Rapport TD et TP

UCE Explicabilité et Interprétabilité
Jean-François Bonastre

Audran BERT
M2 Intelligence Artificielle

21/12/2022

SOMMAIRE

Il est préférable d'ouvrir la version .pptx du rapport car dans la version PDF les enregistrements audios ne fonctionnent pas!

1) Introduction

- Dans ce document, nous allons réaliser un audit de l'explicabilité de la solution de biométrie médicale.
- Nous allons commencer par présenter succinctement le système tel que nous l'avons compris, puis nous allons donc décrire les potentiels forces et faiblesses des différents points ainsi que des propositions d'améliorations.
- Nous allons ensuite présenter différentes méthodes d'explicabilité et plus spécifiquement, nous allons nous concentrer sur l'explicabilité post-hoc (sans besoin de re entrainer le modèle).
- Enfin nous conclurons en résumant l'intérêt de chacune d'elles
- Dans les annexes, nous présenterons une implémentation de l'approche par Counterfactual ainsi qu'une analyse du modèle utilisé

2) Présentation du système analysé

- C'est un système d'authentification à travers la voix de l'utilisateur. L'utilisateur a un login et sa voix pour se connecter à l'application. L'application permet d'accéder aux informations d'une réservation et permet de modifier une réservation (billet) de train/avion jusqu'à 200€. Le créateur de la solution est rétribué à hauteur de 0.2€/par transaction → 0.02€ à 10 Millions de transaction.
- Le système a donc 3 entrées :
  - Un enregistrement audio avec de la parole d'une durée de 3 à 6 sec en majorité. Toutes nationalités, toutes langues, pas de contrainte de contenu, vérification de non-silence, pas de contrainte sur l'environnement. Utilisation mains occupées.
  - Un numéro utilisateur soit en reconnaissant le téléphone qui appelle soit l'utilisateur le rentre.
  - Une signature : la signature de l'utilisateur est un audio de l'utilisateur qui essaye de se connecter. C'est une entrée qui a été stocké dans une database pour permettre l'authentification. Cette signature est donc faite une fois.
- Le système est constitué de :
  - Un extracteur de représentation qui a été entrainé avec le dataset Voxceleb (Anglais pas forcément des natifs, public, 6000 locuteurs, 300k enregistrements)
  - Un comparateur cosinus (distance cosinus)
  - Un score LLR
  - Une Znorm (normalisation avec des enregistrements du locuteur et des enregistrements dont on est sûr que c'est pas le bon locuteur = imposteur ; propre au locuteur)
  - Un module de décision avec seuil

3) Prototypes and criticisms

Cette technique est utilisable sur les vecteurs de représentation obtenu lors de l'entrainement/test. C'est une méthode qui permet de résumer un jeu de données. Elle permet de localiser des locuteurs très bien représenté et des très mal représenté. Cela permet de donner une idée de la réaction de l'extracteur de représentation à des locuteurs/audios. On peut potentiellement trouver des biais de représentation, les biais peuvent être des accents, le sexe, la langue, la durée d'enregistrement, l'environnement au moment de l'enregistrement. On peut appliquer cette technique sur les locuteurs ou sur les enregistrements directement. Pour plus de détails sur la sélection des prototypes voir [6].

- Les forces sont donc : un outil rapide et simple pour visualiser de potentiels biais et forces du système en repérant locuteurs représentatifs. Méthode flexible et simple pour « résumer » des données.
- Les faiblesses sont : une approche basée uniquement sur des exemples qui peut être fortement influencé par la mesure de distance par exemple. Le résultat dépend de la projection dans l'espace de dimensionalité.
- Les opportunités sont : le système contient un extracteur de caractéristiques, il est donc facile d'appliquer cette technique sur la sortie de l'extracteur.
- Les menaces sont : pour comparer les signatures entre elles, il faut accéder et utiliser les données des utilisateurs. Danger au niveau des réglementations en vigueur.

- Dans [5], ils montrent un système permettant de diversifier les prototypes en pénalisant ceux trop proches. Dans leur méthode, il est nécessaire d'entrainer un modèle ce qui ajoute de la complexité et du temps d'exécution.

