Acoustic Features to Predict Topic Change in Instructional Video

Vineet Kumar, Sushant Gupta
Computer Science and Engineering Department, Jadavpur University, Kolkata
{vntkumar8,sushantgupta1996}@gmail.com

1 Introduction

Education is one of the highest priority spends for households in many developed and developing countries and a significant part of this spending goes in the higher education space. In India, the spending has increased consistently in last two decades at an annual growth rate of nearly 21% in both rural and urban areas. In parallel, national governments are also increasing budgetary allocations in the education sector.

Instructional videos have given an open challenge to all current methods of education system. They have high potential of acceptability among all kind of learners. Instructional videos have become a label for many recent course initiatives from higher education institution. We believe that instructional videos are going to be next generation textbooks.

2 Motivation & Contribution

The Goal of this Project is to automatically identify the locations where the topic has changed in an instructional video. We specifically wanted to explore techniques that use acoustic (i.e. spoken) features. This was essentially a classification task. Given acoustic features of particular time, the model has to predict whether at a particular time,there is any topic change.

However in a general video of an hour, we noticed around 15-20 topic changes. If we assume topic change to be a 1 second transition then in 3600 seconds (1 hour) we had only 15-20 seconds where the topic has actually changed. This leads to a class imbalance problem. We have used a nice method to tackle this class imbalance problem by down-sampling.

3 Data Set Preparation & Feature Extraction

There are no standard dataset for preparation of such state of art model. We had to prepare our own curated and labeled data set. This was the most time consuming and bulky task of our project. We had to watch each and every video and manually note down the exact time of topic change as accurate in terms of second as well as the topic that was covered.

We had a dataset of around 120 hours Instructional Video. We extracted the *acoustic* features from these videos[2]. Due to large size of data we used Apache Spark for training our binary classifier. We used algorithms from Ml-Lib library.

4 Model Training

Using Learning Vector Quantization (LVQ) model[3] we ranked the features according to their importance. An LVQ system is represented by prototypes $W = (w(i), ..., w(n))$ which are defined in the feature space of observed data. In training algorithms one determines, for each data point, the prototype which is closest to the input according to a given distance measure. The position of this so-called winner prototype is then adapted, i.e. the winner is moved closer if it correctly classifies the data point or moved away if it classifies the data point incorrectly. The most important features were chosen by using this approach and they were used for further process. We tried few binary classifiers to train our model (with some parameter tuning). Random Forest and SVM performed best. Giving accuracy of 78%.

This is an extended abstract submitted to Paper Presentation Event of Convolution 4.0. The work was done when both authors were intern at Xerox Research Centre India Bangalore

5 Future Directions

In order to improve our accuracy we are planning to train our model on Support Vector Machines for Multiple instance Learning[1]. Multiple-instance learning (MIL) is a variation on supervised learning. Instead of receiving a set of instances which are individually labeled, the learner receives a set of labeled bags, each containing many instances. In the simple case of multiple-instance binary classification, a bag may be labeled negative if all the instances in it are negative. On the other hand, a bag is labeled positive if there is at least one instance in it which is positive. From a collection of labeled bags, the learner tries to either (i) induce a concept that will label individual instances correctly or (ii) learn how to label bags without inducing the concept.

6 Conclusion

We show in this paper how acoustic features can be used as an additional advantage to build intelligent models which will eventually help us to predict close to accurate topic changes in any instructional video. We want to combine the text based (OCR) and acoustic features for much better results. This will enable us to generate Table of Content for any instructional video.

References

[1] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, NIPS'02, pages 577–584, Cambridge, MA, USA, 2002. MIT Press. URL http://dl.acm.org/citation.cfm?id=2968618.2968690.

[2] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 835–838, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2404-5. doi: 10.1145/2502081.2502224. URL http://doi.acm.org/10.1145/2502081.2502224.

[3] Teuvo Kohonen. The handbook of brain theory and neural networks. chapter Learning Vector Quantization, pages 537–540. MIT Press, Cambridge, MA, USA, 1998. ISBN 0-262-51102-9. URL http://dl.acm.org/citation.cfm?id=303568.303833.