Projet Données Multimédia

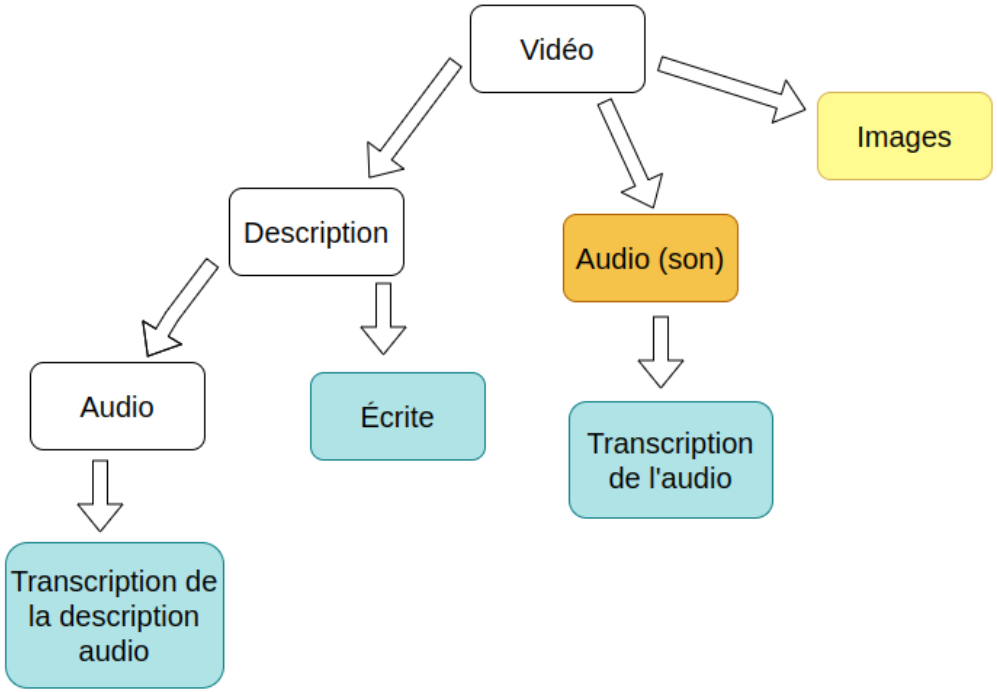
Classification de données multimédias et étude cross-modale d'un corpus