3) Prototypes and criticisms

Si on avait le droit de comparer les clients les uns avec les autres, je calculerai les prototypes et criticisms sur les signatures des utilisateurs. Cela permettrai de repérer les utilisateurs qui sont mal représentés par l'extracteur et donc repérer d'éventuels biais ou problèmes du modèle sur les clients directement. Par exemple voir si une langue est mal représentée vu que le système a été appris à partir de voxceleb qui est en anglais.

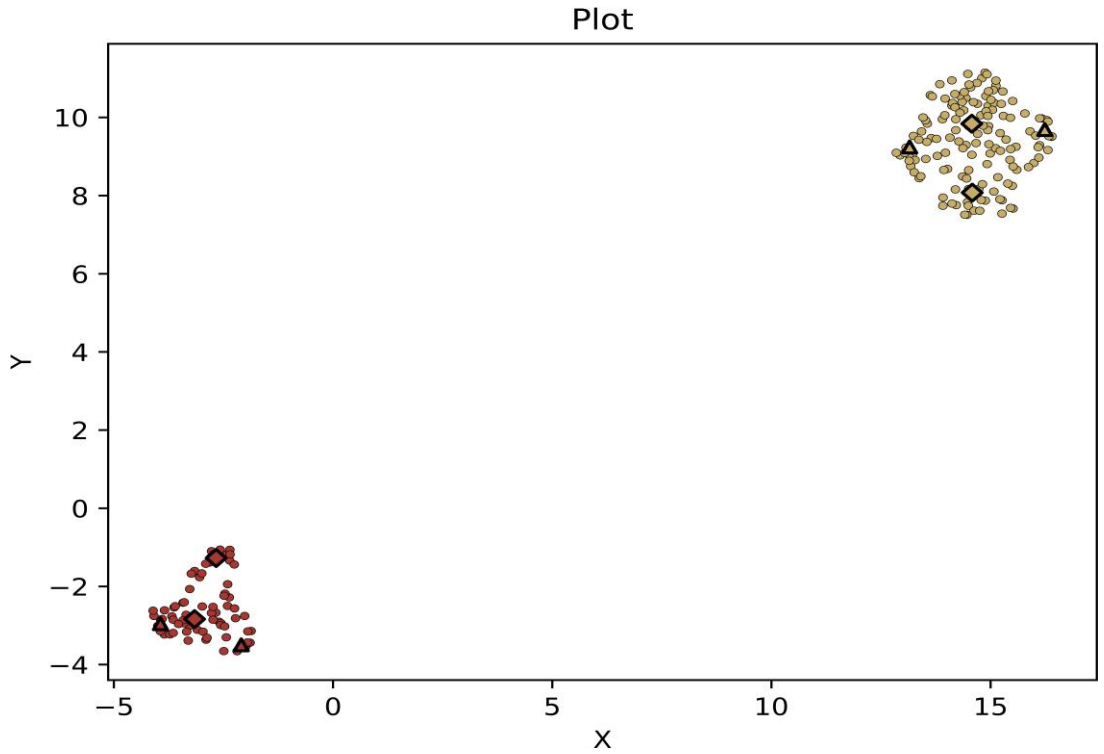


Figure 1. Exemple de calcul des prototypes and criticisms avec 2 coteurs, 2 prototypes et 2 criticisms par locuteur

4) Shapley Value Importance

- SHAP (Shapley Value Importance) permet de donner l'importance de chaque feature pour une prédiction. Cette méthode a été introduite dans [4], dans leur contexte, elle a permis de repérer des biais de genre et de les réduire tout en réduisant de peu la performance sur le dataset original (qui est surement biaisé). Dans notre application, cela pourrait être utilisé pour repérer des biais comme des accents, le sexe, la langue maternelle…
- On pourrait par exemple, comparer des audios provenant de femmes avec ceux d'hommes et regarder si certains pois du modèle sont actifs que dans un cas. De même pour les accents etc…

- Forces : Cela permet de trouver des poids du modèle portant un biais et d'essayer de le corriger simplement sans réapprendre le modèle (par exemple en neutralisant ou diminuant son importance)
- Faiblesses : Les poids relevés comme important dans un biais ne sont pas forcément vecteur du biais, il faut analyser par soi-même le résultat et il faut donc faire attention à ne pas interpréter trop vite. Couteuse en temps humains et d'exécutions.
- Opportunités : Beaucoup de biais possibles que l'on pourrait vérifier. La sortie de l'extracteur de représentation est un espace idéal pour utiliser cette technique
- Menaces : Beaucoup de biais possibles que l'on pourrait vérifier (oui pour moi c'est une opportunité et une faiblesse)

