AUDIO SET CLASSIFICATION WITH ATTENTION MODEL: A PROBABILISTIC PERSPECTIVE

Qiuqiang Kong*, Yong Xu*, Wenwu Wang, Mark D. Plumbley

Center for Vision, Speech and Signal Processing, University of Surrey, UK
{q.kong, yong.xu, w.wang, m.plumbley}@surrey.ac.uk

ABSTRACT

This paper investigates the classification of the Audio Set dataset. Audio Set is a large scale weakly labelled dataset of sound clips. Previous work used multiple instance learning (MIL) to classify weakly labelled data. In MIL, a bag consists of several instances, and a bag is labelled positive if at least one instances in the audio clip is positive. A bag is labelled negative if all the instances in the bag are negative. We propose an attention model to tackle the MIL problem and explain this attention model from a novel probabilistic perspective. We define a probability space on each bag, where each instance in the bag has a trainable probability measure for each class. Then the classification of a bag is the expectation of the classification output of the instances in the bag with respect to the learned probability measure. Experimental results show that our proposed attention model modeled by fully connected deep neural network obtains mAP of 0.327 on Audio Set dataset, outperforming the Google's baseline of 0.314 and recurrent neural network of 0.325.

Index Terms— Audio Set, audio classification, multiple instance learning, attention.

1. INTRODUCTION

Analysis of environmental sounds has been a popular topic which has the potential to be used in many applications, such as public security surveillance, smart homes, smart cars and health care monitoring. Audio classification has also attracted significant research effort due to the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge [1, 2]. Several tasks have been defined for audio classification including acoustic scene classification [1], sound event detection [1] and audio tagging [3, 4]. However, the data sets used in these challenges are relatively small. Recently, Google released an ontology and human-labeled large scale data set for audio events, namely, Audio Set [5]. Audio Set consists of an expanding ontology of 527 sound event classes and a collection of over 2 million human-labeled 10-second sound clips drawn from YouTube videos.

Audio Set is defined for tasks such as audio tagging. The objective of audio tagging is to perform multi-label classification on fixed-length audio chunks (i.e. assigning zero or more labels to each audio chunk) without predicting the precise boundaries of acoustic events. This task was first proposed in DCASE2016 challenge. Deep neural networks (DNNs) [5] and convolutional recurrent neural networks (CRNNs) [3] have been used for predicting the occurring audio tags. Neural networks with an attention scheme was firstly proposed in our previous work [6] for the audio tagging task which provides the ability to localize the related audio events. Gated convolutional neural networks [7] have also been applied in the "Large-scale weakly supervised sound event detection for smart cars" task of DCASE2017 challenge, where our system achieved the 1st place in the audio tagging sub-task[1]. However, the audio tagging data set used in the DCASE2017 challenge is just a small sub-set of Google Audio Set [5]. The number of the audio event classes is only 17 compared with 527 classes in Google Audio Set. In this paper, we propose to use an attention model for audio tagging on Google Audio Set [5], which shows better performance than the Google's baseline. In this work, we have two main contributions, one is that we conduct and explore a large-scale audio tagging on Google Audio Set [5]. Secondly, we explain the attention model from a probability perspective. The attention scheme is also similar to the feature selection process which can figure out the related features while suppressing the unrelated background noise. It is achieved by a weighted sum over frames where the attention values are automatically learned by the neural network model.

In the remainder of this paper, the related works are presented in Section 2. The proposed attention method and explanation from the probability perspective are shown in Section 3. Section 4 presents the experimental setup and results. Conclusions are drawn in section 5.

2. RELATED WORKS

Multiple instance learning (MIL) [8, 9] is a variation on supervised learning, where each learning example contains a *bag* of

---

instances. In MIL, a positive bag contains at least one positive instance. On the other hand, a negative bag contains no positive instances. Each audio clip in Audio Set contains several feature vectors. An audio clip is labelled positive for a class if at least one feature vector belongs to the corresponding class.

