ELL-888: Assignment number 2 - Deep models for spatial data

Prof. Prathosh A. P.

Submission deadline: 30th March 2018

1 Introduction

This goal of this assignment is to get students aquatinted with the use of deep neural network architectures for spatially distributed data (Eg. Images). Specifically, students are expected to employ convolutional deep nets to perform a classification task. This assignment is designed to test wide-range of abilities of the students including (1) judiciously curating the data, (2) handling the data and label noise, (3) casting the problem within the deep-learning framework, (4) choosing and experimenting with the models and architectures, (5) using existing high-level coding frameworks (6) Validating the learned-models and (7) testing the models on unseen data.

2 Problem statement

2.1 Context and background

One of the biggest success of deep learning is in solving several image understanding tasks ranging from object identification to image captioning that have innumerable applications in diverse fields including healthcare, pedagogy, finance, bio-metrics etc. The objective of this task to solve one such image understanding task.

In this era information, there is huge amount of data that gets generated every second. One challenge is curate the data in accordance with several levels of meta-information. A few examples of such tasks include tagging photos based on people and background, classifying music based on genre, arranging items based on utility-category in online-shopping scenario, classifying electronic medical records according to the diseases type and so on (one can create infinite tasks). Here we consider a small but interesting problem of **frame-wise identification of the dominant speaker in short videos** - the final context is to identify "who is speaking when" in a video, given that there are N number of possible speakers (imagine tagging the frames of a long stand-up comedy show video where there are 5 comedians performing).

2.2 Our problem

Since the problem described in the previous section is too broad and complex (both in terms of solution and data-availability) we shall consider a simpler version of it in this assignment. The precise problem is described in the following points.

1. Several short videos (series of static image frames) that only contain a (known from a set of N speakers) **single** human speaker speaking all through out are given.

2. The videos also occasionally contain some irrelevant frames in the sense that there are frames that does not have the speaker (Eg. frames where the crowd is shown in a stand-up-comedy video): this is the source of label noise.

3. Suppose there are videos from N such speakers (6 in the case of this problem).

4. The primary task is to build a deep neural network model that maps each frame in the video to the corresponding speaker (training phase).

5. Once the model is built, during the test phase the task is to identify speaker each frame of an unseen test video as belonging to one of the six speakers.

The problem falls under the broad scope of human recognition and object identification from images. However the challenges arises from the fact that the speaker will have several different spatial orientations, different clothings, different backgrounds in each videos. Additionally, there is label noise (though rare) since the entire video are given one label albeit it contains frames that does not have the object (speaker) being identified.

2.3 Data and labels

We shall consider short youtube videos from **6 speakers - (1) Sadguru Jaggi Vasudev - Indian yogi and mystic (2) Sandeep Mashehwari - motivational speaker, (3) Saurabh pant - standup comedian, (4) Atul Khatri - standup comedian, (5) Shailendra kumar - Guitarist and (6) - Flute Raman - Classical Flutist.** The following are the links from which one can form the training data (note that this is list is not comprehensive, you are free to expand your training dataset).

1. **Sadguru Jaggi Vasudev -** https://www.youtube.com/watch?v=3J-cYxxHQGQ, https://www.youtube.com/watch?v=vQ7ZvPghdy8, https://www.youtube.com/watch?v=fTx9tOmU1sY, https://www.youtube.com/watch?v=CmVQuiT0OTw, https://www.youtube.com/watch?v=nNcFquUuKww https://www.youtube.com/watch?v=rJZjd7rFKws, https://www.youtube.com/watch?v=LNyJgNjCDuU, https://www.youtube.com/watch?v=e2EPuGabgpc

2. **Sandeep Maheshwari -** https://www.youtube.com/watch?v=oB09kuJa-Eg, https://www.youtube.com/watch?v=Y07dnUKwqyw, https://www.youtube.com/watch?v=uWnBTyiSwgE, https://www.youtube.com/watch?v=b_clFB3vL-Q, https://www.youtube.com/watch?v=Ho37w_UFRSg, https://www.youtube.com/watch?v=rGpwRlCOLbY

3. **Saurabh Pant -** https://www.youtube.com/watch?v=tXeZupaycWI, https://www.youtube.com/watch?v=s https://www.youtube.com/watch?v=dkbtHVayA3U, https://www.youtube.com/watch?v=crqlI1Exte4, https://www.youtube.com/watch?v=klP7mKwhweo, https://www.youtube.com/watch?v=4eiyui0dF5o

4. **Atul Khatri -** https://www.youtube.com/watch?v=x09Ft-XdChg, https://www.youtube.com/watch https://www.youtube.com/watch?v=Lc81vSTHGMY, https://www.youtube.com/watch?v=udQ4Igf https://www.youtube.com/watch?v=CHQ3odjT7oY

5. **Shailendra kumar -** https://www.youtube.com/watch?v=Wp_Al0AYylA&list=PLzdjxgz3O7oMZLoZsqAw
   https://www.youtube.com/watch?v=xkX6RU-q4N0&list=PLzdjxgz3O7oMZLoZsqAwX_4lWRLouhZub&inc
   https://www.youtube.com/watch?v=t2P9fDEKCVE&index=14&list=PLzdjxgz3O7oMZLoZsqAwX_4lWRL
   https://www.youtube.com/watch?v=4CCqAdsHl-8&index=19&list=PLzdjxgz3O7oMZLoZsqAwX_4lWRLo
   https://www.youtube.com/watch?v=4o8E5dyvYCk&index=17&list=PLzdjxgz3O7oMZLoZsqAwX_4lWRLc

6. **Flute Raman -** https://www.youtube.com/watch?v=BTvFi5SZJnw, https://www.youtube.com/watch?v=Q
   Kg3M, https://www.youtube.com/watch?v=ir8o5Fxn4yk, https://www.youtube.com/watch?v=_dXZ_dywf
   https://www.youtube.com/watch?v=I3I-ab9FHHA, https://www.youtube.com/watch?v=_olkwh6lQ_s

2.4 Tips on training set curation

1. Once can extract and use every individual frame as a datapoint.
2. Be careful while extracting - some parts (typically in the beginning and ending phase of the video might contain frames without the speaker).
3. You may discard (algorithmically or manually) the spurious frames.
4. You may want to have a seventh category - **Irrelevant frame** (**frames without speakers).** However, it is not straight-forward to get labelled such examples.

2.5 Models, evaluation and grading policy

This is a novel problem (that was created for the purpose of this assignment) there is no one right model. It can be casted in several ways starting from a 6 (or 7) class (frame-level) image object classification problem to a complex audio-visual sequence/gesture modeling problem. The bare-minimum expectation is to build models using convolutional architectures. One can also use an appropriate pre-trained model if you feel fit.

The test set will comprise a set of similar videos on which one has to evaluate the trained models - The evaluation metric will be accuracy of frame-level speaker classification on the test data which will be released **two days before the submission deadline.** Additionally, students are expected to prepare a report (up to 4 pages in two-column format with one additional page for references) that would contain the details of experiments, models and evaluation. Grading is relative and consider several aspects described in the introduction section. Please note that it is not solely based on final accuracy. **The best three teams will get some extra credits too!**

2.6 Softwares, frameworks and resources

Students are free to use any existing deep-learning framework. Some suggestions are Keras with tensorflow backend (my personal choice), tensorflow, pytorch (second best) and Caffe. One may feel free to use the library developed during first assignment. This assignment might not need large scale computing. However, students are encouraged to use HPC access at IITD. Alternatively, Azure credits (likely to be allocated by mid March) can also be used.