4) Relevance heatmaps

- Cette technique consiste à afficher sous forme de heatmap les parties de l'audio qui ont le plus contribué à la décision. Cette méthode est utilisée par-dessus une méthode telle que Lime ou Shap (voir [1]). Cette technique permet d'expliquer localement, c'est à dire entre deux audios et pas sur la globalité des données ce qui a influé sur la décision. Cela permet par exemple de comprendre pourquoi certains faux positifs ou faux négatifs ont eu lieu. A partir de là, on peut par exemple prévoir de potentiels vérification supplémentaire pour les cas litigieux détecté via cette méthode.

- Dans le cadre de cette approche, il faut donc choisir une technique capable d'expliquer localement une décision, j'ai donc parlé de LIME et de SHAP mais il y a aussi Saliency, LRP, DeepLIFT qui sont proposés dans l'article. Ces techniques servent à déterminer l'importance des features localement et donc à construire la heatmap. Dans l'article ils précisent que, SHAP est le plus robuste, il marche pour toutes les architectures mais que pour des architectures plus spécifique, les autres fonctionnent mieux. Excepté LIME qui performe le moins bien dû à la grande dimensionnalité des données.

- Forces : Approche qui permet de comprendre des erreurs du système de manière explicite : on peut aller voir ce qui a induit l'erreur en écoutant l'enregistrement.

- Faiblesses : Approche très locale et nécessite la mise en place d'un autre approche telle que LIME ou SHAP. Temps d'exécutions.

- Opportunités : Dans le contexte de l'identification par la voix il est très intéressant de comprendre ce qui a fait que le modèle a fait faux car domaine sensible.

- Menaces : Besoin d'accéder aux signatures et audio soumis pour tester l'identification, cela peut soulever des problèmes de droits. Approche peut être trop locale.

5) Counterfactual

- Le counterfactual est de base l'explication d'une boite noire en prenant la donnée la plus proche de celle que l'on regarde mais dont la prédiction est inversée. C'est facilement utilisable dans un contexte avec des données tabulaire où il suffit de trouver quel changement sur tel ou telle colonne aurait changé la décision. Par exemple de le cas d'un système de prêt bancaire où la personne a un revenu de 50k, cette approche va consister à dire qu'il fallait qu'il gagne 60k pour que sa demande soit acceptée. Dans [7], ils proposent une méthode générique pour générer des counterfactuals avec lesquels ils obtiennent de bons résultats. Cependant, dans le cas de la reconnaissance du locuteur, c'est beaucoup plus complexe, on ne peut pas juste dire qu'il aurait fallu qu'il gagne plus car on n'a pas des données tabulaires, on a un signal audio. Pour moi, dans ce contexte cette approche consiste à trouver quelle partie de l'audio il aurait fallu supprimer pour inverser la prédiction. On peut regarder quelles sont ces fenêtres qu'il faut supprimer et voir si il y a quelque chose de spécial dans ces fenêtres. Par exemple si c'est toutes des fenêtres où on entend le bruit d'une voiture, ce qui supposerai que le système a un biais ou que cela le perturbe. En appliquant ça, on peut extraire quelque chose de global (même si rien n'est vérifiable) et non plus local et trouvant que ce bruit de voiture influe sur tous les audios où il a un bruit de voiture.
- Forces : Facile à mettre en place, par exemple la méthode de knockout est assez simple et permet de mettre en valeur quelle partie de l'audio a le plus portée la décision. Elle permet une visualisation simple : il suffit d'écouter la version modifiée.
- Faiblesses : Ne résout pas le problème de la boite noire, la version modifiée peut elle-même être biaisée et ne pas correspondre à la réelle explication de la décision. La modification apportée peut-être elle-même biaisée, par exemple ajouter du vide ou du bruit peut influer sur la décision.
- Opportunités : Peut être appliqué aussi bien sur la sortie de l'extracteur que sur la suite du système.
- Menaces : Attention à ne pas généraliser des comportements locaux en comportements globaux. Bien comprendre et décrire ce qui est expliqué

