The Thirty-Third International
FLAIRS Conference (FLAIRS-33)

Experimentation on Hand Drawn Sketches by Children to Classify Draw-a-Person Test Images in Psychology

Ochilbek Rakhmanov,[1] Nwojo Nnanna Agwu,[2] Steve Adeshina[3]
Computer Science Department, Nile University of Nigeria, Abuja, Nigeria[1,2,3]
{ochilbek.rakhmanov, nagwu, steve.adeshina} @nileuniversity.edu.ng

Abstract

Classification of hand drawn sketches with respect to content quality is extremely challenging task, comparing to usual image classification methods. In brief, we need to train computational device to able to classify the images of the same object into different classes with respect their content quality. In this paper we tested several methods of image classification, using machine learning and computer vision algorithms, to classify Draw-a-Person test images sketched by primary school students in Nigeria, aged 4 to 11 years. We collected 1000 original sketches and manually classified them (using guidelines from existing literature) according to the ages (8 classes) before testing this dataset on a computational device. The highest accuracy achieved in this experiment was 62%. We achieved this result with novel method, where we used Bag of Visual Words and K-means algorithm to count keypoints on each sketch. We strongly believe that this challenging task needs further research to improve classification accuracy, we, therefore, release the complete dataset of sketches to the community.

Introduction

The new era in machine learning and image classification brought some new ideas to classify not only real-life images, but also those produced by human through sketching and drawing. Owing to the new opportunities brought about by modern technology, we can now sketch not only on paper but also on touchscreens, tablets, phones and some other devices.

Recognizing free-hand sketches is a very difficult task compared to real life images. This is due to several reasons (see Figure 1):

- Sketches can be very abstract and deformed, but still represent the same object.
- The style and lines of the figure will definitely change with respect to the person drawing it. However, while some people may draw all features of the object, others may miss some features.
- The sketches lack color; mostly black and white.

After their successful experimentations on MNIST hand drawn digits dataset, LeCun *et al* (2010), Eitz *et al* released an open source dataset of 20,000 hand drawn sketches in 2012 (Eitz et al., 2012), consisting of 250 different classes with 80 samples in each. This dataset influenced many researchers to conduct experiments to improve classification accuracy from the initial 55% to 77%, with the latest state-of-art method proposed by Yu *et al* (2017).

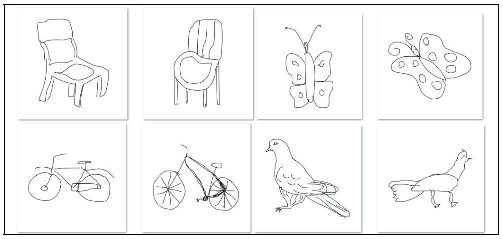

Figure 1: Sample pictures from Eitz *et al* dataset
(chair, butterfly, bicycle and pigeon) (Eitz et al., 2012).

Hand drawn sketch recognition also led to many other combinations. Klare *et al* (2010) achieved good results in matching sketches with photos, using the feature-based approach. Recent researches on the combination of sketches with transfer learning are also important to note. In one of such studies Chugh *et al* (2017) used sketches for face recognition through transfer learning.

It appears likely that sketch recognition and its combinations will continue to flourish in the next few years, spreading to other fields like education, psychology, crime investigations, and so on. In this paper, we conducted one such study; we investigated the possibility of combining sketch recognition with the analysis of cognitive tests. There are many existing cognitive tests being used in education and psychology, but some of them are main actors. They include Goodenough's Draw-a-Person test, Clock drawing, Draw-a-Tree and Rey-Osterrieth's complex figure, among others.

Our goal was to analyze the Draw-a-Person test.

Draw-a-Person test

First conceived by Dr. Florence Goodenough in 1926 (Goodenough, 1926), the Draw-a-Person test is a skill test that is designed to measure a child's mental age through a figure drawing task. It estimates the progress of learning visual, cognitive, and motor skills by having the candidate draw a human figure, scoring the drawing for presence and quality of figure features, and comparing the score to children's typical rate of acquisition of figure features (Goodenough, 1926). Throughout the years, this test underwent many discussions, whereby the researchers tested the validity and reliability of the test, usually resulting in supportive conclusions. The instrument is among the top 10 tools used by practitioners, according to Yama *et al* (1990). It is widely used in early childhood education; primary school counselors can use it to monitor children's mental development. In Psychology, for instance, it has been used during comparison between healthy patients and those with mental disorders. There are many supportive researches and case studies for these aforementioned fields, summarized by Naglieri *et al* in 2004 (Naglieri et al., 2004). Due to the wide adoption and use of DAPT, there is a need to conduct some experiments on DAPT sketches, to see if we can develop a model to classify images automatically. This was the main reason why we conducted a number of image classification tests (using machine learning and computer vision algorithms) to determine whether we would be able to develop such model.