A multi instance dataset consists of many pairs $\{B_n, d_n\}, n = 1, ..., N$, where N is the number training pairs. Each bag B_n consists of several instances $B_n = \{x_{n1}, ..., x_{nL}\}$, where x_{nl} is an instance in a bag and L is the number of instances in each bag. We denote d_n as the label of the n-th bag. In Audio Set classification, a bag is a collection of L features from an audio clip. Each instance $x_{nl} \in \mathbb{R}^M$ is a feature, where M is the dimension of the feature. The label of a bag is $d_n \in \{0, 1\}^K$ where K is the number of audio classes and 0 and 1 represent the negative and positive label, respectively. For a specific class k, when the label of the n-th bag $d_{nk} = 1$ then $\exists x_{nl} \in B_n$ so that x_{nl} is positive. Otherwise if $d_{nk} = 0$ then $\forall x_{nl} \in B_n$ so that x_{nl} is negative. Assume we have a classifier f on each instance, we want to obtain a classifier F on each bag. There are several ways to obtain bag level classifier from instance level classifier described as follows.

2.1. Collective assumption

The *collective assumption* [10] states that all instances in a bag contribute equally and independently to the bag's label. Under this assumption, the bag level classifier F is obtained by using the sum as the aggregation rule:

$$F(B) = \frac{1}{L} \sum_{x_l \in B} f(x_l). \tag{1}$$

The collective assumption is simple and assumes that the instances contribute equally and independently to the bag-level class labels. However the collective assumption assumes all the instances inherit the label from its corresponding bag, which is not the same as the MIL assumption.

2.2. Maximum selection

The *maximum selection* [11] states that the prediction of a bag is the maximum classification value of each instance in the bag described as follows:

$$F(B) = \max_{x_l \in B} f(x_l). \tag{2}$$

Maximum selection has been used in audio tagging using convolutional neural networks (CNNs) [12] and audio event detection using weakly labelled data [13]. Maximum selection corresponds to a global max pooling layer [12] in a convolutional neural network. Maximum selection performs well in audio tagging [12] but is sometimes inefficient in training because only one instance with the maximum value in a bag is used for training, and the gradient will only be computed from the instance with the highest classification value.

2.3. Weighted collective assumption

The *weighted collective assumption* is a generalization of the collective assumption, where a different weight $w(x)$ is allowed for each instance x [9]:

$$F(B) = \frac{1}{\sum_{x \in B} w(x)} \sum_{x \in B} w(x) f(x). \tag{3}$$

The weighted collective assumption asserts that each instance contributes independently but not necessarily equally to the label of a tag. This is achieved by incorporating a weight function $w(x)$ into the collective assumption. Equation (3) has the same form as our joint detection-classification (JDC) model [14] and our attention model [6] proposed for audio tagging and sound event detection. The difference is that the work in [14, 6] model both $w(x)$ and $f(x)$ using neural network.

3. ATTENTION A PROBABILISTIC PERSPECTIVE

Although Equation (3) has been used in many previous works [9, 14, 6], the explanation for this equation is not clearly presented. In this paper we explain this attention model in Equation (3) from a probabilistic perspective, which is helpful to guide the selection of f and w in Equation (3).

3.1. Measure space

For any instances x in a bag, they should contribute differently to the classification of a bag. In MIL, a bag is labelled positive if at least one instance in the bag is positive. To solve this problem, the positive instances should be attended to and the negative instances should be ignored. We first assign a *measure* on each $x \in \Omega$ where Ω is a set x laid in, for example Euclidean space. To assign the measure on each instance x, we introduce the *measure space* [15] in probability theory.

Definition 1. Let Ω be a set, $\mathscr{F}$ a Borel field [15] of subsets of Ω. A *measure* μ on $\mathscr{F}$ is a numerically valued set function with domain $\mathscr{F}$, satisfying the following axioms:
1. $\forall E \in \mathscr{F} : \mu(E) \geq 0$
2. If $\{E_j\}$ is a countable collection of disjoint sets in $\mathscr{F}$, $\mu(\bigcup_j E_j) = \sum_j \mu(E_j)$, then we call the triple $(\Omega, \mathscr{F}, \mu)$ a *measure space*.

