AUDIO EVENT DETECTION FROM WEAK ANNOTATIONS: WEIGHTED GRU VERSUS MULTI-INSTANCE LEARNING

Léo Cances, Thomas Pellegrini, Patrice Guyot,

IRIT, Université de Toulouse, CNRS, Toulouse, France
{leo.cances,thomas.pellegrini,patrice.guyot}@irit.fr

ABSTRACT

In this paper, we address the detection of audio events in domestic environments in the case where some of the data is weakly annotated. By weak labels, we mean that a single label describes a whole audio recording. We report experiments in the framework of the task four of the DCASE 2018 challenge. The objective is twofold: detect audio events (multi-categorical classification at recording level), localize the events precisely within the recordings. We explored two approaches:1) a "weighted-GRU" one, in which we train a Convolutional Recurrent Neural Network for classification and then exploit its frame-based predictions at the output of the time-distributed dense layer to perform localization. We propose to lower the weight of the cell states to avoid predicting a same score allover a recording. 2) An approach inspired by Multi-Instance Learning, in which we train a CNN to give predictions at frame-level, using a custom loss function based on the weak label and statistics of the frame-based predictions.

Index Terms— One, two, three, four, five

1. INTRODUCTION

The guidelines given below, including complete descriptions of the fonts, spacing,and related information for producing your proceedings manuscripts, are critical to produce the DCASE 2018 proceedings with a more uniform look. If you have any questions, please email them to dcase.workshop@gmail.com.

2. FORMATTING YOUR PAPER

All manuscripts must be submitted electronically as PDF files. All manuscripts must be formatted for white US letter paper (8.5 $\times$ 11 inches). Please do **not** use A4-size papers. All printed material, including text, illustrations, and charts, must be kept within a print area of 7.0 inches (178 mm) wide by 8.9 inches (226 mm) high. Do not write or print anything outside the print area. The top margin must be 1 inch (25 mm), except for the title page, and the left margin must be 0.75 inch (19 mm). All *text* must be in a two-column format. Columns are to be 3.29 inches (83.5 mm) wide, with a 0.31 inch (8 mm) space between them. Text must be fully justified.

3. NUMBER OF PAGES

You are allowed a total of up to 4+1 pages for your DCASE 2018 Workshop submission, with up to 4 pages for technical content including figures and possible references, with the optional 5th page containing only references.

4. PAGE TITLE SECTION

The paper title (on the first page) should begin 0.98 inches (25 mm) from the top edge of the page, centered, completely capitalized, and in Times 14-point, boldface type. The authors' name(s) and affiliation(s) appear below the title in capital and lower case letters. Papers with multiple authors and affiliations may require two or more lines for this information.

5. TYPE-STYLE AND FONTS

We strongly encourage you to use Times-Roman font. In addition, this will give the proceedings a more uniform look. Use a font that is no smaller than nine point type throughout the paper, including figure captions.

In nine point type font, capital letters are 2 mm high. **If you use the smallest point size, there should be no more than 3.2 lines/cm (8 lines/inch) vertically.** This is a minimum spacing; 2.75 lines/cm (7 lines/inch) will make the paper much more readable. Larger type sizes require correspondingly larger vertical spacing. Please do not double-space your paper. True-Type 1 fonts are preferred.

The first paragraph in each section should not be indented, but all the following paragraphs within the section should be indented as these paragraphs demonstrate.

6. MAJOR HEADINGS

Major headings, for example, "1. Introduction", should appear in all capital letters, bold face if possible, centered in the column, with one blank line before, and one blank line after. Use a period (".") after the heading number, not a colon.

6.1. Subheadings

Subheadings should appear in lower case (initial word capitalized) in boldface. They should start at the left margin on a separate line.

6.1.1. Sub-subheadings

Sub-subheadings, as in this paragraph, are discouraged. However, if you must use them, they should appear in lower case (initial word capitalized) and start at the left margin on a separate line, with paragraph text beginning on the following line. They should be in italics.

7. PAGE NUMBERING, HEADER, AND FOOTER

Please do **not** paginate your paper. In addition, please do **not** change and remove the header and footer.

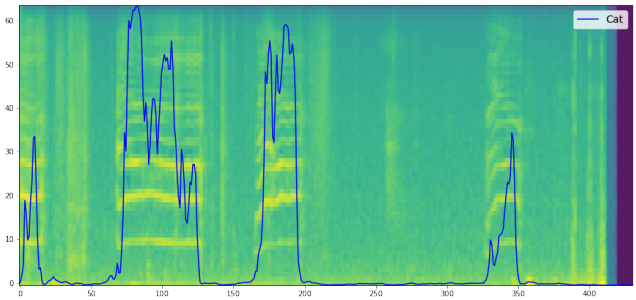

Figure 1: Example of a figure with experimental results.

8. ILLUSTRATIONS, GRAPHS, AND PHOTOGRAPHS

Illustrations must appear within the designated margins. They may span the two columns. If possible, position illustrations at the top of columns, rather than in the middle or at the bottom. Caption and number every illustration. All halftone illustrations must be clear black and white prints. Colors may be used, but they should be selected so as to be readable when printed on a black-only printer.

Since there are many ways, often incompatible, of including images (e.g., with experimental results) in a LaTeX document, an example of how to do this is presented in Fig. 3.

9. EQUATIONS

Equations should be placed on separate lines and consecutively numbered with equation numbers in parentheses flush with the right margin, as illustrated in (1) that gives the homogeneous acoustic wave equation in Cartesian coordinates [1],

$$\Delta^2 p(x,y,z,t) - \frac{1}{c^2}\frac{\partial^2 p(x,y,z,t)}{\partial t^2} = 0, \qquad (1)$$

where $p(x,y,z,t)$ is an infinitesimal variation of acoustic pressure from its equilibrium value at position (x,y,z) and time t, and where c denotes the speed of sound.