How the DAPT scored?

This section will provide a brief explanation of the DAPT picture (as it was proposed by (Goodenough, 1926)). In the DAPT image evaluation, we counted the number of features, like eyes, nose, mouth, hair, hands, shoulder, fingers, etc. Yet, some more scores came from the comparison of the geometrical positioning of the features. For instance, a child can draw one leg short and other long or provide a sketch of eyes whose proportion on the face is abnormal. The sum of these scores results in a total score (minimum 0, maximum 51 points). We look up for corresponding mental age from score ~ mental age scale proposed by Goodenough (Goodenough, 1926). This mental age is compared to biological age, and if the difference is high, then scholar and practitioners advice that the child needs special assistance or training on recognizing the functions of the objects surrounding him or her in his or her daily life (Goodenough, 1926). Figure 2 shows some pictures from our dataset and their corresponding mental ages.

Major differences between datasets

At this point, it is very important to underline the difference between the dataset of Eitz *et al* and our DAPT dataset.

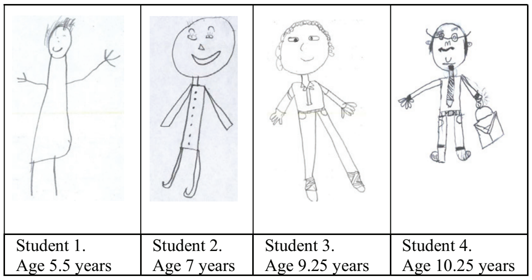

Figure 2: Sample sketches of the students.

Figure 1 clearly shows that even though the shapes are deformed, there is still space for computer algorithms to identify differences between objects to classify them. However, this is not always the same for DAPT dataset. Figure 3 demonstrates some human figures which belong to the same class, even though they are drawn in very different ways.

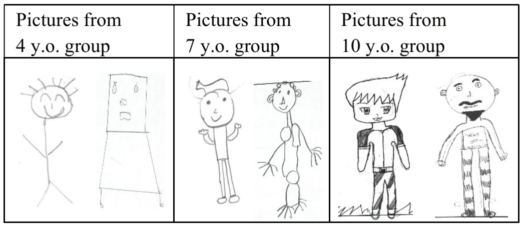

Figure 3: Different human shapes from same group.

Thus, the biggest challenge in this dataset as it can be observed from Figure 3 is that sketches may appear in very different shapes. Unlike in most common image classification tasks, we are not actually looking for the difference between the images, but for the common features and how frequently they appear. The exploration of any possible solution for this challenge is the main goal of this research.

Another important difference is that the Eitz *et al* dataset was collected through touchscreen devices and mostly participants were adults. All pictures are clear and sharp. However, it was impossible to do so in our case as we had kids aged 4-7 years who were unfamiliar with touchscreens or other electronic devices. Thus, our dataset consists of sketches drawn on a plain sheet. Each student tried to draw his or her best picture; so, there are many pictures where students tried to erase previous sketches and replace them with new ones. They also tried to draw extra elements and the lines in their drawings were clashing or not straight. All this created significant noise on images, thus requiring extra image processing job. Nonetheless, for any type of image processing step, there is always a danger of losing some features in the picture. This affected our classification accuracy during the experiment.

Purpose of the study

The objectives of this study are as follows:

1. To collect new dataset of DAPT and conduct experiments on classification of this dataset with existing state-of-art methods for hand drawn image classification.
2. To test some computer vision algorithms (feature extraction, identifying descriptors, etc.) on DAPT dataset, to determine how well they improve classification accuracy.
3. To conduct and present comparative analysis on classification methods for future researches on DAPT or any similar type of dataset.

