0
13	2	0	1	1	0	4	0	0	0	10	2	1	0	0	3	0	0	0	13	0
14	5	0	0	10	0	2	0	1	0	10	2	0	1	2	5	3	3	0	15	1
15	0	1	0	4	0	1	0	0	0	11	1	0	0	0	5	1	0	1	15	0
16	2	1	0	4	0	2	0	1	0	8	0	0	1	0	0	2	1	0	14	0
17	2	0	1	3	0	2	1	0	0	6	2	0	0	1	7	2	1	0	19	0
18	5	0	1	3	0	1	1	0	1	8	0	0	0	0	0	3	1	1	18	0
19	6	0	2	5	0	0	0	1	0	7	0	0	0	0	1	0	0	1	20	1
20	8	0	1	5	0	1	0	1	1	7	0	2	0	0	3	1	1	0	17	0

Table 24: Confusion matrix - VGG-P-SVM - validation - entire dataset

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	52	11	24	29	4	19	12	4	5	6	23	9	5	11	9	27	2	15	10	19
2	35	23	12	19	2	12	18	5	11	7	24	9	5	8	9	22	5	8	6	18
3	32	12	22	40	2	17	21	3	10	3	29	7	11	7	8	18	4	18	8	12
4	32	12	14	31	3	13	15	2	7	5	22	8	12	15	9	19	3	13	6	16
5	31	4	11	32	8	14	19	2	6	1	28	9	14	6	7	18	0	16	6	15
6	23	10	12	22	2	13	22	5	12	3	34	5	11	7	6	24	4	11	12	12
7	14	12	17	30	0	22	30	6	7	7	31	8	9	16	12	24	3	16	10	22
8	30	6	9	15	3	17	19	10	12	6	31	9	14	11	15	17	7	17	6	10
9	28	4	12	18	1	26	23	4	19	10	28	9	13	10	14	22	7	18	7	16
10	24	3	14	16	2	21	18	4	12	4	33	13	13	8	11	23	3	19	7	15
11	16	7	10	18	3	15	23	6	8	6	28	14	16	7	14	20	7	23	6	14
12	24	8	16	12	2	18	18	2	16	6	29	7	14	15	8	27	6	12	11	16
13	23	11	14	35	3	13	23	6	10	3	37	4	13	11	12	26	3	22	7	13
14	26	6	16	21	0	18	28	4	7	4	28	7	18	13	17	29	5	24	3	12
15	13	14	12	26	4	16	20	5	5	6	25	13	16	8	14	36	3	24	6	21
16	21	6	13	23	2	17	15	8	6	5	32	9	10	10	9	38	8	15	7	20
17	22	9	14	17	2	17	23	7	10	4	23	17	19	11	12	22	7	22	9	14
18	36	5	8	20	3	20	22	6	5	6	33	8	18	7	12	22	4	26	11	17
19	18	8	6	27	5	21	23	4	5	6	22	5	11	6	5	27	2	24	10	17
20	15	12	14	28	2	20	20	4	2	3	22	11	9	10	6	18	2	11	10	52