ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2015

Acoustics Recognition and Video Sound-Track Classification using CNN

Venkata Sai Sriharsha Sammeta[1], Naveen Y[2], C Suresh[3]

Student, Computer Science, Vasavi College of Engineering, Osmania University & Intern@Oracle India [1]

Student, Computer Science, BVRIT, Hyderabad, India [2]

Senior Software Manager, Oracle India R&D Pvt.Ltd[3]

ABSTRACT: For effective classification of image and audio Convolutional Neural Networks (CNNs) are used. CNN architectures help to classify videos that contain large data. To classify 70M large training videos various CNN architectures shows its effectiveness.In this, we will test AlexNet, VGG, Inception, ResNet and Deep Neural Networks (DNNs).In this, we traverse different sized subset of training videos and labels. Dataset carry Audio set, AcousticEvent Detection (AED)and Video-level labels. Derivatives of image classification network works good on audio classification network to increases producing number of labels. Datasets like Alex Net, Performance of models ameliorate with increase in training set size and model using from video level task perform well then a baseline on Audio set classification.

KEYWORDS: Text detection, Inpainting, Morphological operations, Connected component labelling, Recurrent Neural Networks, Classification, Time-Series.

I. INTRODUCTION

VGG, Inception and ResNet help to improve image classification using Convolutional Neural Network (CNN) architecture. Dataset contains 70milloin training videos total 5.24million hours tagged from 30k labels.The mission is to speculate the video level labels by utilizing audio information. In presentation of Lyon, teaching machines to hear and understand video can upgrade our capability to “categorize, organize, and index them”.
In this paper we utilize the YouTube-100M dataset to explore: Importance of Deep Neural Networks (DNNs) when compared with video soundtrack classification; how performance varies in training set and vocabulary sizes; and if our trained videos models can be used for AED.

AED consists of features like MFCCs and classifiers based on GMMs, HMMs,NMF, or SVMs. Advance work has been done on datasets like TRECVid, ActivityNet, Sports1M,and TUT/DCASE Acoustic scenes 2016. For Large Vocabulary Continuous Speech Recognition (LVCSR) RNNs and CNNs are applied. RNNs and CNNs do not contain labeled sequence data like LVCSR so we have not yet tried. Acoustic Scene Classification (ASC) task, involves assigning a single label to an audio clip containing many events. The ASC used i-vector features to feed a VGG classifier.Our objective is to analogize the behavior of different architectures.

The principle used in the visual-based video classification is found simple averaging of single-frame CNN classification outputs. By correlation, we apply a classifier to series of non-overlapping segments, and then average all the sets of classifier outputs. Contemplate AED in a dataset with video level labels as a Multiple Instance Learning (MIL) problem, but observe that that scaling such approaches remains an open problem.

By antithesis, we are exploring the limit s of training with weak labels for very large datasets. We anticipate that many of the individually labeled segments to be ambiguous about their labels, but one can find out the necessary cues with the net, which is given by training sufficiently. We cannot capacitate how “weak” the labels are and for the bulk of

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2015

classes it's not clear how to decide applicability. For some classes' background environs is itself informative, and this maybe be quite common.

Our dataset size allows us to inspect networks with large model quantity, taking assert of the substantial literature on image classification architectures. By figure out log mel spectrograms of different frames, we create 2D image-like patches which can be presented to the image classifiers. Although the time and frequency axes have definite meanings, which might argue for audio-specific model architectures, this work scrutinizes, the accomplishment of minimally altered image classification networks such as Inception-V3 and ResNet-50. We instruct with subsets of YouTube-100M spanning 23K to 70M videos to assess the impact of training set size on presentation, and we explore the effects of label set size on generalization by training models using subsets of labels, spanning 400 to 30000, that are evaluated on a single common subset of labels.

