Pré-rapport:
Reconnaissance étages de batiment à partir des échantillons sonores

Massih-Reza Amini [a] and Junior N'nane [b]
[a] Department, University, City, Country
[b] Department, University, City, Country

Abstract

Keywords: Machine Learning; Sound; Classification format;

1 Introduction

Le but de ce projet est de pouvoir déterminer, à partir d'un enregistrement sonore, l'étage du batiment dans lequel l'enregistrement a été effectué.

Nous nous servirons de méthodes de machine Learninge et de traitement de signal.

2 Notions

Le son peut se définir comme étant une sensation auditive provoquée par des vibrations de l'air. VanDerveer proposa les critères suivants pour définir un son de l'environnement :

- un événement le produit ;
- il est le reflet d'un ou d'une série d'événements causaux ;
- son traitement est plus compliqué qu'un son pur généré en laboratoire ;
- il ne relève pas de la reconnaissance de parole (plus généralement de communication selon sa définition, la parole n'étant pas le seul type de son nous permettant de communiquer, par exemple les interjection ne font pas partie de la parole, mais sont pour autant un moyen de communication).

Les sons d'un environnement peuvent être catégorisés suivant plusieurs catégories: bruit, son naturel, son artificiel, parole, musique. Le son peut qualifé d'impulsionnel ou stationnaire, mais aussi périodique ou non-périodique.

L'enregistrement sonnore se présente comme un vecteur unidimensionnel possédant un grand nombre d'échantillons parseconde. Il est possible d'utiliser une fenêtre pour observer un signal sur une durée finie, on le multiplie par une fonction fenêtre d'observation. La fonction de Hammming permet d'améliorer les lobes secondaires.

Une fois le fénétrage effectué, il est possible d'extraire les paramètres acoustiques divisés en deux catégories: temporels et fréquentiels.

Les paramètres temporels peuvent s'obtenir avec des méthodes telles que : ZCR. Les paramètres fréquentiels eux s'obtiennent avec les méthodes telles que MFCC (Mel-Frequency Cepstral Coefficients) et SFRF (Sectral Rollof Point) et SC (Spectral Centroid).

La transformation de Fourier est une opération qui transforme une fonction intégrable sur R en une autre fonction, décrivant le spectre fréquentiel de cette dernière. elle permet d'obtenir une représentation Temps-Fréquence d'un enregistrement sonore, appelée spectrogramme.

La classification d'évenements sonores se fait en général avec des méthodes de Machine Learning, on peut utiliser des méthodes supervisées et non-supervisées

3 Methodes à explorer

Les methodes à explorer sont:

- entrainement de modèle à modalité unique
- early fusion (avec un modèle U-net pour apprendre une representation conjointe)
- late fusion (entrainer un modèle "on top" des modèles déjà existants

Les sources utilisées étant différentes mon parle de "multimodal learning", où nous allons entrainer des modèles à partir d'entrées differentes et les fusionner. la stratégie la plus utilisée consiste à fusionner informations au niveau des fonctionnalités, également appelées fusion précoce. L'autre approche est le niveau de décision fusion ou fusion tardive qui fusionne plusieurs dans l'espace sémantique.

3.1 Niveaux de fusion

3.1.1 Feature level ou early fusion

Dans l'approche du niveau des features ou de la fusion précoce, les fonctionnalités extraites des données d'entrée sont d'abord combinées puis envoyées en entrée d'une seule unité d'analyse (AU) qui effectue la tâche d'analyse.

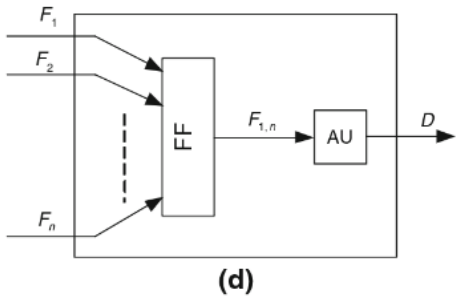

Figure 1: Early fusion, Atrey et al. (2010)

La fusion au niveau des fonctionnalités est avantageuse en ce qu'elle peut utiliser la corrélation entre plusieurs caractéristiques de différentes modalités à un stade précoce, ce qui aide à mieux accomplir de la tâche. De plus, il ne nécessite qu'un seul apprentissage phase sur le vecteur de caractéristique combiné. Mais, les features à fusionner doivent être représentées dans le même format avant la fusion.

3.1.2 Decision level ou late fusion

Dans l'approche du niveau de décision ou de la fusion tardive, des unités de décision fournissent d'abord les décisions locales D1 à Dn qui sont obtenus à partir des caractéristiques individuelles F1 à Fn . La les décisions locales sont ensuite combinées à l'aide d'une fusion de décision.