Symbols in your equation should be defined before the equation appears or immediately following. Use (1), not Eq. (1) or equation (1), except at the beginning of a sentence: "Equation (1) is ..."

10. FOOTNOTES

Use footnotes sparingly and place them at the bottom of the column on the page on which they are referenced. Use Times 9-point type, single-spaced. To help your readers, avoid using footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence).

11. REFERENCES

List and number all bibliographical references at the end of the paper. The references should be numbered in order of appearance in the document. When referring to them in the text, type the corresponding reference number in square brackets as shown at the end of this sentence [2], [3]. For LaTeX users, the use of the BibTeX style file IEEEtran.bst is recommended, which is included in the LaTeX paper kit available from the DCASE 2018 website [4].

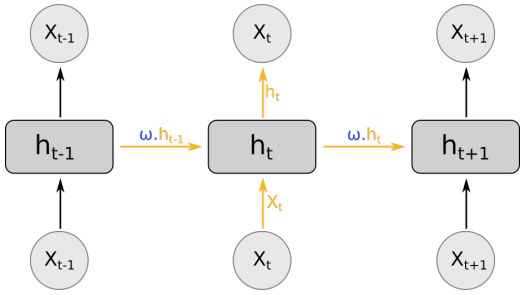

Figure 2: Insertion of a weight ω on the recurrent links between GRU's cells

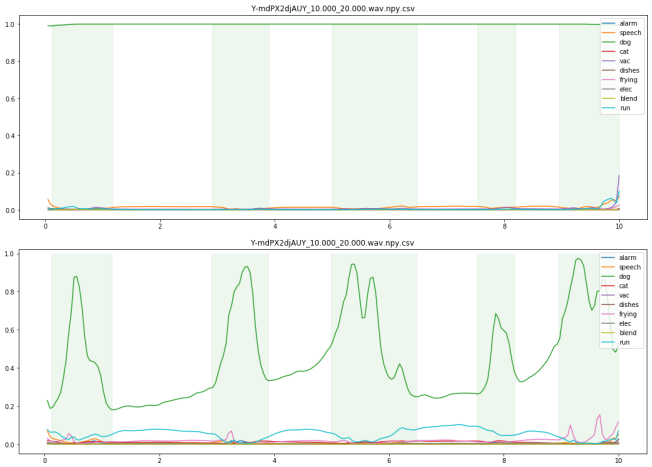

Figure 3: The curves are the outputs of the last time distributed layer. Green for dog. The vertical rectangle come from the ground thruth and reprense the segments where the dog should be detected. *Top:* Localization results of the CRNN with GRU. The class dog is detected but during the entire clip. *Bottom:* The prediction of the CRNN with WGRU on the same file with a temporal weight of 0.25. The dog is detected and properly localated.

12. ACKNOWLEDGMENT

The preferred spelling of the word acknowledgment in America is without an "e" after the "g." Try to avoid the stilted expression, "One of us (R. B. G.) thanks ..." Instead, try "R.B.G. thanks ..." Put sponsor acknowledgments in the unnumbered footnote on the first page.

13. WEIGHTED GATE RECURENT UNIT

13.1. General RNN [5]

In principle, RNN are more suitable for capturing relationships among sequential data types. This makes them efficient for localization of sound event due to their temporal relation. They can be define as below:

$$h_t = g(Wx_t + U_{t-1} + b) \qquad (2)$$

where x_t is the (external) m-dimensional input vector at time t, $_t$ the n-dimensional hidden state, g is the (point-wise) activation

function, such as the logistic function, the hyperbolic tangent function, or the rectified Linear Unit (ReLU) CITATIOn HERE, and , are the appropriately sized parameters (two weights and bias). Specifically, in this case, W is an $n \times m$ matrix, abd U is an $n \times n$ matrix, and b is an $n \times 1$ matrix (or vector).

13.2. Weighted RNN

During the different experimentation done, the CRNN model using a GRU layer show that stationary sounds (vacuum cleaner, running water) were properly detected and classified but, more occasional one (speech, dog, cat) were not. The idea behind the weighted GRU is to reduce the impact of the temporality between the recurrent cells in order to increase the performance of the model to locate such event.

The introduction of the weight ω is done as bellow or can be seen on Figure 2

$$h_t = \omega * g(W x_t + U_{t-1} + b) \tag{3}$$

13.3. Weight impact on localization

	Weight	F1	ER
Baseline	/	14.06 %	1.54
CRNN with GRU	/	6.68 %	2.55
CRNN with WGRU	0.50	4.69 %	2.92
CRNN with WGRU	0.30	8.24 %	3.18
CRNN with WGRU	0.20	11.35 %	3.37

14. REFERENCES

[1] E. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustic Holography*. London, UK: Academic Press, 1999.

[2] C. Jones, A. Smith, and E. Roberts, "A sample paper in conference proceedings," in *Proc. IEEE ICASSP*, vol. II, 2003, pp. 803–806.

[3] A. Smith, C. Jones, and E. Roberts, "A sample paper in journals," *IEEE Trans. Signal Process.*, vol. 62, pp. 291–294, Jan. 2000.

[4] http://dcase.community/workshop2018/.

[5] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks," *CoRR*, vol. abs/1701.05923, 2017. [Online]. Available: http://arxiv.org/abs/1701.05923