Prior Label	Example Labels
0:1:::0:2	Song, Game, Sports, Performance, Music
0.01:::0:1	Singing, Car, Speech, Chordophone
~10^-5	Custom Motorcycle, Retaining Wall
~10^-6	Cormorant, Lecturer

Table 1: Table 1: Labels from the 30K set.

We additionally investigate the usefulness of our networks for AED by exploring the presentation of a model trained with i The YouTube-100Million data set composed of 100 million YouTube videos: 10Million evaluation videos, 70Million training videos and a pool of 20Million videos that we use for affirmation. Videos average 4.6 minutes each for a total of 5.4Million training hours. Each of these videos is labeled with 1 or more topic qualifiers from a set of 30,871 labels. There are averages of around 5 labels per video. The labels are allocated automatically based on amalgamate of metadata context, and image content for each video. The labels relate to the entire video and range from very generic to very specific Table 1 shows a few examples.

Being machine initiated, the labels are not 100% meticulous and of the 30K labels, some are clearly acoustically applicable and others are less so Videos are often commented with similar labels with varying degrees of precision. For example, videos labelled with "Trumpet" will tend to be labeled "Entertainment" as well, but no hierarchy is accomplished plant from one of our networks on the Audio Set dataset.

II. RELATED WORK

The YouTube-100M data set composed of 100 million YouTube videos: 70Million training videos, a pool of 20Million videos and 10Million evaluation videos, that we use for affirmation. Videos average 4.6 minutes each for a total of 5.4M training hours. Each of these videos is labeled with 1 or more topic qualifiers from a set of 30,871 labels. There are averages of around 5 labels per video. The labels are allocated automatically based on amalgamate of metadata context, and image content for each video. The labels relate to the entire video and range from very generic to very specific Table 1 shows a few examples.
Being machine initiated, the labels are not 100% meticulous and of the 30K labels, some are clearly acoustically applicable and others are less so Videos are often commented with similar labels with varying degrees of precision. For example, "Trumpet" label can be "Entertainment" labeled as well, but no hierarchy is accomplished.

A. *TRAINING*

The audio is split into non-overlapping 960ms frames. This gave about 20 billion examples from the 70M videos. Each frame obtains all the labels of its parent video. The 960 ms frame are perished with a short-time Fourier transform applying 25 ms windows every 10 ms. The derived spectrogram is united into 64 mel-spaced frequency

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2015

bins, and the magnitude of each bin is log converted after adding a small offset to avoid numerical issues. This gives a log-mel spectrogram will form an input to other classifiers using patch of 96 × 64 bins. 128 input samples are fetched at the time of training by sampling all the patches equivalently.

All experiments used TensorFlow and were trained on various GPUs asynchronously using the Adam optimizer. We implemented grid inquires over learning rates, batch sizes, number of GPUs, and parameter servers. Batch regularizing was applied after all convolutional layers. All models used a final sigmoid layer rather than a softmax layer since each example can have various labels. Loss function will be considered as a Cross-function. In view of the large training set size, we did not use dropout weight decay, or other common normalizing methods. For the models trained on 7Million or more examples, we saw no confirmation of over fitting. During training, we continually calculated 1-best accuracy and mean Average Precision (MAP) over a subset to monitor progress.

B. EVALUATION

From the pool of 10M assessment videos we created three balanced assessment sets, each with approximately 33 examples per class: 1M videos, 100K videos for the 3087 (henceforth 3K) most persistent labels, and 12K videos for the 400 most persistent labels.

To calculate video-level apparently for each class, we passed each non-overlapping 960 ms frame from each assessment video with the help of the classifier. The classifier output scores are then averaged for each class across all sectors in a video. For our metrics, we calculated the equitable average across all classes of AUC and mean Average Precision (MAP).