Related literature and guideline for methodology

During the experimentation, we followed literature on the classification of Eitz *et al* dataset. We tested all proposed state-of-art methods to classify this dataset as it is the dataset most similar to ours.

In 2012, Eitz *et al* released a dataset of hand drawn images with 20,000 samples; 250 different objects with 80 samples in each (Eitz et al., 2012). They used Support Vector Machines (SVM) classifier, after calculation of Histogram of Oriented Gradients (HOG) for each picture. They achieved a 56% classification accuracy.

In the next 3 years, some significant studies were conducted by the group of researchers led by Li *et al* (2015), where they used some ensemble learning and multi-kernel image processing in their researches, subsequently increasing classification accuracy to 68%. The Bag of Visual Words concept was successfully used during these researches to improve the accuracy.

When the CNN (Convolutional Neural Network) became one of the most powerful image classifiers, Yu *et al* proposed a CNN structure for hand drawn image classifications in 2015 (Yu et al., 2015). This proposed CNN structure reached 74% of classification accuracy. Two years later, the same group proposed an updated version for their CNN structure (Yu et al., 2017), with some additions, and managed to increase accuracy by 3%. Today, this method remains the state-of-art method for the classification of hand drawn images.

During our methodology, the literature listed above served as a guideline for us.

Theoretical framework and instruments

In this section, we briefly introduce the algorithms and literature we used during the experiment.

• *Computer vision and image processing*. Apart from common image processing functions like resizing, threshold (to reduce noise), binary inverse and dilation, we used several computer vision concepts during our experiment. HOG - Histogram of Oriented Gradients (Dalal & Triggs, 2005) was successfully used by Eitz *et al* (Eitz et al., 2012). The HOG descriptor technique counts occurrences of gradient orientation in localized portions of an image detection window, or region of interest. Another important algorithm for image feature detection is BoVW (Yang et al., 2007). The general idea of BoVW is to represent an image as a set of features. These features consist of key-points and descriptors. Keypoints are the points in an image; so, no matter how much the image is rotated, shrunk, or expanded, its keypoints will always be the same. The descriptor is the description of the keypoint. We used the keypoints and descriptors to construct vocabularies and represented each image as a frequency histogram of features that are in the image. From the frequency histogram, we can find other similar images or predict the category of the image . To calculate keypoints and descriptors, we used ORB (Rublee et al., 2011).

• *Support Vector Machines.* One of most used machine learning algorithms in such cases is Support Vector Machines (SVM). In short, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. During the experiment, we also used the Gaussian kernel to determine whether it results in better accuracy. During the experiment, we tested the dataset with both linear and Gaussian kernel (Eichhorn & Chapelle, 2004). During training with SVM we used 10-fold cross validation to avoid overfitting.

• *Neural networks.* Artificial Neural Network (ANN) consists of simple processing elements that are interconnected via weights. The network is first trained using an appropriate learning algorithm for the estimation of interconnected weights. Once the network is trained, unknown test signals can be classified. The class of neural networks used most often for classification tasks is the multilayer perceptron network. A Convolutional Neural Network (CNN) is comprised of one or more convolutional layers (often with a subsampling step) and then followed by ANN (Gulli & Pal, 2017).

In our experiment, all hidden layers of network are configured with Rectifier Linear Units (ReLUs), which have proven to be faster than their equivalent tanh or sigmoid units. Dropout was performed after some layers to avoid codependences between different nodes. Cross entropy was used as loss function and weights were optimized using both Stochastic gradient decent (SGD) or Adam optimizer, depending on accuracy performance. Unlike other layers, final layer used Softmax function for final probability prediction. To avoid over fitting of the model, we used data augmentation techniques and tried to determine if batch normalization will improve the accuracy (Gulli & Pal, 2017).

• *Programming instruments*. We used Python as it is one of most popular programming languages for image classification. Image processing and computer vision algorithms were implemented by using open source library OpenCV (Gary & Adrian, 2008). While Scikit-Learn was our main library during machine learning training (Pedregosa et al., 2011), Keras and Tensorflow was used during ANN and CNN training (Gulli & Pal, 2017).

Data collection and dataset formation

Data collection

• *Ethics and regulations*. We followed all ethics and regulations during data collection. All parents were properly informed of test regulations and their consents were obtained. Students were guided by their teachers during process. It was an extra-curricular activity for students with minimum level of stress.