5) Counterfactual - Implémentation

- L'approche par Counterfactual implémenté en TP a permis de montrer certains points à surveiller :
  - Le modèle est impacté par les bruits de fonds, les hésitations…
  - Il se base en partie sur certaines particularités vocales de la personne comme les accents
  - Certaines données sont probablement mal classées : certains audios semblant être labélisé à la mauvaise personne
- Cette analyse a été construite à partir de quelques exemples locaux et il est donc dangereux de la généraliser
- Voir les annexes pour plus de détails
- Le code source est donné avec ce rapport

6) Conclusion

- Il y a donc plusieurs techniques d'explicabilité post-hoc qui peuvent être utilisé dans ce contexte d'identification par la voix des utilisateurs.
- Les prototypes permettent notamment d'avoir quelque chose d'un peu global en extrayant des représentants et des outliers parmi les audios ou locuteurs.
- SHAP permet de donner l'importance de chaque feature pour une prédiction. Cela permet de trouver des poids biaisés et de réduire leur impact.
- L'approche par relevance heatmaps est très local et permet pour un audio donné de voir quelle partie de l'audio a servi dans la décision. Elle nécessite d'appliquer une autre technique avant comme SHAP.
- Enfin, l'approche par counterfactual permet elle aussi localement de trouver ce qu'il fait que le système a prédit telle ou telle chose.

- La page [2] permet d'avoir un bon aperçu des différentes méthodes et permet de trouver de bonnes références sur le sujet
- Un très bon livre [3] sur l'interprétabilité de manière plus général

- Il existe d'autres méthodes que je n'ai pas abordé comme :
  - LIME qui permet d'apprendre un modèle autour d'une prédiction. Je ne vois pas trop l'intérêt dans notre application
  - Influential Instances qui permet de récupérer des instances qui influencent la décision du modèle. Il me parait compliquer dans notre cade applicatif d'aller chercher les instances qui ont influé sur la décision

7) Bibliographie

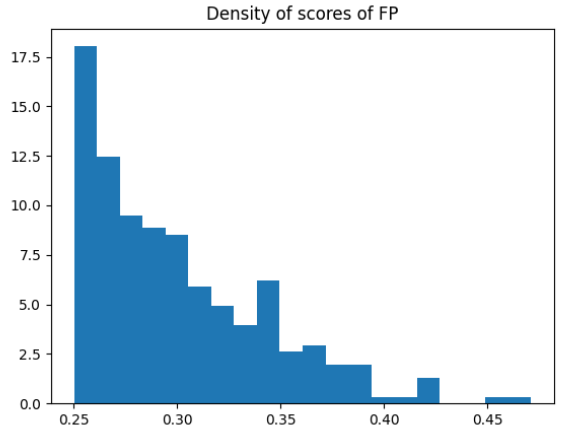
- [1] Udo Schlegel, Hiba Arnout, Mennatallah El-Assady, Daniela Oelke et Daniel A. Keim. « Towards a Rigorous Evaluation of XAI Methods on Time Series ». In : (2019). doi : 10.48550/ARXIV.1909.07082. url : https://arxiv.org/abs/1909.07082.
- [2] https://spectra.mathpix.com/article/2021.09.00007/demystify-post-hoc-explainability#user-content-Mothilal%20et%20al.,%202020
- [3] https://christophm.github.io/interpretable-ml-book/
- [4] Amirata Ghorbani et James Y Zou. « Neuron Shapley : Discovering the Responsible Neurons ». In : Advances in Neural Information Processing Systems. Sous la dir. De H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan et H. Lin. T. 33. Curran Associates, Inc., 2020, p. 5922-5932. url : https://proceedings.neurips.cc/paper/2020/file/41c542dfe6e4fc3deb251d64cf6ed2e4-Paper.pdf.
- [5] Alan H. Gee, Diego Garcia-Olano, Joydeep Ghosh et David Paydarfar. « Explaining Deep Classification of Time-Series Data with Learned Prototypes ». In : (2019). doi :10.48550/ARXIV.1904.08935. url : https://arxiv.org/abs/1904.08935.
- [6] Jacob Bien et Robert Tibshirani. « Prototype selection for interpretable classification ». In : The Annals of Applied Statistics 5.4 (déc. 2011). doi : 10 . 1214 / 11 - aoas495. url : https://doi.org/10.1214%2F11-aoas495.
- [7] Emanuele Albini, Antonio Rago, Pietro Baroni et Francesca Toni. « Relation-Based Counterfactual Explanations for Bayesian Network Classifiers ». In : Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20. Sous la dir. de Christian Bessiere. Main track. International Joint Conferences on Artificial Intelligence Organization, juill. 2020, p. 451-457. doi : 10.24963/ijcai.2020/63. url : https://doi.org/10.24963/ijcai.2020/63.

