Data collection scheme for training Deep Neural Network by voluntary individual contribution on Social Media

Mohammad Sohorab Hossain

Abstract— One of the recent prominent achievement of computer science community is the advancement in deep neural network. However, efficiency of these deep neural network models is highly dependent on the availability of huge amount of training data. Though in the age of internet data availability in many fields is not an issue, there are also many other fields (which can be radicalized by using deep neural network) facing a critical problem - lack of huge well-organized collection of data. We have proposed hereon a data collection scheme where individuals will voluntarily contribute by providing data over social media. As a possible application, we have proposed hereon a scheme to collect handwritten Bengali digit using a popular social media in Bangladesh named Facebook.

Keywords: Data collection, Training data, Artificial Neural Network (ANN), Deep Neural Network, Social media, Facebook.

1 Introduction

One of the recent prominent achievement of computer science community is the advancement in deep neural network. However, efficiency of these deep neural network models is highly dependent on the availability of huge amount of training data. Though in the age of internet data availability in many fields is not an issue, there are also many other fields (which can be radicalized by using deep neural network) facing a critical problem - lack of huge well-organized collection of data. Therefore, there should be some viable ways to collect these data. On one side, there are classical approach to collect data, such as, launching survey, organizing specific programs where participants will be paid to give personalized data, and so on. On other side, there are illegal practices to hack user accounts on social media to collect sensitive personal data. So, it is obvious that to collect data using current approaches, we have to either spend money or follow illegal means. Therefore, there is a clear need for a free and legal approach to collect huge, yet, well-organized data.

Social media or social network has completely changed the way we communicate and share our thoughts with each other. It has also enabled people to perform huge collaborative tasks by helping each other in new ways. People can now respond to any request within seconds. And, it has also created a positive vibe among people to volunteer in good collaborative tasks. Noticing these influences of social media on people's behavior and knowing the need of new approach to collect data, we have proposed hereon a data collection scheme where individuals will voluntarily contribute by providing data over social media. As a possible application, we have proposed hereon a scheme to collect handwritten Bengali digit using a popular social media in Bangladesh named Facebook.

The rest of the paper is organized as follows: Section 2 describes artificial neural network, deep neural network, and social media. Section 3 outlines our proposed data collection procedure. Section 4 describes a proposed application of our proposed framework, followed by the conclusion in section 5 and references in section 6.

2 Review Section

2.1 Artificial Neural Network and Deep Neural Network

Artificial Neural Network (ANN) is a mathematical model to mimic the biological brain, i.e., the biological neural network. A biological neuron has three main parts: Dendrites, Axon, and Axon terminals. Dendrites are used to collect signal from other connected neurons, axon is used to output signal from one neuron to another, and axon terminals make the actual connection between two neurons. Similarly, ANN has many neurons typically arranged in layers: input layer, hidden layer, and output layer.

Neurons from each layer are connected to the neurons of previous layer. These connections are actually a weights (real numbers) which are multiplied by the output of neurons to make input signals of next layers neuron. A Non-linear function is applied on the sum of these weighted inputs to a particular neuron, and that neuron fires, i.e., produces an output signal if

• *Author Mohammad Sohorab Hossain is currently working as a Lecturer of the Department of Energy Science and Engineering at Khulna University of Engineering & Technology, Bangladesh.*
Email: sohorabkuet@ese.kuet.ac.bd.

the output of the non-linear function is above a threshold. These weights are adjusted so that the error, i.e., difference between the calculated output and the target output decreases. This adjusting is done in training stage of the ANN. The weight values determine the contribution of input signal to the output of a neuron. There may be multiple hidden layer. Each layer in an ANN does some kind of unique transformation on the input data. ANN is very useful in classification tasks. But to build a working ANN model a huge collection of training data is required.

Deep neural network (DNN) is a special type of ANN. Deep means more hidden layer in the ANN architecture. This approach is opposite of Shallow network approach, where more neurons are added in one hidden layer instead of adding new hidden layer. Each new hidden layer learns more abstract representation of the previous layer's output, and thus learns to map input to output in steps. It is believed that this approach significantly improves learning rate of the ANN [1-3].