• *Participants*. Students from private educational institution were selected. They were aged from 4 to 11 years, from Nursery 1-2 and Primary 1-2-3-4 grades. More than 1000 pictures were collected, and some meaningless sketches were eliminated, resulting with 1000 pictures remaining.

• *Process*. Students were given a plain sheet and told to sketch a man. No further assistance was given as the sketch required the authentic work of the child. About 10-15 minutes were given to the children to finish the job. After the sketching process was completed, all pictures were collected and passed for further steps.

• *Manual classification*. Three trained persons (school counselor and PhD students) followed Goodenough's 51-point scoring criteria to calculate the score of every picture (Goodenough, 1926). Once scoring was done, we classified them according to mental ages, from 4 to 11 years, using a total of 8 classes.

Dataset formation

To make our work diverse and find best possible option, we created several versions of the dataset.

a) *All_8*: this is primer dataset, all sketches divided to 8 classes, according to the ages, from 4 to 11 years. Pictures were cropped and resized to 120x240 pixels.

b) *Inv_8*: we cropped pictures, inversed color to binary (background is black, sketch is white), and dilated it to not lose important features during resizing. Another purpose for this dataset was to reduce the calculation cost during machine learning training.

c) *Double*: in this dataset, we merged the age groups; 4 with 5, 6 with 7, 8 with 9 and 10 with 11. This resulted in a total of 4 classes.

d) *Double_inv*: Inverse 8 is grouped into 4 classes, just like Double.

e) *Reference*: in this small dataset, we joined only 10 and 11. This dataset was used during the BoVW experiment to propose a novel method.

The formation of different datasets is not a desirable condition and our expectation is to get the best possible result on All_8 dataset. However, during the experiment we discovered that the classification of All_8 is a very difficult task, which forced us to look for different methods of classification.

As a rule of thumb, we used 75% of the data for training and the remaining 25% for testing, 750 images versus 250.

Methodology

Testing without feature extractors

To determine how well the computer vision feature extractors were performing, we first conducted a classification test only using pixel values of the pictures. Even though we did not expect that accuracy would be high, we expected it, at least, to give us insight into how well we were performing using image feature extractor methods. A picture with a size of 100 x 100 pixels was flattened to vector [10000,1] and fed as input for classification algorithm.

Classifier	Parameters	Dataset	Acc.
SVM	kernel=linear	Inv _8	18%
ANN	1,2 or 3 HLs	Inv _8	17%

Table 1: Test data prediction accuracy with only pixel value.

Testing with HOG

Firstly, we used HOG to extract features from pictures. It is well known that HOG is a very a successful tool for human detection on images; so, we believe that it would be useful for our study. We extracted HOG features from all pictures in All_8 dataset and used SVM and ANN to classify them. Grid search was conducted during the search of best parameters for SVM and several ANN structures (1, 2 and 3 hidden layers (HL)) were tested. Table 2 is a summary for classification accuracies using HOG. We tested many options. The best results are presented on Table 2.

Classifier	Parameters	Dataset	Acc.
SVM	Kernel=linear	Inv _8	22%
SVM	Kernel=linear	Double_inv	48%
ANN	1,2 or 3 HLs	All_8, Inv_8	17%
ANN	1,2 or 3 HLs	Double	32%

Table 2: Test data prediction accuracy with HOG.

During the training of ANN, Adam optimizer was inefficient as it suffered from local minima, training accuracy was constant after some trainings. There can be several reasons for that, such data inconsistency, curse of dimensionality or other reasons, and this phenomenon can be investigated in future researches. The dimensions of the input vector to ANN after HOG operation was very large, [453600x1], causing network to suffer the curse of dimensionality. Double data performed better, and the highest accuracy was obtained with SVM (linear kernel) trained on Double_inv dataset.

Testing with BoVW

The next step was to use BoVW for feature extraction and classification. The biggest challenge in this task was that the dimensions of input vectors were not same as descriptors may vary too much if we try to show them in matrix form.

Therefore, we tried two different ways of using BoVW to classify our dataset: manual and formal methods.

• *Manual method.* We first used the Reference dataset to calculate 1000 key-points and their descriptors from each picture. Figure 4 displays some of selected images and their respective key-points (1000) with red circles.