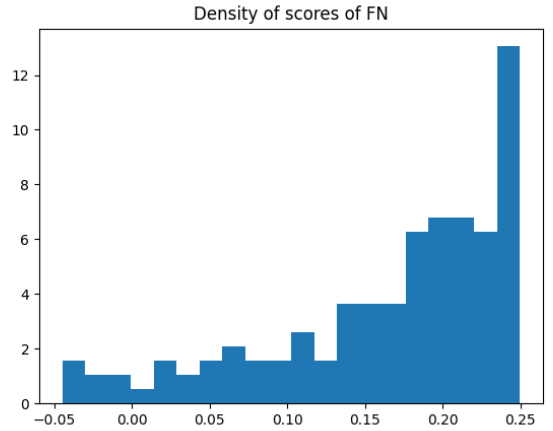
8) Annexes – Présentation TP

- Le modèle est un modèle de reconnaissance du locuteur : https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb
- Modèle appris sur Voxceleb (1 et 2)
- Le seuil (T) du modèle est 0,25 : en dessous, les enregistrements sont considérés comme étant de locuteurs différents et au-dessus, du même locuteur
- Nomenclature :
  - True Positive (TP) : le nombre de test de tar pour lesquels la décision a été positive (score> seuil)
  - True Negative (TN) : le nombre de test de non pour lesquels la décision a été négative (score<= seuil)
  - False Negative (FN) : le nombre de test de tar pour lesquels la décision a été négative (score<= seuil)
  - False Positive (FP) : le nombre de test de non pour lesquels la décision a été positive (score> seuil)
- Nombre de données :
  - FP = 276
  - TP = 18730
  - FN = 130
  - TN = 18584
  - Total = 37720
- Les trois taux d'erreurs :
  - False Acceptance (FA) : FA = FP/(FP+TN) = 0.985
  - False Reject (FR) : FR = FN/(TP+FN) = 0.993
  - La moitié du taux total : HTER = (FR+FA)/2 = 0.989
  - On voit que le système fait peu d'erreurs et dans le cadre de notre application c'est souhaitable

TP – Seuil

- Le seuil (T) du modèle est 0,25: en dessous, les enregistrements sont considérés comme étant de locuteurs différents et au dessus, du même locuteur
- Rien qu'en regardant les valeurs de FP et FN on comprend que le seuil n'est pas optimal pour l'application du TD.
  - Rappel : identification par la voix
- Dans l'application du TD, on chercherait à minimiser les FP quitte à avoir plus de FN. Il est plus important d'être sur que la personne qui essaye de s'authentifier est bien la bonne quitte à ce qu'elle essaye 2 fois.
- Ici avec T=0,25, on a 2 fois plus de FP (276) que de FN (130)
- Un meilleur seuil pour cette problématique serait vers 0,35
- La répartition des scores est donnée dans les plots ci-dessous
  - Il y a évidemment beaucoup d'erreurs autour de 0,25 ce qui est logique car c'est le seuil
  - Dans les FPs, il y a une erreur avec un score de 0,45 ce qui est loin du seuil (nous analyserons cette erreur lors du knockout)
  - De même dans les FNs, il y a quelques erreurs en dessous de 0 ce qui est très éloigné su seuil
  - On constate que la majorité des FP se trouvent entre 0,25 et 0,3, déplacer le seuil vers 0,3/0,4 supprimerait beaucoup de faux positifs





