Master's internships 2021/2022

*A limited number of positions for a Master's internship are this year. These internships should be for **A MINIMUM OF 5-6 MONTHS**. Please send your application by e-mail (CV + LM required) to **syntheticlearner@gmail.com.***

1. Self-supervised learning at the word level

Supervision in machine learning is a paradigm that requires labelled datasets which is often the result of substantial human time and efforts. For that reason, unsupervised or self-supervised methods are becoming increasingly used in several areas of machine learning: vision, text, and very recently speech. For instance, contrastive predictive coding or Wav2vec2.0 have been used to discover speech representations without supervision [1,2]. These models can embed fixed duration (usually 10ms) of speech into a vector but cannot represent variable-length speech sequences. These latter representations can be very useful in a variety of tasks ranging from information retrieval to speech segmentation into words (i.e can you find word boundaries in an audio recording without label nor prior knowledge of the language?).
The aim of this internship is to search for new and more robust methods to build variable length speech embeddings. You will implement new loss functions and regularisation schemes in a pre-existing deep learning model to improve its performance. The resulting model will be used as input to a speech segmentation model and hopefully improve the current state-of-the-art in that domain. This internship will be done in collaboration with researchers at Facebook AI Research.

[1] Aaron van den Oord, Yazhe Li, Oriol Vinyals (2019). Representation Learning with Contrastive Predictive Coding https://arxiv.org/pdf/1807.03748.pdf
[2] https://arxiv.org/abs/2006.11477
[3] https://arxiv.org/abs/2007.13542

2. Universal speech synthesis

Text-to-speech synthesis (like the voice of Siri, or of Google Translate) has seen stunning advances in the last five years, sounding extremely natural. However, like many other speech tools, it requires massive amounts of labelled training data to construct a good speech synthesis model, and this means concentrating on single language, a single, usually highly standardized, accent, and, for best results, a single speaker. As a consequence, there are not many languages with high-quality speech synthesis. A new wave of speech synthesis attempts to move towards *multilingual* speech synthesis, so that, on the basis of training data in a handful of languages, we would have a single model

that could generate speech in other languages. The goal of this internship is to advance *universal* speech synthesis, a single model which can synthesize speech in any language in the world.
Several papers have attempted tasks going in this direction, including recent work in our lab. A promising approach is that taken by [https://arxiv.org/abs/2008.04107], which uses meaningful articulatory-inspired features as input, to allow for a more general model, as well as the task of cross-lingual voice transfer [https://arxiv.org/abs/1907.04448], which takes speech in an unknown language as input and re-speaks it, without knowledge of this language. Tasks for the intern may thus include attempting to take the articulatory features approach, and seeing whether it continues to be useful for the voice transfer task, or developing a more rigorous evaluation in order to push the limits of these models, depending on the interests and aptitudes of the intern.

3. Domain discovery for blind distributional robustness in deep learning models

Current neural network-based machine learning models are able to attain high accuracy across a wide range of modalities and tasks. However, they often fail when used on data from different domains: for example, a machine translation model trained on news articles will tend to perform poorly when translating social media comments, a very different kind of text [1]. This has unfortunate implications for their use in real-world scenarios where distribution shifts abound.
There has been a recent surge of interest in developing models that are robust to domain shift. However, most such approaches rely on having examples from a variety of different domains at training time to learn invariant representations [2], or make strong assumptions on the nature of the shift (eg. that the target domain overlaps with the training data [3]).
The subject of this internship will be to develop a training algorithm that is able to yield models that are more robust to domain shift, without the aforementioned limitations. A potential idea to explore would be to uncover domains present in the training data, and use this information to learn invariant features. However, other approaches could be explored depending on the interests of the intern. For inquiries, please contact pmichel31415@gmail.com.

[1] Michel, P., & Neubig, G. (2018). MTNT: A testbed for machine translation of noisy text. https://arxiv.org/pdf/1809.00388.pdf
[2] Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. https://arxiv.org/pdf/1907.02893.pdf
[3] Oren, Y., Sagawa, S., Hashimoto, T. B., & Liang, P. (2019). Distributionally robust language modeling. https://arxiv.org/pdf/1909.02060.pdf

4. Curriculum design for emergent communication

There has been growing interest in recent years in studying emergent communication between two agents (generally neural network models) which have to cooperate using language to solve simple games [1]. Such simple communication games provide interesting testbeds for studying the emergence of language under controlled conditions (difficulty of the underlying task, capacity of the agents, etc..).
In most experimental settings however, both the difficulty of the game and the agent capacity are fixed throughout the experiment. This limits the range of games to those that can be solved by bootstrapping a communication protocol essentially from scratch.
The goal of this internship is to design curriculum learning [2] approaches for learning communication protocols in games of increasing complexity. Depending on the intern's interests, we will explore ways to make simple image-based referential games more (or less) difficult and how this affects learning. Another complementary approach will be to vary the capacity of the agent (eg. by using increasingly larger architectures, or by expanding the agent's vocabulary). We will look at whether such curricula enable us to train agents on more difficult games and study the properties of the communication protocols that emerge.

[1]: Lazaridou, A., & Baroni, M. (2020). Emergent multi-agent communication in the deep learning era. https://arxiv.org/pdf/2006.02419.pdf
[2] Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009, June). Curriculum learning. https://ronan.collobert.com/pub/matos/2009_curriculum_icml.pdf

5. Using deep learning to study children's multimodal behavior in face-to-face conversation

If you want to apply for this internship subject, please send your inquiries and application directly to Abdellah Fourtassi (abdellah.fourtassi@gmail.com)

The study of how children develop their conversational skills and how these skills help them learn from others is an important scientific frontier at the crossroad of social, cognitive, and linguistic development with important applications in health (e.g., mitigating communicative difficulties), education (e.g. improving teaching practices), and child-oriented AI (e.g., virtual learning companions). Recent advances in Natural Language Processing and Computer Vision allow going beyond the limitations of traditional research methods in the lab and advance formal theories of conversational development in real-life contexts. In this internship, we will leverage some of these recent techniques (e.g., multiscale recurrent neural network, see [1]) to build a model that mimics how children behave in face-to-face conversations with their caregivers and how this behavior develops across middle childhood. The intern will have access to the child-caregiver conversation data collected by our team [2]. The data has already been hand-annotated for non-verbal behavior (e.g., nods, smiles, and frowns) and is currently being transcribed for verbal data and processed for extraction of vocal/acoustic features. The interns will contribute to the development of a model (building on an existing pipeline in PyTorch) that aims at studying

how multimodal cues from the vocal, visual, and verbal dimensions contribute to predicting the child's coordination behavior in conversation (e.g., turn-taking management, negotiating shared understanding with the interlocutor, and the ability for a coherent/contingent exchange). The intern will collaborate closely with several members of our team, involving computer scientists, psychologists, and linguists (see our website www.cocodev.fr) as well as members from the CoML team.

[1] Roddy, M., Skantez, & Harte (2018). Multimodal Continuous Turn-Taking Prediction Using Multiscale RNNs. *In Proceedings of the 20th ACM International Conference on Multimodal Interaction*

[2] Bodur, K., Nikolaus, M., Kassim, F., Prévot, L., & Fourtassi, A. (2021). ChiCo: A Multimodal Corpus for the Study of Child Conversation. *In Proceedings of the International Workshop on Corpora and Tools for Social Skills Annotation. 23rd ACM* International Conference on Multimodal Interaction