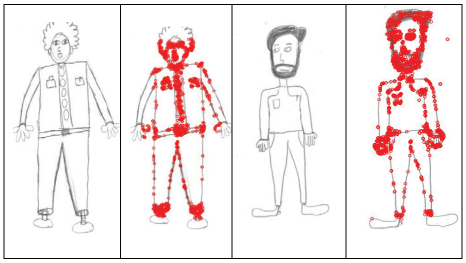

Figure 4: Sample pictures and their key-points for BoVW.

Next, we used K-means to find the 400 most likely descriptors to appear in every picture. We formed a vocabulary of visual words from these 400 descriptors and used matching method to determine if these vocabulary elements are appearing in other classes of All_8, and how frequently they appeared. If a descriptor was appearing on picture (probability of appearing is bigger than zero) we counted it as 1 point. We calculated the total points for all pictures. Table 3 is summary of the mean for every class and standard deviation (rounded to integer).

Class	Mean	Std.		Class	Mean	Std.
4	113	37		8	158	33
5	135	37		9	170	30
6	141	36		10	183	32
7	152	34		11	192	30

Table 3: Appearance of the vocabulary

Table 3 clearly shows that there is difference between the mean of descriptor appearance per class. However, the problem is that standard deviation is too high, which was the main reason of low prediction accuracy. Consequently, when we tried to classify pictures according to mean/std intervals, our classification accuracy was as low as 20%. To extract more info from Table 3, we tested the so called 'maximum class' concept. We put an upper limit for every class and checked if the number of descriptors was not exceeding the limit. In other words, we predicted the maximum class value of tested picture. We set limits as 127-138-145-155-163-177-188-250. Here, we got some promising results, 54% of prediction accuracy. This accuracy reached 62% when we tested the Double dataset with this method. Table 4 is summary of 'maximum class' prediction.

All_8									Double			
Class	*N*	*Upper bound*	*Acc.*		*Class*	*N*	*Upper bound*	*Acc.*	*Class*	*N*	*Upper boun*	*Acc.*
4	142	127	0.63		8	125	163	0.58	45	273	130	0.6
5	131	138	0.52		9	124	177	0.51	67	312	148	0.56
6	152	145	0.52		10	79	188	0.47	89	249	168	0.6
7	160	155	0.52		11	38	250	0.63	1011	117	168+	0.7

Table 4: Manual method accuracies

• *Formal method.* After testing with the manual method, we tried to use the BoVW descriptors with classification algorithms. However, as previously stated, descriptors are in different shapes. To solve this challenge, just like in manual method, we used the vocabulary method to standardize the input shape, through histogram calculation, which resulted in a very low accuracy of below 20% (with SVM). The main reason for this was that standardization operation was seriously affecting the data, thus misleading the classifier.

Testing with CNN

We already stated that we used several types of CNN structures and optimization algorithms. On this dataset, Adam optimizer was suffering local minimum; thus, after some training, validation accuracy was stuck at local minimum. Hence, SGD was the preferred optimizer during training. Table 5 presents all the different options and parameters we have tested to reach highest accuracy.

Parameters	Tested values
Convolutional layers	1,2 or 3
Fully connected layers	1,2 or 3
Batch normalization	True, False
Data Augmentation	True, False
No of filters	16,32,64
Strides	(1,1), (2,2)
Optimizer	SGD, Adam
Dropout	True, False

Table 5: All tested parameters during CNN training

During training, our criterion to stop training was that loss value should be less than 0.01 with training accuracy bigger than 95%. Nonetheless, in most cases this led to increment of validation loses. Thus, our goal was to obtain an optimal network. We have arrived at the network structure presented in Figure 5, which reached the highest prediction accuracy, while keeping validation loss as low as possible. Batch normalization was not used, while data augmentation really helped the network. We applied drop-out only once, after Hidden layer-1, and optimization function was SGD.

As was expected, we had to do a tradeoff between prediction accuracy and validation loss. Since our testing pictures vary in shapes, validation loss never managed to reach the lowest value, usually stopping decrement after several epochs of training, but this does not necessarily lead us to overfit the model. We tested both All_8 and Double dataset with CNN structure presented in Figure 5, in sequential order. We reached 32.33% of prediction accuracy on All_8 and 52.22% on Double dataset.