In addition, if we have:
3. $\mu(\Omega) = 1$
then we call the triple $(\Omega, \mathscr{F}, \mu)$ a *probability space*.

3.2. Probability space

When classifying a bag, different instances contribute differently. We define a probability space for each bag B_n for each class k. As $B_n \subset \Omega$, we may define a probability space $(B_n, \mathscr{F}_{B_n}, p_{nk})$ on B_n where $\mathscr{F}_{B_n} = \mathscr{F} \bigcap \mathscr{F}(B_n)$ and $\mathscr{F}(B_n)$ is the Borel filed of the set B_n. The probability
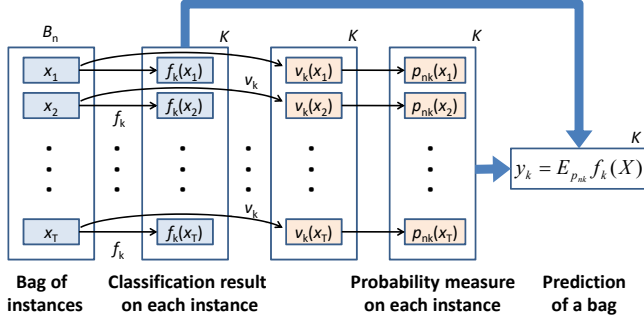
Fig. 1. Attention model a probabilistic perspective where f_k is the classification result on each instance and p_{nk} is the probability measure of each instance in a given bag. The prediction is the expectation of f_k with respect to the probability measure p_{nk}.

measure p_{nk} on B_n satisfies $\sum_{x \in B_n} p_{nk}(x) = 1$, so Definition 1 Axiom 3 is satisfied. We call $(B_n, \mathscr{F} \bigcap B_n, p_{nk})$ a probability space for the k-th class. For an instance x in a bag, the closer $p_{nk}(x)$ to 1 the more this instance is attended. The closer $p_{nk}(x)$ to 0 the less this instance is attended.

3.3. Expectation

Assume for the k-th class, the classification prediction and the probability measure on each instance $x \in B_n$ are $f_k(x)$ and $p_{nk}(x)$, respectively. To obtain the classification result on the bag B_n, we apply the expectation of the classification result $f_k(\cdot)$ with respect to the probability measure p_{nk}:

$$F(B_n) = E_{p_{nk}}(f_k(X)) = \sum_{x \in B_n} f_k(x) p_{nk}(x) \qquad (4)$$

where X is a random variable. Equation (4) represents the instances $x \in B_n$ contributes differently to the classification of the bag B_n. The probability measure $p_{nk}(\cdot)$ controls how much an instance x is attended. Large p_{nk} and small p_{nk} represents the instance is attended and ignored, respectively.

3.4. Modeling attention

For a dataset with $\Omega = \mathbb{R}^M$. A mapping $f_k : \mathbb{R}^M \mapsto [0, 1]$ is used to model the presence probability of the k-th class of an instance x. On the other hand, modeling the probability measure $p_{nk} : \mathbb{R}^M \mapsto [0, 1]$ is difficult because of the constraint that the sum of the probability of the instances in a bag should be equal to 1:

$$\sum_{x \in B_n} p_{nk}(x) = 1. \qquad (5)$$

So instead of modeling p_{nk} directly, we start from modeling μ_k in the measure space $(\mathbb{R}^M, \mathscr{F}, \mu_k)$ because in the measure

space μ_k does not need to satisfy Definition 1, Axiom 3. To model μ_k, we use a mapping $v_k : \mathbb{R}^M \mapsto \overline{\mathbb{R}}^+$, where $\overline{\mathbb{R}}^+ = \mathbb{R}^+ \bigcup \{0\}$. Then for each bag B_n and $x \in B_n$, we may define the probability measure of any instance x of the k-th class as:

$$p_{nk}(x) = \mu_k(\{x\})/\mu_k(B_n) = v_k(x)/\sum_{x \in B_n} v_k(x) \qquad (6)$$

where $\mu(\{x\})$ and $\mu(B_n)$ are the measure of $\{x\}$ and B_n, respectively. From Definition 1 Axiom 2, $\mu_k(B_n)$ can be calculated by $\mu_k(B_n) = \sum_{x \in B_n} \mu_k(\{x\})$. So the constraint in Equation (5) is satisfied. After modeling f_k and p_{nk}, the prediction of the k-th class can be obtained by using Equation (4). The framework of the attention model is shown in Fig. 1.