Léa Fabriol
Valentin Lafargue
Théotime Dmitrašinović

M2 SID

23/10/2023 - 10/11/2023

Corpus de vidéo QuerYD



Augmentation des données

La **transcription de l'audio** est obtenue avec le modèle **Whisper small**

↪ temps de calcul : 15 h

Figure: Schéma explicatif de la nature des données

Répartition des labels Youtube

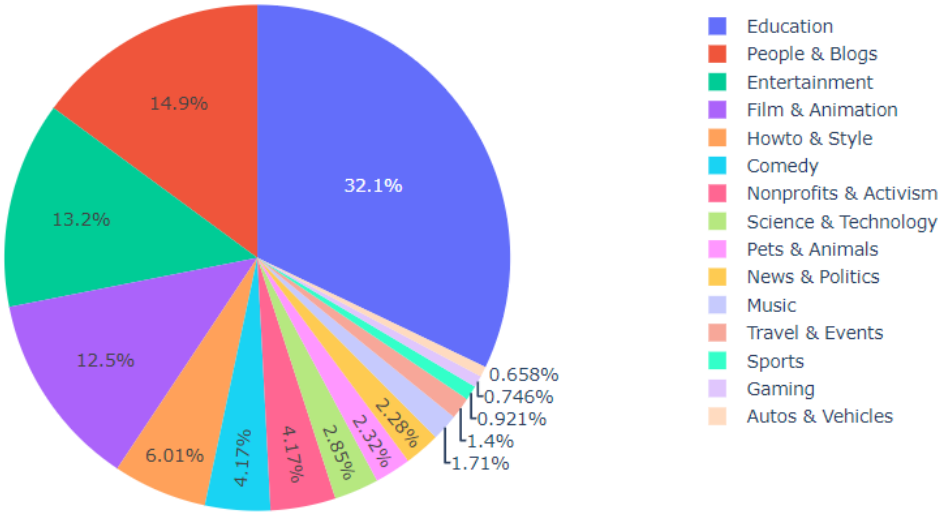


Figure: Répartition des labels YouTube dans le jeu de données

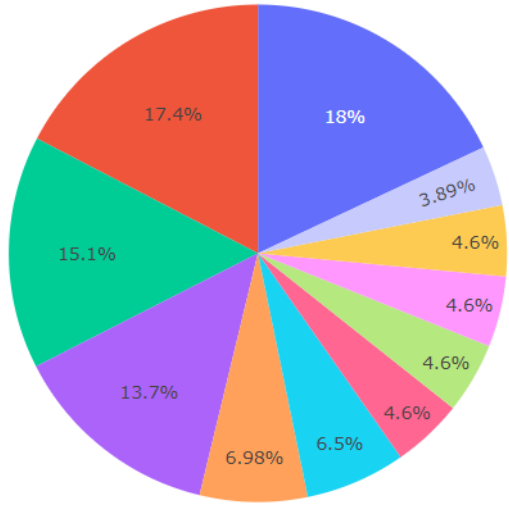


