INSTITUT POLYTECHNIQUE DE PARIS

Research internship at Telecom Paris (French version below)

Unsupervised data selection for knowledge distillation of self-supervised speech models.

Topic details :

Knowledge distillation is the process of transferring knowledge from a large teacher model to a smaller student one, to reduce inference time and computational cost. In this internship, we want to explore training data selection for distilling self-supervised speech representation models. Proper distillation requires access to the full dataset used for training the teacher model, which is problematic for two reasons :

- Training data may not be publicly available, in the cases of private data or unclear data manipulation during training.
- Training data consists of huge datasets, leading to costly distillations.

We want to explore techniques coming from unsupervised data selection for a better selection of training data in distillation.

References :

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., & Wei, F. (2021). WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal on Selected Topics in Signal Processing*, *16*(6), 1505–1518. https://doi.org/10.1109/JSTSP.2022.3188113

Lee, Y., Jang, K., Goo, J., Jung, Y., & Kim, H. (2022). FitHuBERT: Going Thinner and Deeper for Knowledge Distillation of Speech Self-Supervised Learning. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2022-September*, 3588–3592. https://doi.org/10.48550/arxiv.2207.00555

Lu, Z., Wang, Y., Zhang, Y., Han, W., Chen, Z., & Haghani, P. (n.d.). *Unsupervised Data Selection via Discrete Speech Representation for ASR*. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2022-September*, https://arxiv.org/abs/2204.01981

Duration : 5 months

Audio Data Analysis and Signal Processing (ADASP) - département Images, Données, Signal, Télécom Paris (https://adasp.telecom-paris.fr/)

Supervisors : Slim ESSID, Salah ZAIEM (Junior supervisor)

Start : April
This is a paid internship.

Contacts :
slim.essid@telecom-paris.fr , salah.zaiem@telecom-paris.fr

If you are interested in this offer, please send us an email with your CV and a few lines on your motivation for the project.

Version Française :

Sélection de données non supervisée pour la distillation de connaissances de modèles de parole auto-supervisés.

Détails du sujet :

La distillation de la connaissance est le processus de transfert de la connaissance d'un grand modèle “enseignant” vers un plus petit modèle "élève", afin de réduire le temps d'inférence et le coût de calcul. Dans ce stage, nous voulons explorer la sélection de données d'entraînement pour la distillation de modèles de représentation de la parole auto-supervisés. Une distillation correcte nécessite l'accès au jeu de données complet utilisé pour l'entraînement du modèle de l'enseignant, ce qui est problématique pour deux raisons :

- Les données d'entraînement peuvent ne pas être accessibles au public, dans les cas de données privées ou de manipulation peu claire des données pendant l'entraînement.
- Les données d'entraînement consistent en d'énormes ensembles de données.

Nous voulons explorer des techniques provenant de la sélection de données non supervisée pour une meilleure sélection des données d'entraînement dans la distillation.

Références :

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., & Wei, F. (2021). WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal on Selected Topics in Signal Processing*, *16*(6), 1505–1518. https://doi.org/10.1109/JSTSP.2022.3188113

Lee, Y., Jang, K., Goo, J., Jung, Y., & Kim, H. (2022). FitHuBERT: Going Thinner and Deeper for Knowledge Distillation of Speech Self-Supervised Learning. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2022-September*, 3588–3592. https://doi.org/10.48550/arxiv.2207.00555

Lu, Z., Wang, Y., Zhang, Y., Han, W., Chen, Z., & Haghani, P. (n.d.). *Unsupervised Data Selection via Discrete Speech Representation for ASR*. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2022-September*, https://arxiv.org/abs/2204.01981

Durée : 5 mois
Analyse des données audio et traitement du signal (ADASP) - département Images, Données, Signal, Télécom Paris (https://adasp.telecom-paris.fr/)

Encadrants : Slim ESSID, Salah ZAIEM

Début : Avril

Stage rémunéré.

Contact :
slim.essid@telecom-paris.fr , salah.zaiem@telecom-paris.fr

Si vous êtes intéressé par cette offre, veuillez nous envoyer un email avec votre CV et quelques lignes sur votre motivation pour le projet.