3.5. Mini batch balancing

The Audio Set dataset is highly unbalanced. Some classes have tens of thousands samples while other classes only contain hundreds of samples. We therefore propose a mini batch balancing strategy, where the occurrence frequency of training samples of the different classes in a mini-batch are kept the same.

4. EXPERIMENTS

4.1. Dataset

We experiment on the Audio Set dataset [5]. Audio Set contains over 2 million 10 seconds audio clips extracted from YouTube videos. Audio Set consists of 527 classes of audio with a hierarchy structure. The original waveform of the 2 million audio clips are not published. Instead, we use the published bottleneck feature vectors extracted from the embedding layer representation of a deep CNN trained on the YouTube-100M dataset [16]. The bottleneck feature vectors are extracted at one feature per second, that is, there are 10 features in an 10 seconds audio clip. Then the bottleneck feature vectors are post-processed by a principle component analysis (PCA) to remove the correlations and only the first 128 PCA coefficients are kept.

4.2. Model

The source code of this system is available here[2]. We apply a simple fully connected deep neural network to verify the effectiveness of the proposed attention model. We first apply fully connected layers on the input feature vectors to extract high level representation. We call this mapping as *embedded mapping* and denote as g. We call $h = g(x)$ as *embedded instance*. The embedded mapping g is modeled by three fully connected layers, with 500 hidden units in each layer followed by ReLU [17] non-linearity and dropout [18] rate of

---

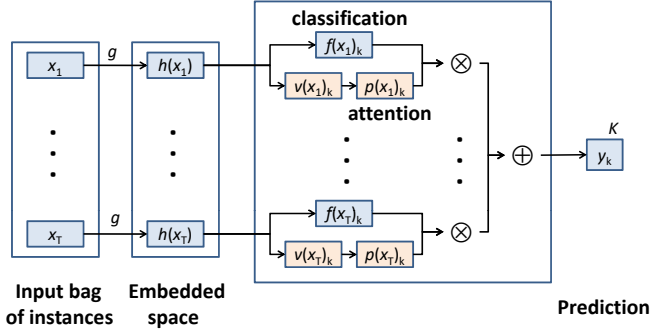[2]https://github.com/qiuqiangkong/ICASSP2018_audioset

Fig. 2. Model for Audio Set classification. The input space is mapped to an embedded space followed by a classification and an attention (probability measure) branch. Final prediction is the expectation of the classification output with respect to the learned probability measure.

0.2 to reduce the risk of over-fitting. These configurations are chosen empirically. Then we model the classifier f_k and the measure v_k on each embedded instance h by the following equation:

$$f_k(h) = \sigma(W_f h + b_f)_k \qquad (7)$$

$$v_k(h) = \phi(W_v h + b_v)_k \qquad (8)$$

where σ is sigmoid non-linearity $f(z) = 1/(1 + e^{-z})$. The sigmoid non-linearity ensures that the probability $f_k(\cdot)$ is between 0 and 1. The non-linearity ϕ can be any non-negative function and we investigate ReLU [17], sigmoid and softmax functions in our experiment.

Then we may obtain p_{nk} in the n-th bag by:

$$p_{nk}(x) = v_k(g(x))/\sum_{x \in B_n} v_k(g(x)) \qquad (9)$$

Finally the prediction of the k-th event in bag B_n is obtained by using Equation (4).

4.3. Experiment analysis

We evaluate using mean average precision (mAP), area under curve (AUC) and d-prime used in [5]. These values are computed for each of the 527 classes and averaged across the 527 classes to obtain the final mAP, AUC and d-prime. Higher mAP, AUC and d-prime lead to better performance.

Table 1 shows the results of with and without data balancing strategy using collective assumption in Equation (1). The data balancing strategy is described in Section 3.5. Table 1 shows using balancing strategy performs better than without data balancing strategy in all of mAP, AUC and d-prime.