Figure: Répartition des labels YouTube apres sub et sur-sampling

Relabélisation

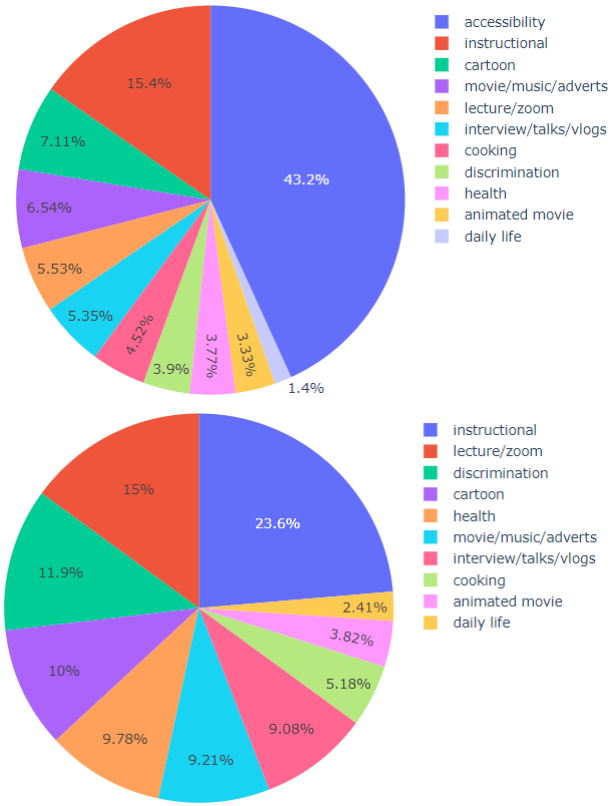


Figure: Labels zero-shot

zero-shot

- 37 h de calcul
- Relabélisation avec les catégories du papier

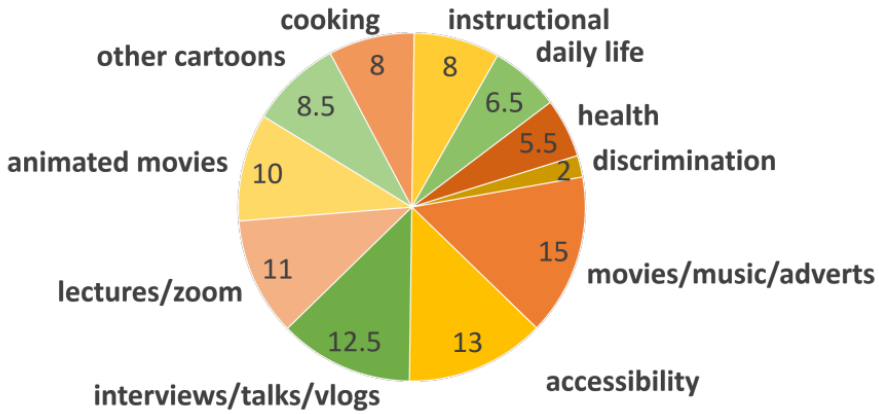
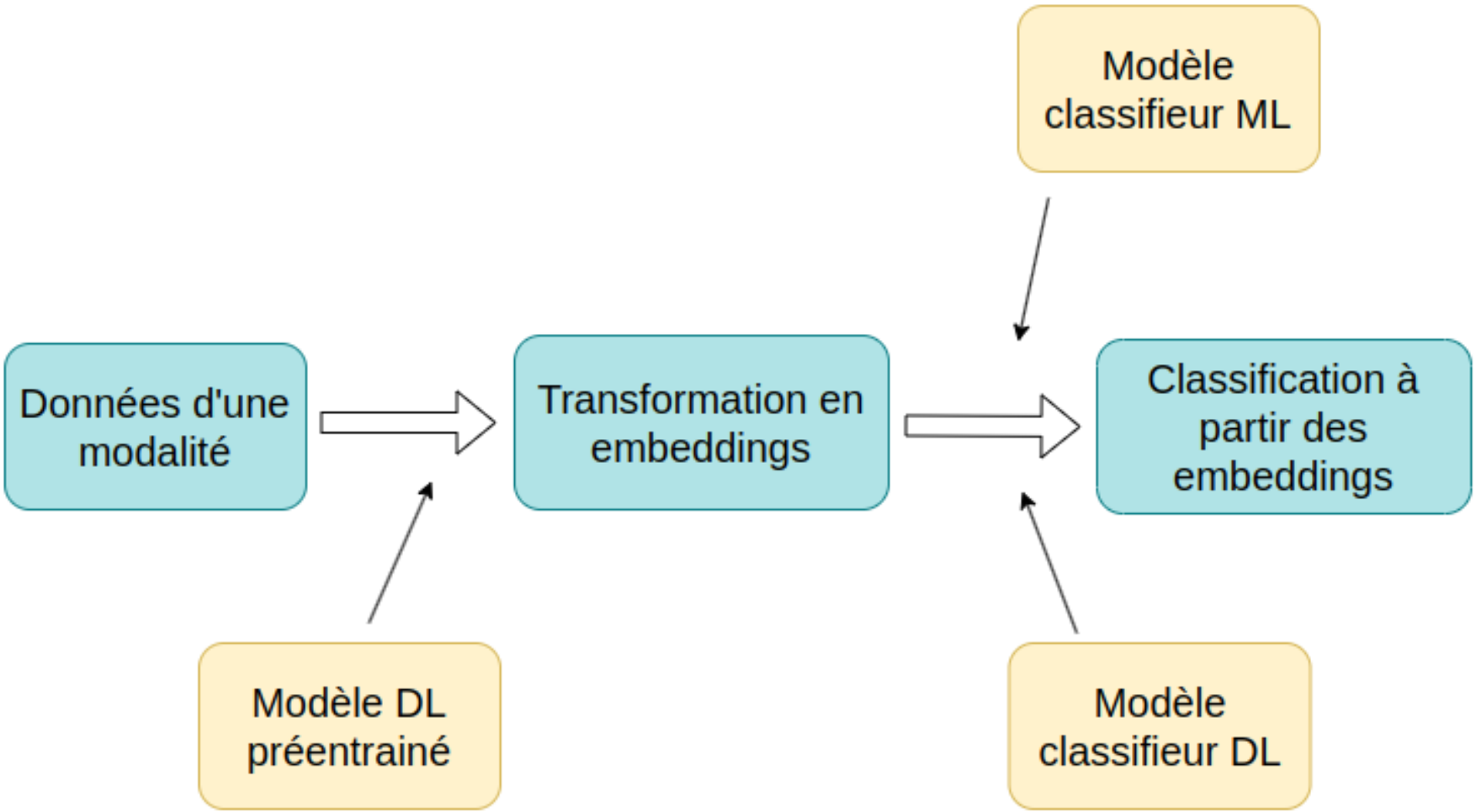


Figure: Labels du papier

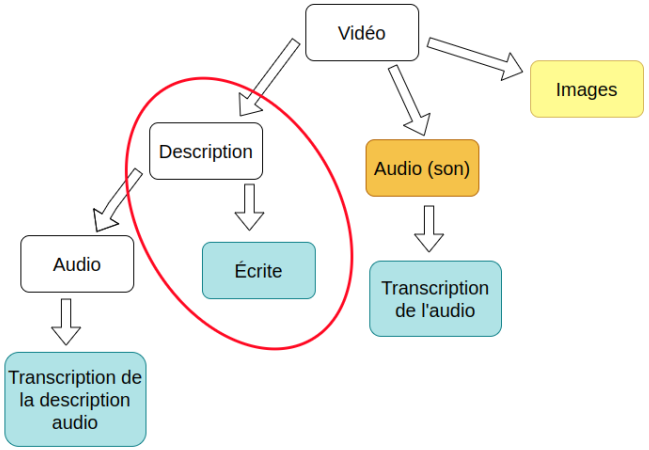
Présentation de la stratégie globale en classification