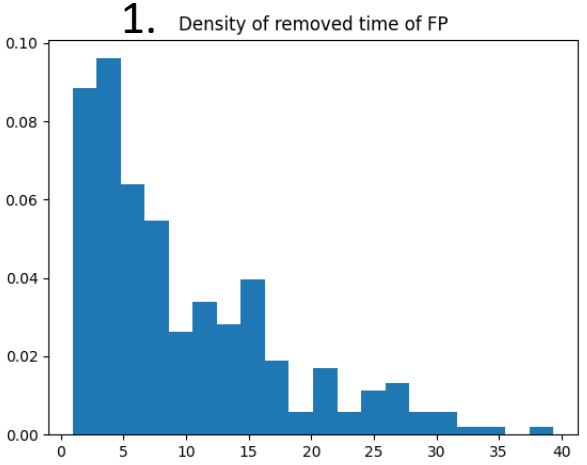
TP – Présentation de l'Approche par Counterfactual

- Dans le TP, nous avons mis en place l'approche counterfactual. Elle consiste à trouver le moyen d'inverser la décision, dans le cadre de la reconnaissance du locuteur, nous enlevons des parties de l'audio jusqu'au changement de la décision.
- Nous intéressons particulièrement aux erreurs, c'est-à-dire les faux positifs et faux négatifs car c'est là que nous pouvons probablement trouver les biais et limites du modèle.
- Notre but est de faire changer la prédiction si cela est possible (sinon faire changer le score significativement) en effectuant le knockout sur les différentes paires.
    - Dans notre cas, on a une paire d'enregistrements provenant de locuteurs et on effectue le knockout seulement sur un des deux enregistrements
    - J'utilise des fenêtres de 0,16 secondes
- 2 approches :
    - Knockout non continu : on prend la fenêtre qui fait changer le plus le score et ainsi de suite jusqu'à changer la prédiction
    - Knockout continu : on prend la fenêtre continue la plus courte possible qui fait changer la prédiction
- Nous nous intéresserons plus spécifiquement à quelques exemples et essayerons d'en comprendre l'erreur

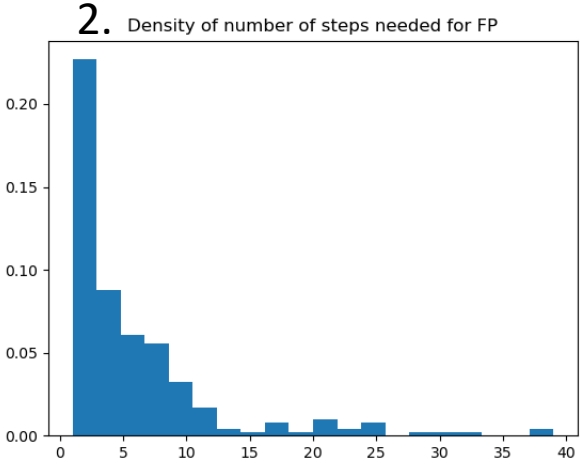
TP – Knockout sur des faux positifs

- L'histogramme 1 donne la proportion de temps qui a été enlevé. En axe x, c'est le % de temps retiré.
  - Pour une bonne partie, seulement 10% ou moins de l'audio ont été enlevé
- L'histogramme 2 donne le nombre de fenêtres qui ont été enlevé. En axe x, c'est le nombre de fenêtres.
  - On peut voir que ce graphique ne ressemble pas totalement au 1. Ces deux histogrammes sont tout de même corrélé (valeur de corrélation :0,71). Cela veut dire que plus un fichier audio est long, plus il faudra enlever de fenêtres pour changer la prédiction

1.



2.



TP – Knockout sur des faux positifs avec des exemples

a. 

b. 

bk. 

bc. 

- Premier exemple :
  - Durées : a: 10,20sec et b: 17,96sec
  - Le score de base était 0,47 (très élevé par rapport au seuil)
  - Il a fallu enlever 25 fenêtres (4,48sec) sur l'audio b pour que la prédiction change
    - Ce qui représente 25% du temps de b
- A et b sont deux hommes avec un accent semblable, peut-être un accent pas très fréquent
  - L'intonation est similaire
  - A a du bruit de fond alors b non, l'enregistrement B est beaucoup plus propre
- Analyse du knockout :
  - bk est le résultat du knockout sur b
  - Tous les « Just » ont été enlevé
  - Ce sont des mots qui ont été enlevés ce qui conforte l'hypothèse de l'accent
- Analyse du knockout en version continue :
  - bc est le résultat du knockout continu sur b
  - Cela a enlevé de 0,96sec à 13,92sec donc le début.
  - Cela a enlevé tous les « Just »
  - Ça a laissé la dernière phrase
  - Toutes les hésitations étaient dans la partie enlevée

TP – Knockout sur des faux positifs avec des exemples

a. 

b. 

bk. 