Table 2 shows the results of modeling the measure function $v_k(\cdot)$ using different non-negative functions including ReLU, sigmoid and softmax functions. Softmax non-negative

performs slightly better than sigmoid non-negative and better than ReLU non-negative function.

Table 3 shows the comparison of different pooling strategies. Average pooling and max pooling along time axis are described in Equation (1) and (2), respectively. The Google baseline uses a simple fully connected DNN [5]. Table 3 shows that RNN with global average pooling performs better than Google baseline. Using DNN with attention achieves better performance than Google baseline and RNN.

Table 1. Classification result with and without data balancing strategy.

	mAP	AUC	d-prime
w/o balancing	0.275	0.957	2.429
with balancing	**0.296**	**0.960**	**2.473**

Table 2. Classification results of measure $v_k(\cdot)$ modeled by ReLU, sigmoid and softmax functions.

	mAP	AUC	d-prime
DNN ReLU attention	0.306	0.961	2.500
DNN sigmoid attention	0.326	0.964	2.547
DNN softmax attention	**0.327**	**0.965**	**2.558**

Table 3. Classification results with different pooling strategy.

	mAP	AUC	d-prime
DNN max pooling	0.284	0.958	2.442
DNN avg. pooling	0.296	0.960	2.473
Google baseline	0.314	0.959	2.452
RNN avg. pooling	0.325	0.960	2.480
DNN softmax attention	**0.327**	**0.965**	**2.558**

5. CONCLUSION

In this paper, an attention model in audio classification is explained from a probability perspective. Both the classifier and the probability measure on each instance are modeled by a neural network. We apply fully connected neural network with this attention model on Audio Set and achieves mAP of 0.327 and AUC of 0.965 outperforming the Google baseline and recurrent neural network. In the future, we will explore more on modeling probability measure using different non-negative functions.

6. ACKNOWLEDGEMENT

7. REFERENCES

[1] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *EUSIPCO*. IEEE, 2016, pp. 1128–1132.

[2] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of DCASE2017 Workshop*.

[3] Y. Xu, Q. Kong, Q. Huang, W. Wang, and Mark D. Plumbley, "Convolutional gated recurrent neural network incorporating spatial features for audio tagging," in *IJCNN*. IEEE, 2017, pp. 3461–3466.

[4] Y. Xu, Q. Huang, W. Wang, P. Foster, S. Sigtia, PJB Jackson, and MD Plumbley, "Unsupervised feature learning based on deep models for environmental audio tagging," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 25, no. 6, pp. 1230–1241, 2017.

[5] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017, pp. 776–780.

[6] Y. Xu, Q. Kong, Q. Huang, W. Wang, and Mark D. Plumbley, "Attention and localization based on a deep convolutional recurrent model for weakly supervised audio tagging," in *INTERSPEECH*. IEEE, 2017, pp. 3083–3087.

[7] Y. Xu, Q. Kong, W. Wang, and Mark D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," *arXiv preprint arXiv:1710.00343*, 2017.

[8] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Advances in Neural Information Processing Systems*, 1998, pp. 570–576.

[9] J. Foulds and E. Frank, "A review of multi-instance learning assumptions," pp. 25(1):570–576, 2010.

[10] X. Xu, "Statistical learning in multiple instance problems," in *Master Thesis, University of Waikato, 2003*.

[11] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artificial Intelligence*, pp. 25(1):570–576, 2013.

[12] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," *arXiv preprint arXiv:1606.00298*, 2016.

[13] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 1038–1047.

[14] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "A joint detection-classification model for audio tagging of weakly labelled data," in *ICASSP*. IEEE, 2017, pp. 641–645.

[15] K. L. Chung, "A course in probability theory," *Academic Press*, pp. 16–21, 2001.

[16] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, and M. Slaney, "CNN architectures for large-scale audio classification," in *ICASSP*. IEEE, 2017, pp. 131–135.

[17] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning*. IEEE, 2010, pp. 807–814.

[18] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," in *Journal of Machine Learning Research*. IEEE, 2014, pp. 1929–1958.