C. *ARCHITECTURES*

We used a fully associated DNN as our baseline and analysed this with several networks that were based as closely as possible on successful image classification models. For our baseline architecture analyses, we trained and assessed using only the 10% most constant labels of the original 30K. For all experiments, we made a bristly search for the optimal number of GPUs and learning rate for the frame level classification accuracy. The excellent number of GPUs represents accommodates between overall computing power and communication overhead, and thus varies across network architectures.

i. Fully Connected

Our baseline network is a fully associated model with RELU actions, N layers, and M units per layer. We clear over N = [2, 3, 4, 5,] and M = [500, 1000, 2000, 3000]. Our finest intending model had N = 3 layers, M = 1000 units, learning rate of 3×10−5, 10 GPUs and 5 parameter servers. This network has relatively 11.2M weights and 11.2M multiplies.

ii. *AlexNet*

The authentic AlexNet architectures were composed with a stride of 4 and with an initial 11 × 11 convolutional layer for a 227 ×227 × 3 input. Since information is 96 × 64, we use a stride of 2 × 1 so that the number of activations is similar after the initial layer. We also use batch regularizing after each convolutional layer rather of local response normalization (LRN) and change the final 1000unit layer with a 3087unit layer. While the authentic AlexNet has nearly 62.4M weights and 1.1G multiplies, our version has 37.3M weights and 767M multiplies. Also, for simplicity, unlike the Authentic AlexNet, we do not split filters across various devices. We trained with 20 GPUs and 10 parameter servers.

iii. *VGG*

The only changes we made to VGG were to the eventual layer as well as the use of batch normalization rather of LRN. While the authentic network had 144M weights and 20B multiplies, the audio alternative uses 62M weights and 2.4B multiplies. We tried another alternative that reduced the original strides, but found that not adjusting the strides

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2015

consequence in faster training and better behaviour. With oursetup, comparing above10 GPUs did not help compellingly, so we trained with 10 GPUs and 5 parameter servers.

Architectures	Steps	Time	AUC	d-prime	mAP
Fully Connected	5M	32h	0.856	1.478	0.059
AlexNet	5M	83h	0.874	1.784	0.215
Inception V3	5M	137h	0.913	1.965	0.184
ResNet-50	5M	129h	0.956	1.962	0.192
ResNet-50	17M	316h	0.826	2.051	0.232

Table 2: Analysing function of several DNN architectures trained on 700M videos, each marked with labels from a set of 3K.The last row accommodates results for a model that was trained much longer than the others, with devaluation in learning rate after 13 million steps.

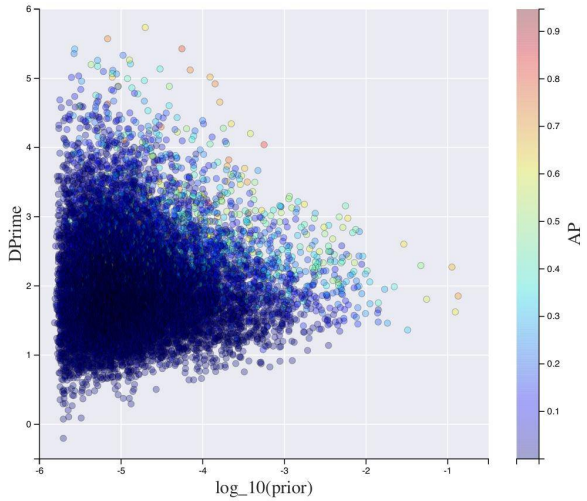

Fig. 1: Scatter plot of ResNet-50's per-class d-prime versus log prior Credible. Each point is a distinct class from a random 20% subset of the 30K set. Color reflects the class AP.

iv. *Inception V3*

We changed the inception V3 network by abolishing the first four layers of the stem, up to and combining the MaxPool, as well as abolishing the auxiliary network. We modified the Average Pool size to 10 × 6 to reflect modifies in activations. We tried combining the stem and abolishing the first stride of 2 and MaxPool but found that it practiced worse than the alternate with the abbreviated stem. The authentic network has 27M weights with 5.6B multiplies, and the audio alternate has 28Million weights and 4.7Billion multiplies with 40 GPUs and 20 parameter servers.

v. *ResNet-50*

We changed ResNet-50 [4] by abolishing from the first 7×7 convolution with 2 stride so that the total activations was not too disparate in the audio version. We modified the Average Pool size to 6 × 4 to follow the adjustment in activations. The authentic network has 26M weights and 3.8B multiplies. The audio alternate has 30M weights and 1.9B multiplies with 20 GPUs and 10 parameter servers.