There are many kinds of DNN which are basically variants of basic DNN architecture. Among them Feed-forward networks are the most used architecture. This architecture does not have any loop in the data flow path from input layer to output layer. Beside Feed-forward network recurrent neural networks (RNNs) are also used in many applications, such as, natural language processing [4-9]. Convolutional neural networks (CNNs) which are developed recently have brought radical changes in Computer vision and automatic speech recognition (ASR) tasks [10-11].

2.2 Social Media

Social media is internet based technology to nurture interpersonal relationships, business relations, and it is a way to express personal ideas, thoughts about social and political issues. On social media people can do all of these by posting comments, thoughts, photos, and videos. Now-a-days there are a lot of different types of social media which only complicates the definition of social media. However, there are some common traits in all kind of social media which are listed below,

- Social media is almost always interactive and based on Web 2.0 web system [12-13].
- Users have their specific account which are maintained by them, i.e., they can change their profile informations at any time [12] [14].
- User created contents are the main driving force of social media. Users share their comments, photos, and videos on social media which enrich it [12-13].
- Social media spreads or grows as a network by linking user profiles depending on users' interpersonal relationships [12] [14].

Social media has radically changed the communication system. It has brought noticeable and long lasting changes in the way organizations, communities, and individuals communicate with each other [15].

The quality [16], interactivity, reach, and participants are different among social media and electronic media like TV broadcasting or paper-based media. Social media is a dialogic transmission system (i.e., many sources to many receivers) whereas other two medias mentioned above are monologic transmission systems (i.e., one source to many receivers) [17]. People usually use social media on desktop or laptop computers, and smartphones with internet connection. Most popular social media sites today are Facebook, Twitter, Google+, WhatsApp, Instagram, Snapchat, and Myspace. In Bangladesh the most popular one is Facebook, followed by Twitter.

3 Data Collection Procedure

To collect data for training deep neural network we propose the following steps:

1. Find Social media users (seed users) with good responsive connections/friends on social media.
2. Post status explaining the goal, importance and procedure of data collection by the seed users.
3. Include a request to participate and spread news about this event in the post.
4. Repost the status frequently.
5. Make simple data storage system and give right to access the data by anyone.
6. Make a system to process the data.
7. Showcase the result or outcome of the deep neural network.
8. Give appropriate credit to the participants.

To get good response and to maximize data collection, users should have enough friends on social media, and should have few mutual friends. If there are more mutual friends, target audience will be small. However, there should be some mutual friends so that news can spread more efficiently. Moreover, it is very important to explain the goal, importance and procedure

to collect data to gain public faith so that they become interested to participate. Also, it is important to explicitly request others to participate in this action and spread news about it. To get better response from others, it is imperative that users should repost the same status frequently, otherwise participants' interest in the action will evaporate with time. It is very important to save and organize the collected date. For low budget application, free or very cheap cloud storage system, e.g., google drive, dropbox, can be used. The organizers should also make sure that anyone can access those data to maintain transparency and credibility of the data collection scheme. It should be noted that, most of the data collected in this way will be unorganized and inefficient to store in raw format. Therefore, some kind of processing has to be performed on these raw data. After this step, the data will be ready to use for training deep neural network (DNN). It is imperative that the result or outcome of the DNN have to publish publicly. It can be done by publishing journal or conference paper. However, to reach everyone results should be summarized in a public post on the social network used to collect the data. As a good practice and to inspire the participants, proper credit should be given to the participants.

4 Example Application

As an application of our proposed data collection scheme, we will discuss about possible scenarios to collect data on handwritten Bengali numerical digit using a popular social network in Bangladesh namely Facebook.