La stratégie de fusion au niveau décisionnel présente de nombreux avantages sur la fusion de fonctionnalités. Par exemple, contrairement au niveau des features, où les caractéristiques de différentes modalités (par ex. audio et vidéo) peuvent avoir des représentations différentes, le décisions (au niveau sémantique) ont généralement le même représentation. Par conséquent, la fusion des décisions devient plus facile. De plus, la stratégie de fusion

au niveau de la décision offre l'évolutivité (c'est-à-dire la mise à niveau ou la dégradation progressive) dans termes des modalités utilisées dans le processus de fusion, qui est difficile à réaliser dans la fusion au niveau des fonctionnalités. Une autre L'avantage de la stratégie de fusion tardive est qu'elle nous permet d'utiliser les méthodes les plus appropriées pour analyser chaque modalité, telle que le modèle de Markov caché (HMM) pour l'audio et prend en charge la machine vectorielle (SVM) pour l'image. Ceci offre beaucoup plus de flexibilité que la première fusion.

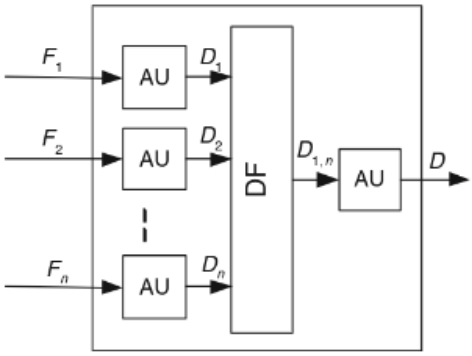

Figure 2: Late fusion, Atrey et al. (2010)

3.2 Methodes de fusion

3.2.1 Règles

Les méthodes basées sur les règles consistent à définir "à la main" un algorithme d'interpretation des décisions ou des features fournis par les modèles antécedants.

3.2.2 Fusion linéaire

La fusion pondérée linéaire est l'une des plus simples et des plus méthodes largement utilisées. Dans cette méthode, les informations obtenus à partir de différentes modalités est combiné dans un linéaire mode. Pour combiner les informations, on peut attribuer des poids de normalisation aux différentes modalités. En littérature, il y a diverses méthodes de normalisation du poids telles que min– max, mise à l'échelle décimale, z-score, estimateurs tanh et sigmoïde fonction.

3.3 Association des modalités

Il existe de nombreuses façon d'évaluer l'impact de l'ajout d'une modalité ou la quantité d'information partagée d'une modalité à une autre.

On note:

- calcul de la correlation entre les features
- calcul de l'information mutuelle (voir page 366 fusion_survey)

La fusion des caractéristiques se traduit généralement par un grand vecteur de caractéristiques, qui devient un goulot d'étranglement pour un particulier tâche d'analyse. C'est ce qu'on appelle la malédiction de la dimensionnalité. Pour surmonter ce problème, différentes techniques de réduction de données niques sont appliqués pour réduire le vecteur de caractéristiques.

- Latent semantic analysis (LSI,SFA) ces methodes sont utilisés pour obtenir une representation conjointe des différents espaces de feature.
- PCA, SVD, LDA

3.4 Selection des modalités

4 Methodologie

Etant donné 03 modalités : MFCC,raw audio image, spectrogramme, il faut prédire le niveau de l'immeuble.

´

5 Données

Les données mises à notre disposition sont de trois types:

- spectrogramme/sonogramme (image)
- fichier audio brut (.wav)
- enregistrement amplitude (image)