Texte

Modèles utilisés pour la transformation en embeddings

- Sentence-BERT
- DistillBERT



Label utilisées	**Méthode embeddings**	**Méthode de classification**	**Temps calcul**	**Accuracy**
Youtube	s-bert mpnet	MLP	2h	54 %
Youtube	DistillBERT	Model spécialisé	10 min	60 %
Sur+Sub Sampling	DistillBERT	Model spécialisé	10 min	48 %

Table: Meilleurs résultats des différents tests de classification à partir des données textuelles

Audio

Modèles utilisés pour la transformation en embeddings

- YAMNet
- VGGish: → Average VGGish → Max pool VGGish

Labels	**Méthode embeddings**	**Méthode de classification**	**Temps calcul**	**Accuracy**
Youtube	Average VGGish et Max pool VGGish	1vsRestSVC	25h	53 %
Youtube	VGGish et YAMNet	1vsRestSVC	25h	50 %

Table: Meilleurs résultats des différents tests de classification à partir des données audio

Image

Modèles utilisés pour la transformation en embeddings

- YOLOv7 + S-BERT ou proba moyenne/max
- ViViT

Labels	**Méthode embeddings**	**Méthode de classification**	**Temps de calcul**	**Accuracy**
zero-shot	YOLOv7	SVC	61h	35 %
Youtube	ViViT	MLP (mean)	23h	41 %

Table: Meilleurs résultats des différents tests de classification à partir des images des vidéos

Passage d'une modalité à une autre

Texte-Image

- CLIP
  → Image de la video et résumé de la description écrite
- Dalle-mini
  → Génération d'image à partir des description écrites

Embeddings-Embeddings

- Passage direct
- Réduction de dimension avec un encodeur
- Moyennes d'embeddings
  → manque de diverdité des résultats

Agrégation des modalités pour la classification

Plusieurs stratégies d'agrégations mises en place

Agrégation précoce
↪ Concaténation avant d'entrer dans le modèle de classif

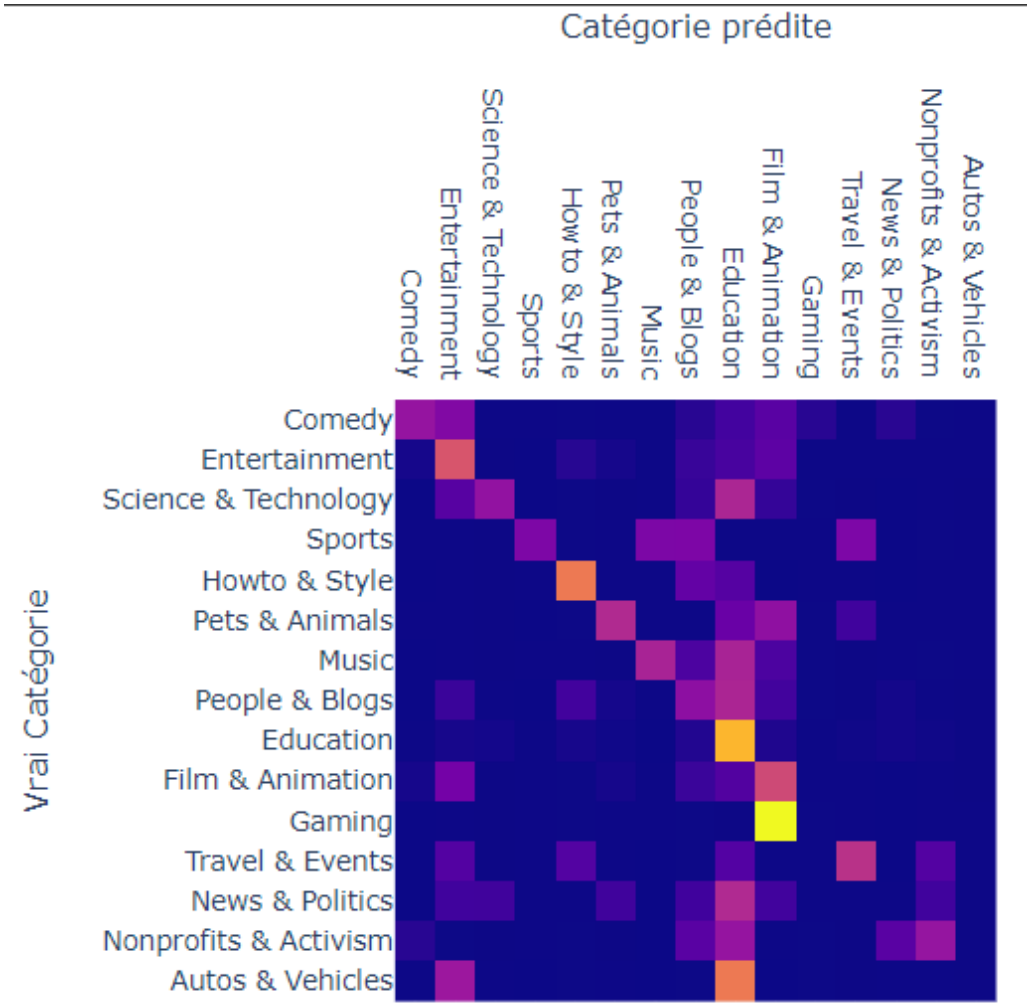
Agrégation intermédiaire
↪ reduction de dimention avant la prédiction