First step is to find seed users with good connection who will post a request on their profile to submit pictures of hand written digits voluntarily. Assume, there are N number of such seed users in this case. Let, number of connection of each individual seed user is C_i, where, $i = 1, 2, \dots, N$. Let, percentage of connection of each seed user willing to participate is η_i, where, $i = 1, 2, \dots, N$. In this case, number of data classes is, $L = 10$, i.e., 10 Bengali numerical digits. Now after a time T, number of data collected is given by the following proposed equation,

$$\text{Number of data collected, } S = \sum_{i=1}^{N} e^{-iC_i\eta_i T} \quad -----(1)$$

Participants can post pictures of hand written digits as comments on the original posts or an individual page on Facebook can be created and used to collect those pictures. In our view, the later approach is more convenient. To make data collection and post-processing steps easier, it is advisable to ask participants to write the digits in sequence, i.e., from 0 to 9 sequentially. Some of the critical post-processing steps are converting color images to gray scale images, resizing the images into fixed predefined size, crop image segments each containing one individual digit, resize cropped image into fixed predefined size, and saving those images into ten categories as input from one individual participant. Google drive offers free storage of 15 GB which can be used to save these images.

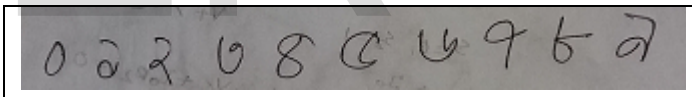
4.1 Image color space conversion

Color images have mainly 3 channels (matrices): red (R), green (G), and blue (B). Each channel can contain pixel values from 0 to 255. A gray scale image has only one channel and can contain pixel values from 0 to 255. To convert color image to gray image, pixel values from 3 channels are averaged and rounded to get a single value ranging from 0 to 255. Therefore, pixel value in gray image is given by,

$$G = \frac{R + G + B}{3} \quad -----(2)$$

We can also use popular open source computer vision library OpenCV to do this simple task. OpenCV has a function called 'cvtColor' to transform image from one color space to another. This function uses following standard equation instead of equation (2) to convert RGB image to Gray scale image.

$$G = 0.299\,R + 0.587\,G + 0.114\,B \quad -----(3)$$



(a)

(b)

Figure 1. (a) An example of input image containing hand written Bengali digits. (b) Binary image of the input.

After having the gray scale images for further processing we have to convert this images to binary images, i.e., black and white images. This is done by comparing every pixel of an image with a threshold. If pixel value is above threshold, it is set to 255 (i.e., making it a foreground white pixel), and if pixel value is below threshold, it is set to 0 (i.e., making it a background black pixel). OpenCV has a function called 'threshold' to transform gray image into binary image.

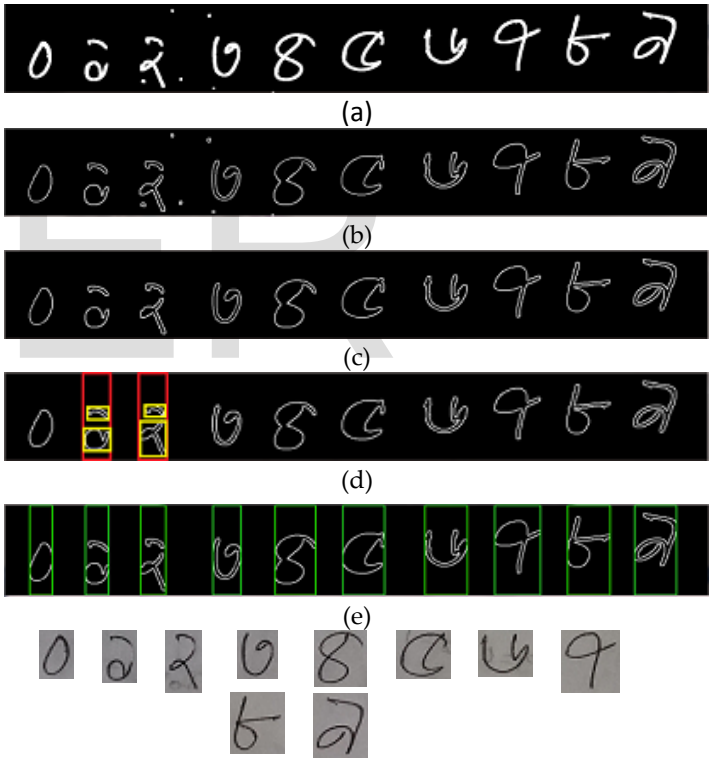
4.2 Cropping image into segments containing individual digit