III. EXPERIMENTAL RESULTS

A. ARCHITECTURE SIMILARITY

Network architectures trained with 3K labels and 70M videos and analysed after 5 million mini-batches of 128 inputs. Because some networks trained faster than others, analysing after a constant wall-clock time would give little

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2015

contrasting results but would not modify the analogous ordering of the architectures 'implement. We attach numbers for ResNet after training for 17 million minibatches (405 hours) to show that achieve continues to recover. Learning rate can be reduced by a factor of 10 after 13 million mini-batches.

Table 2 shows the assessed results calculated over the 100K balanced videos. All CNNs beat the fully-connected baseline. Inception and ResNet accomplish the best performance; they administer high model scope and their convolutional units can calmly capture common structures that may occur in different areas of the input array for both images, and, we ascertain, our audio representation.

To investigate how the previous likelihood of each label affects it's operate, Fig. 1 shows a discard plot of the 30K classes with label frequency on the x axis and ResNet-50's d-prime on the y axis. d-prime seems to stay focused around 2.0 across label prior, although the changes of d-prime increases for less-common classes.

B. ***SIZE OF LABEL SET***

Out of 30K labels 400 label sets were used to explore how training with different subsets of classes can affect behaviour; perhaps by embolden intermediate representations that better generalize to unseen examples even for the evaluation classes. In addition to investigating three label set sizes (30K, 3K, and 400), we analysed patterns both including and excluding the 128 units from a bottleneck layer placed right in front of the final output layer. We initiating the bottleneck layer to speed up the training model of 30K labels. Without a bottleneck, the larger output layer increased the number of weights from 30M to 80M and reduced training speed. We do not address metrics on the 30K label model without the bottleneck because it takes lot of time to train. Among the label set size experiments, we used trained for 5 million mini-batches of 128 inputs (about 120 hours) on 70M videos and the ResNet-50 model.

Tables 3 show the results, when we analyse models with the bottleneck, we see that present does indeed increase slightly as we increase the number of labels we trained on, although networks without the bottleneck have higher implement overall. The bottleneck layer is correspondingly small compared to the 2048 activations coming out of ResNet-50's Average Pool layer and so it is consequence a substantial decrease in information.

C. TRAINING SET SIZE

Having a very large training set possible allows us to explore how training set size affects behaviour. With 70M videos and regular of 4.6 minutes per video, we have around 20 billion 960 ms training examples. Given ResNet-50's training speed of 11 minibatches per second with 20 GPUs, 23 weeks are required to find out each pattern once (one epoch) for the network. However, if all videos were parallel length wise and completely randomized, we assume to detect one frame per video in only 14 hours. We consider that, even if we cannot get through an entire epoch, 70M videos will provide choice over 7M by virtue of the greater diversity of videos underlying the limited number of training patterns gained. We trained 16 million mini-batches of 128 inputs using ResNet-50 model (about 380 hours) on the 3000 label set with 70Million, 7Million, 70000, and 23000 videos. Results are shown in Table 4. The 70000 and 23000 models show worse behaviour but the confirm plots presented that they likely adjust from overfitting. Regularization techniques (or data augmentation) might have encouraged the numbers on these smaller training sets. The 700K, 7M, and 70M models are mostly very close in achieve although the 700K model is marginally inferior.