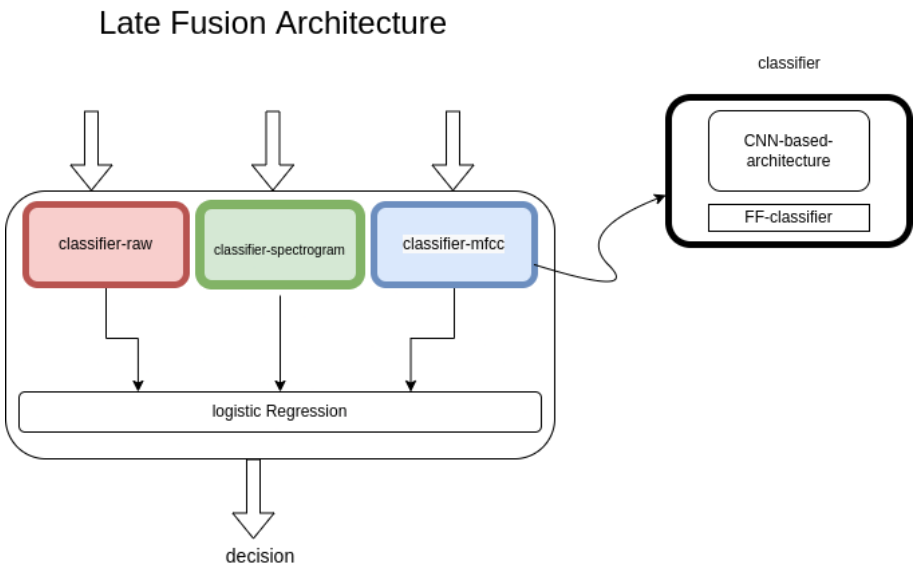

Figure 3: Late fusion

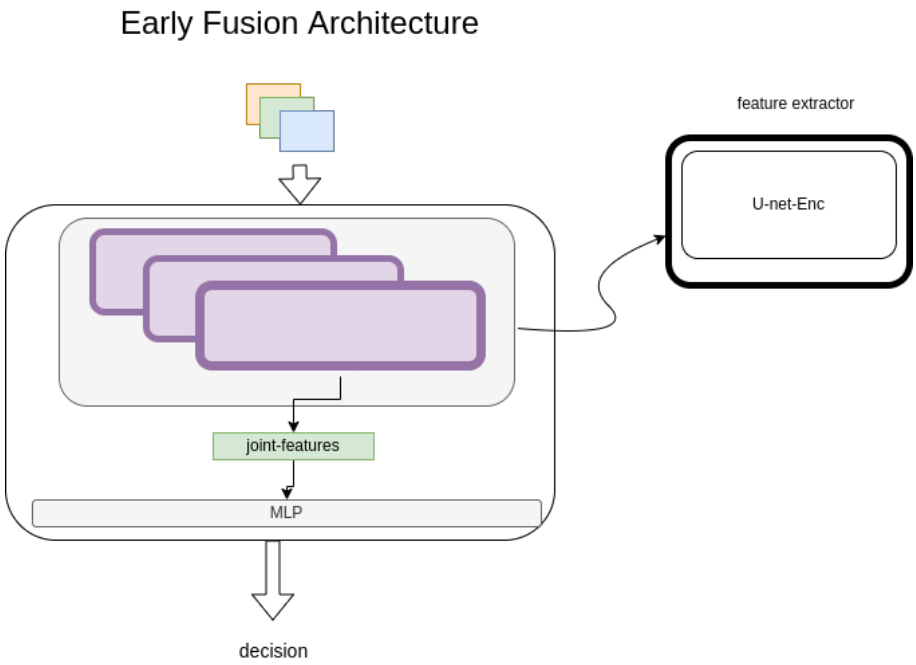

Figure 4: Early fusion

6 Outils/Ressources disponibles

Ici, nous allons lister les outils python disponibles pour traiter les enregistrements sonores notamment ceux intégrés dans la librairie pytorch.

MFCC [pytorch]

CNN with Pytorch using Mel features[pytorch]

How to Detect COVID-19 Cough From Mel Spectrogram Using CNN [keras]

Audio feature extraction [pytorch]

Environment Sound Event Classification With a Two-Stream Convolutional Neural Network

7 Questions

- Dans le cas d'une fusion early, est-ce qu'on va utiliser le même U-net Ronneberger et al. (2015) pour obtenir les represenations pour tous les modes?
- De ce que je lis l'approche early, signifie de créer un unique modèle qui donne les representations pour tous les modes, ne serait-ce pas plus interessant d'avoir un modèle pour extraire les features de chaque modalité individuelement ?
-

8 Travaux relatifs

Ici, nous presentons des travaux similaires (utilisant le spectrograme/sonograme et la representation MFCC pour de la classification).

Honk: A PyTorch Reimplementation of Convolutional Neural Networks for Keyword Spotting

Environmental sound classification using temporal-frequency attention based convolutional neural network

Rethinking CNN Models for Audio Classification

CFA

Fast environmental sound classification based on resource adaptive convolutional neural network

IMPORTANT Environment Sound Event Classification With a Two-Stream Convolutional Neural Network

References

Atrey, P. K., M. A. Hossain, A. E. Saddik, and M. S. Kankanhalli (2010). Multimodal fusion for multimedia analysis: a survey.

Ronneberger, O., P. Fischer, and T. Brox (2015). U-net: Convolutional networks for biomedical image segmentation. *CoRR abs/1505.04597*.