Participants will asked to give single image with all of their 10 hand written digits in a single row to increase convenience in their part and to maximize participation. So, to store data into individual categories, it is required to crop image segments from input images each containing individual digit. To do this, we can flow the following steps,

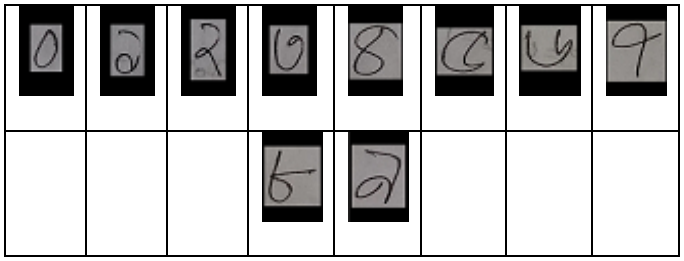
1. Apply morphological operations, i.e., erosion twice and dilation once to remove noise from the binary images.
2. Initially take 10 equally wide segments assuming each segment contains individual digit. So, there will be 20 vertical lines that represents 20 vertical sides of the 10 segments.
3. For each odd numbered vertical line, consider a rectangle whose left side is that vertical line and whose width is some predefined value, named 'search_window_size' = 20 pixels. Now, we have to calculate the number of pixels of a digit enclosed by this rectangle. If it is lower than some predefined or tunable value (say, 3), shift the odd numbered vertical line in the right direction by half of the 'search_window_size', i.e., by 10 pixels in this case, and repeat step 4 again. Otherwise, shift the vertical line in left direction by 10 pixels and take this location as its fixed position.
4. For each even numbered vertical line, consider a rectangle whose right side is that vertical line and whose width is some predefined value, named 'search_window_size' = 20 pixels. Now, we have to calculate the number of pixels of a digit enclosed by this rectangle. If it is lower than some predefined or tunable value (say, 3), shift the even numbered vertical line in the left direction by half of the 'search_window_size', i.e., by 10 pixels in this case, and repeat step 5 again. Otherwise, shift the vertical line in right direction by 10 pixels and take this location as its fixed position.
5. Consecutive odd and odd numbered vertical lines will enclose a digit, and now we can crop these segments containing individual digit.

An alternative and recommended procedure to achieve the same above mentioned result is discussed below.

1. Apply morphological operations, i.e., erosion twice and dilation once to remove noise from the binary images.
2. Find the contours of the binary images.
3. To remove noise exclude the contours whose area is smaller than a minimum area.
4. It is possible that several contours are very close to each other, which may indicate that they are of the same digit. Consider such contours as a single contour. Now, find the bounding boxes of all the unique contours which represent individual digits.
5. Crop images from the input images according to the bounding boxes which will contain individual digits.
6. Resize all the cropped images to predefined size using zero padding.



(a) (b) (c) (d) (e) (f)



(g)

Figure 2. Outputs of different processing steps mentioned in the 2nd method, a) Binary image after applying morphological operations, (b) All contours of the binary image, (c) Image after removing noise, i.e., contours having very small area, (d) Close contours are considered as one and bounded in same Rectangle, (e) Bounding boxes around unique contours which represent individual digits, (f) Cropped images, and (g) Resized images.

4.3 Resize cropped image into fixed predefined size

Following the procedure explained in section 4.2, we will have cropped segments or images containing individual digit, but each segments will have different size. In a database, all the images should have same size, i.e., width and height. So, all these mismatched segments should be resized to a fixed predefined size. We can use OpenCV's 'resize' method to do this task.

4.4 Saving images into ten categories

Following the procedure explained in section 5.4, for each participant we will have 10 images of 10 digits. Now, we can create 10 folders on Google drive, each for different category, and save images in respective folders. We can also mark which user has given which images.

5 Conclusion

This paper presents a novel approach to collect data for training artificial neural network and deep neural network. It proposes a new way to harness the power of social media or social network to improve the efficiency and availability of artificial neural network and deep neural network. We detailed the steps of data collection using our technique for a proposed application. In the future, we will conduct actual experiment to collect data for the proposed application using our proposed data collection scheme.