Agrégation tardive
↪ on apprend à partir des des probas des modèles de classif

Labels	**Méthode embeddings**	**Méthode d'agrégation**	**Model**	**Acc**
Youtube	VGGish-Yolo-MPNet	Précoce	VsSVC	54 %
SurSub Youtube	VGGish-Yolo-MPnet	Inter	MLP	48 %
Youtube	VGGish-Yolo-MPNet	Tardive	VsSVC	55 %

Table: Meilleurs résultats des différents tests de classification à partir des vidéos en agrégeant les 3 modalités

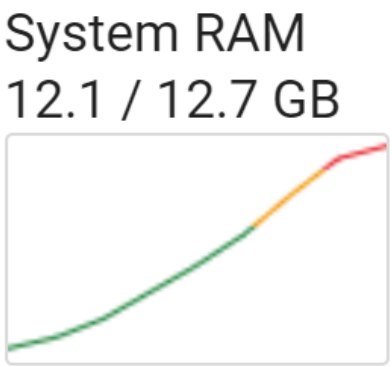
Matrice de confusion approche tardive



Bilan

Les -

- Les badges
- Source label
- Colab Pro
  → RAM
  → GPU



Les +

- Bonne cohésion
- Projet très enrichissant
  → log des résultats
- Accompagnement

Conclusion

Axes progrès

- Proba sur meilleur resultat de texte pour agrégation tardive
- VitMAE mais pb d'autentification google
- Pas pu faire de deep à chaque étape
  $\rightarrow$ convolution sur les embeddings

Merci pour votre attention.

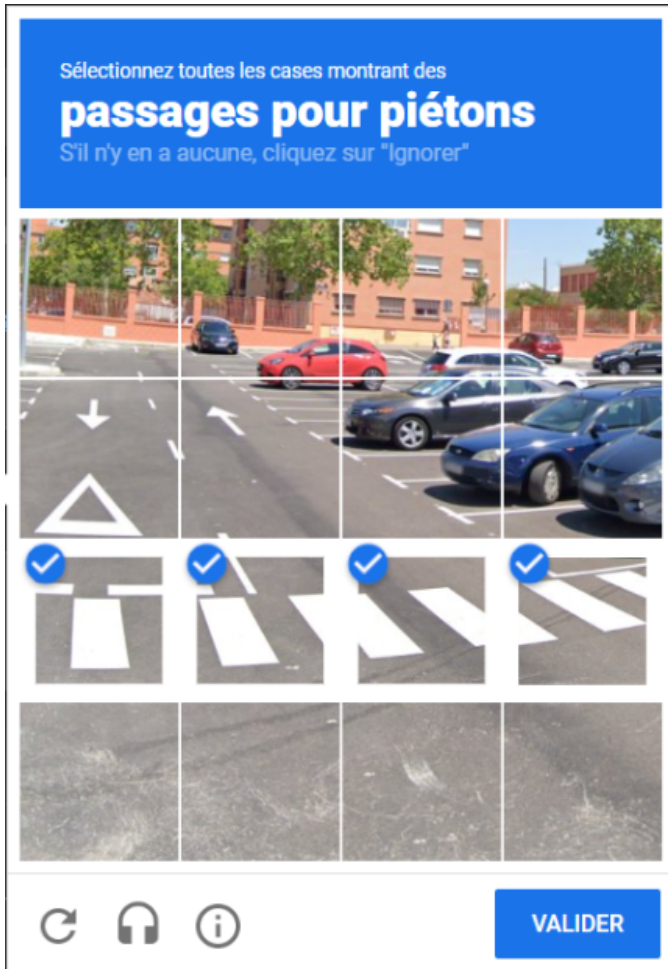


Figure: Captcha recontré au cours du projet