Bneck	Labels	AUC	d-prime
no	30K	—	—
no	3K	0.93	2.087
no	400	0.928	2.067
yes	30K	0.925	2.035
yes	3K	0.919	1.982
yes	400	0.924	2.026

Table 3: Results of changing label set size, figure out over 400 labels. All models are alternative of ResNet-50 trained on 70M videos. The bottleneck, if present, is 128 dimensions.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2015

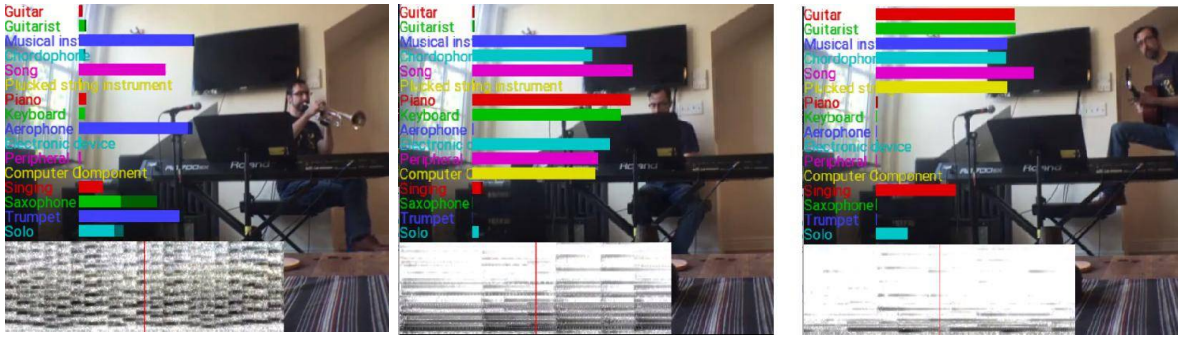

Fig. 2: Three example extract from a video classified by ResNet-50 with rapid model outputs overlaid. The 16 classifier outputs with the extreme peak values across the entire video were chosen from the 30K set for display.

Training Videos	AUC	d-prime	mAP
70M	0.923	2.019	0.206
7M	0.922	2.006	0.202
700K	0.921	1.997	0.203
70K	0.909	1.883	0.162
23K	0.868	1.581	0.118

Table4: Results of training where all rows used the same ResNet-50 architecture trained on videos attached with labels from a set of 30K.

IV. CONCLUSION

The decisions in Section 4.1 show that state-of-the-art image networks are adept of excellent results on audio allocation when to a simple fully connected network or earlier image architectures. In Section 4.2 we saw results showing that training on larger label set vocabularies can improve performance, albeit quietly, when calculating on smaller label sets. In Section 4.3 we saw that escalating the number of videos up to 7M boosts performance for the best-conducting ResNet-50 architecture. We note that regularization could have diminished the gap between the models trained on smaller datasets and the 7M and 70M datasets. In Section 4.4 we see an important increase over our baseline when training a model for AED with ResNet inserts on the Audio Set dataset.

In addition to these capable results, we can subjectively explore the management on segments of video the model. Fig. 2 explains the outcome of using the best classifier and overlaying the frame-by-frame results of the 16 classifier outputs with the extreme peak values across the entire video. The distinct sound sources present at recognizable points in the video are clearly distinguished.

REFERENCES

[1] I. Sutskever A. Krizhevsky, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in leades in neural information processing systems, 2012, pp.1097–1105.
[2] A. Zisserman, and K. Simonyan "Very deep convolutional networks for large-scale image observation," arXiv preprintarXiv:1409.1556, 2014.
[3] J. Shlens, C. Szegedy, V. Vanhoucke, S. Ioffe, and Z. Wojna, "Rethinking the inception architecture for computer vision," arXiv preprint arXiv:1512.00567, 2015.
[4] S. Ren, K. He, X. Zhang, and J. Sun, "Deep residual learning for image recognition," arXiv preprint arXiv: 1512.03385, 2015.
[5] D. P. W. Ellis, J. F. Gemmeke, D. Freedman, A. Jansen,W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-defined dataset for audio events," in IEEE 2017.