- Deuxième exemple :
  - Durées : a: 7,96sec et b: 14,68sec
  - Le score de base était 0,259(très proche du seuil)
  - Il a fallu enlever 1 fenêtre (0,16sec) sur l'audio b pour que la prédiction change
    - Ce qui représente 1% du temps de b
- A et b sont deux femmes
  - B semble zozoter
  - A a un accent qui ressemble légèrement à du zozotement
- Analyse du knockout :
  - bk est le résultat du knockout sur b
  - C'est la toute fin qui a été retiré
  - Le mot « fainted » a été coupé et l'allongement de la dernière syllabe a été enlevé
  - Enlever une syllabe allongée a permis au système de voir que ce n'était pas la même personne
  - La ressemblance entre les deux personnes était donc basée sur le sexe et le genre de zozotement

TP – Knockout sur des faux positifs avec des exemples

a. 

b. 

bk. 

bc. 

- Troisième exemple :
  - Durées : a: 6,56sec et b: 4,48sec
  - Le score de base était 0,34
  - Il a fallu enlever 6 fenêtres (0,96sec) sur l'audio b pour que la prédiction change
    - Ce qui représente 21% du temps de b
- A et b sont deux femmes qui ont un léger accent où elles allongent certaines syllabes
- Analyse du knockout :
  - bk est le résultat du knockout sur b
  - Les parties enlevés concernent les syllabes allongées justement : « right », « tries » « said »
  - De plus, ces syllabes concernent des sons « i »
- Analyse du knockout en version continue :
  - bc est le résultat du knockout continu sur b
  - Cela a enlevé de 3,84sec à 4,48sec (la fin).
  - Cela a enlevé moins qu'avec le knockout de base. Je ne sais pas pourquoi, soit une mauvaise implémentation. Soit le fait que sur le knockout continu, les fenêtres testées se chevauchent et donc cela teste possiblement de meilleurs placements de fenêtres.
  - Cela enlève « I said »

TP – Knockout sur des faux négatifs

- Il n'est pas forcément possible d'inverser la prédiction comparé aux FP. Il est plus simple de faire baisser le score que de le faire monter.
- Il y a 7 Faux négatifs avec un score inférieur à 0 (soit à plus de 0,25 du seuil) et pour ces 7 exemples, il a été impossible de changer la prédiction. En regardant de plus près, on s'aperçoit que tous ces pairs où il y a des erreurs viennent d'un seul et même locuteur : « id10304 ».

```
7
('id10304/HTL8iLI75TY/00001.wav', 'id10304/HTL8iLI75TY/00004.wav')
('id10304/HTL8iLI75TY/00004.wav', 'id10304/GRv7pEnTwUc/00001.wav')
('id10304/HTL8iLI75TY/00004.wav', 'id10304/JQtDfEz08aU/00014.wav')
('id10304/HTL8iLI75TY/00004.wav', 'id10304/FJOvOooCIvs/00005.wav')
('id10304/W-5dGkFnQxM/00001.wav', 'id10304/HTL8iLI75TY/00004.wav')
('id10304/W-5dGkFnQxM/00001.wav', 'id10304/0TJ9sdJiDvA/00005.wav')
('id10304/hDBMV_0Vz4E/00010.wav', 'id10304/o9bD_E99Jpg/00015.wav')
```

A 

B 

- 'id10304/HTL8iLI75TY/00004.wav' apparait dans beaucoup d'eux.
  - A est 'id10304/HTL8iLI75TY/00001.wav'
  - B est 'id10304/HTL8iLI75TY/00004.wav'
    - On se rend compte que dans cet enregistrement, ce n'est pas le bon locuteur qui parle, une femme au lieu d'un homme
  - Même problème pour : 'id10304/W-5dGkFnQxM/00001.wav' ce n'est pas la bonne personne
  - Même problème pour : 'id10304/o9bD_E99Jpg/00015.wav' ce n'est pas la bonne personne
    - Il y a donc un problème avec les enregistrements provenant de ce locuteur