6 References

[1] Bengio, Yoshua (2009). "Learning Deep Architectures for AI." Foundations and Trends in Machine Learning. 2 (1): 1–127. doi:10.1561/2200000006.

[2] Schmidhuber, J. (2015). "Deep Learning in Neural Networks: An Overview." Neural Networks. 61: 85–117. doi:10.1016/j.neunet.2014.09.003. PMID 25462637.

[3] Szegedy, Christian; Toshev, Alexander; Erhan, Dumitru (2013). "Deep neural networks for object detection." Advances in Neural Information Processing Systems.

[4] Gers, Felix A.; Schmidhuber, Jürgen (2001). "LSTM Recurrent Networks Learn Simple Context Free and Context Sensitive Languages." IEEE Trans Neural Netw. 12 (6): 1333–1340. doi:10.1109/72.963769.

[5] Sutskever, L.; Vinyals, O.; Le, Q. (2014). "Sequence to Sequence Learning with Neural Networks." Proc. NIPS.

[6] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, Yonghui Wu (2016). Exploring the Limits of Language Modeling.

[7] Dan Gillick, Cliff Brunk, Oriol Vinyals, Amarnag Subramanya (2015). Multilingual Language Processing From Bytes.

[8] Mikolov, T.; et al. (2010). "Recurrent neural network based language model."

[9] Hochreiter, Sepp; Schmidhuber, Jürgen (1997-11-01). "Long Short-Term Memory." Neural Computation. 9 (8): 1735–1780. doi:10.1162/neco.1997.9.8.1735. ISSN 0899-7667. PMID 9377276.

[10] LeCun, Y.; et al. (1998). "Gradient-based learning applied to document recognition." Proceedings of the IEEE. 86 (11): 2278–2324. doi:10.1109/5.726791.

[11] Sainath, T. N.; Mohamed, A. r; Kingsbury, B.; Ramabhadran, B. (May 2013). "Deep convolutional neural networks for LVCSR." 2013 IEEE International Conference on Acoustics, Speech and Signal Processing: 8614–8618. doi:10.1109/icassp.2013.6639347. ISBN 978-1-4799-0356-6.

[12] Obar, Jonathan A.; Wildman, Steve (2015). "Social media definition and the governance challenge: An introduction to the special issue." Telecommunications policy. 39 (9): 745–750. doi:10.1016/j.telpol.2015.07.014. SSRN 2647377 Freely accessible.

[13] Kaplan Andreas M., Haenlein Michael (2010). "Users of the world, unite! The challenges and opportunities of social media." Business Horizons. 53 (1): 61. doi:10.1016/j.bushor.2009.09.003.

[14] Boyd, danah m.; Ellison, Nicole B. (2007). "Social Network Sites: Definition, History, and Scholarship." Journal of Computer-Mediated Communication. 13 (1): 210–30. doi:10.1111/j.1083-6101.2007.00393.x.

[15] Kietzmann, Jan H.; Kristopher Hermkens (2011). "Social media? Get serious! Understanding the functional building blocks of social media." Business Horizons. 54 (3): 241–251. doi:10.1016/j.bushor.2011.01.005.

[16] Agichtein, Eugene; Carlos Castillo. Debora Donato; Aristides Gionis; Gilad Mishne (2008). "Finding high-quality content in social media." WISDOM – Proceedings of the 2008 International Conference on Web Search and Data Mining: 183–193.

[17] Pavlik & MacIntoch, John and Shawn (2015). Converging Media 4th Edition. New York, NY: Oxford University Press. p. 189. ISBN 978-0-19-934230-3.

[18] Hajirnis, Aditi (2015-12-01). "Social media networking: Parent guidance required." The Brown University Child and Adolescent Behavior Letter. 31 (12): 1–7. doi:10.1002/cbl.30086.

[19] Tang, Qian; Gu, Bin; Whinston, Andrew B. (2012). "Content Contribution for Roddue Sharing and Reputation in Social Media: A Dynamic Structural Model." Journal of Management Information Systems. 29 (2): 41–75. doi:10.2753/mis0742-1222290203.