Conv. layer	Input layer. 240x120x3.
	Layer 1. 32 filters. 3x3 kernel. 2x2 stride.
	Max-pooling 2x2. 2x2 stride.
	Layer 2. 32 filters. 3x3 kernel. 1x1 stride.
	Max-pooling. 2x2 stride. 1x1 stride.
FC layer	Flatten layer.
	Hidden layer 1. 1024.
	Hidden layer 2. 128.
	Output layer. 8.

Figure 5: CNN structure with highest accuracy.

Results and discussion

In this paper our main objective was to conduct some experiments on DAPT dataset to determine how it would perform with some state-of-art classification methods. Throughout the experimentation, we discovered that this kind of images should use some specific way of approach, rather than common classification methods. We observed that the All_8 dataset is really difficult to classify with high accuracy, while joining 2 classes like in Double, can minimize this burden. However, our primary goal should be to develop a method to classify All_8 dataset.

We introduced a novel Manual BoVW method, which seems to be simple but achieved good results with this type of dataset. We can underline that further experimentation with DAPT dataset should include experimentations with BoVW. Prediction accuracy might improve if future studies can lower the standard deviation for BoVW method. As we have tested a unique dataset, which was not tested before for classification by computational device, we believe that our resulted prediction accuracy, %62, is worthy finding and can be improved with further studies.

Unlike for Yu *et al* (2017), by using CNN, we could not achieve such promising results they have achieved on Eitz *et al* dataset. This requires further experimentations with CNN to improve prediction accuracy. We conclude our paper with Table 6, the summary of all experiments we conducted, reference for further studies.

Method	All_8/ Inv_8	Double/Double_Inv
HOG+SVM	22%	48%
HOG+ANN	17%	32%
BoVW+Max.Class	54%	62%
CNN	32%	52%

Table 6: Summary of all classification experiments

References

Chugh, T., Singh, M., Nagpal, S., Singh, R., & Vatsa, M. (2017). Transfer learning based evolutionary algorithm for composite face sketch recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 117–125.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection.

Eichhorn, J., & Chapelle, O. (2004). Object categorization with SVM: Kernels for local features.

Eitz, M., Hays, J., & Alexa, M. (2012). How do humans sketch objects? ACM Trans. Graph., 31(4), 44–1.

Gary, B., & Adrian, K. (2008). Learning OpenCV: Computer vision with the OpenCV library. In O'Reilly Media, Inc.

Goodenough, F. L. (1926). Measurement of intelligence by drawings.

Gulli, A., & Pal, S. (2017). Deep Learning with Keras. Packt Publishing Ltd.

Klare, B., & Jain, A. K. (2010). Sketch-to-photo matching: A feature-based approach. Biometric Technology for Human Identification VII, 7667, 766702.

LeCun, Y., Cortes, C., & Burges, C. J. (2010). Mnist handwritten digit database. AT&T Labs.

Li, Y., Hospedales, T. M., Song, Y.-Z., & Gong, S. (2015). Free-hand sketch recognition by multi-kernel feature learning. Computer Vision and Image Understanding, 137, 1–11.

Naglieri, J. A., McNeish, T. J., & Achilles, N. (2004). Draw a person test. Tools of the Trade: A Therapist's Guide to Art Therapy Assessments, 124.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. R. (2011). ORB: An efficient alternative to SIFT or SURF. ICCV, 11, 2.

Yama, M. F. (1990). The usefulness of human figure drawings as an index of overall adjustment. Journal of Personality Assessment, 54(1–2), 78–86.

Yang, J., Jiang, Y.-G., Hauptmann, A. G., & Ngo, C.-W. (2007). Evaluating Bag-of-visual-words Representations in Scene Classification. Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval, 197–206. https://doi.org/10.1145/1290082.1290111

Yu, Q., Yang, Y., Liu, F., Song, Y.-Z., Xiang, T., & Hospedales, T. M. (2017). Sketch-a-Net: A Deep Neural Network that Beats Humans. International Journal of Computer Vision, 122(3), 411–425. https://doi.org/10.1007/s11263-016-0932-3

Yu, Q., Yang, Y., Song, Y.-Z., Xiang, T., & Hospedales, T. (2015). Sketch-a-Net that Beats Humans. ArXiv:1501.07873 [Cs].