TP – Knockout sur des faux négatifs

A 

B 

Bk 

- Premier exemple :
  - Durées : a: 4,20sec et b: 4,12sec
  - Le score de base était 0,247
  - Il a fallu enlever 4 fenêtres (0,64 sec) sur l'audio b pour que la prédiction change
    - Ce qui représente 15% du temps de b
- A et b proviennent d'un homme
  - Dans les deux cas, il y a de la musique, assez forte dans B
  - Dans A, la personne a un accent qui n'est pas présent dans B
- Analyse Knockout :
  - Encore une fois les hésitations ont été enlevés

A 

B 

Bk 

- Deuxième exemple :
  - Durées : a: 3,96sec et b: 3,96sec
  - Le score de base était 0,189
  - Il a fallu enlever 5 fenêtres (0,8 sec) sur l'audio b pour que la prédiction change
    - Ce qui représente 20% du temps de b
- A et b proviennent d'un homme
  - Dans A, l'enregistrement est propre tandis que dans B, il est très bruité
  - Dans B, le locuteur est plus hésitant
- Analyse Knockout :
  - « Starbucks » et du blanc ont été enlevés
  - « Starbucks » on dirait que la personne est loin du micro lorsqu'elle le dit, ce qui altère le son donc peut-être que ça a mis en difficulté le modèle
  - Le blanc a été enlevé car cela enlève des parties avec seulement du bruit de fond

TP – Knockout sur vrais positifs

A 

B 

Bk 

- Premier exemple :
  - Durées : a: 14,76sec et b: 8,12sec
  - Le score de base était 0,82
  - Il a fallu enlever 30 fenêtres (4,8 sec) sur l'audio b pour que la prédiction change
    - Ce qui représente 59% du temps de b
- A et b viennent d'une femme
  - Dans les deux cas, les enregistrements sont propres, sans bruits
- Analyse du knockout :
  - Cela a laissé que des syllabes, aucun mot de plus de 1 syllabe est entier

- Deuxième exemple :
  - Durées : a: 18,8sec et b: 4,48sec
  - Le score de base était 0,375
  - Il a fallu enlever 6 fenêtres (0,96 sec) sur l'audio b pour que la prédiction change
    - Ce qui représente 21% du temps de b

A 

- A et b viennent d'une femme, elle ne semble pas avoir anglais comme langue maternelle
  - Dans A, l'enregistrement est très légèrement bruité
  - Dans B, la voix est un peu déformée par le micro

B 

- Analyse du knockout :
  - « Los angeles » et d'autres syllabes ont été enlevés
  - Les hésitations ont été laissés
  - Beaucoup de son « e » dans A et justement « Los angeles » correspond à un son « e »
  - Le modèle doit se baser en partie sur son accent et donc lui enlever des mots avec un accent prononcé lui rend la tâche plus difficile

Bk 

- Dans les 2 cas, A est un audio très long et B assez court

TP – Conclusion & Regard Critique

- Regard critique :
  - Je n'ai pas pu explorer toutes les pistes
    - Analyses locales sur quelques exemples et difficilement généralisable
  - Je n'ai pas testé le changement de la valeur du seuil
  - Qu'est-ce qu'un changement « significatif » lorsque l'on ne peut pas changer la prédiction ?
  - Difficultés d'analyse du modèle et de ces résultats

- Conclusion :
  - Le modèle est impacté par les bruits de fonds, les hésitations…
  - Il se base en partie sur certaines particularités vocales de la personne comme les accents
  - Certaines données sont probablement mal classées : certains audios semblant être labélisé à la mauvaise personne
  - Le seuil de 0,25 n'est pas adapté à l'application (il est nécessaire de l'augmenter)

Merci d’avoir pris le temps de lire

AVIGNON
UNIVERSITÉ

UCE Explicabilité et Interprétabilité
Jean-François Bonastre

Audran BERT
M2 Intelligence Artificielle

21/12/2022