INF642 : Socio-emotional Embodied Conversational Agents

Lab 2: Developing the Deep Learning Model for Upper Facial Gestures Generation

22/01/2020 – 1:30 pm

Note: The submission deadline is: 27/01/2020, 11:59pm. You need to submit a report that explains your work, and your source code. Add all the deliverables in a ZIP file, **save it with your first and last name**, and send it to **mf.upmc@gmail.com**

Objective: The objective of this lab session is to develop a sequence to sequence model that can predict the upper facial action units AU01, AU02, AU04, AU05, AU06, and AU07, based on the acoustic and prosodic features from Lab 1.

Sequence to Sequence Model

Your deep learning model will be an Encoder-Decoder Sequence to Sequence (Seq2Seq) model [1], that should be implemented with **Keras** [2].

The idea is to develop a Seq2Seq model composed of 2 Recurrent Neural Networks (1 encoder and 1 decoder), this method is a **machine translation method** that maps an input sequence to an output sequence.

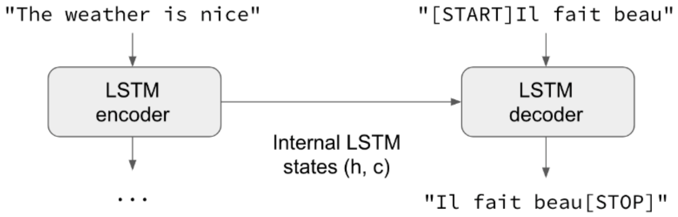

Figure 1 Canonical Sequence to Sequence [2]

In this lab, your model should predict the 6 action units AU01, AU02, AU04, AU05, AU06 and AU07 based on the Fundamental Frequency (F0). These features are the ones that you have extracted in your Lab 1.

The input sequence will be a sequence of Fundamental Frequencies (F0), and the generated output will be composed of 6 sequences of the 6 action units. Therefore, your network can be composed of 1 encoder to encode the sequence of F0s, and 6 decoders to decode each sequence of Action Unit.

The input sequence of F0 should have a sequence length equal to **100 frames**, and each output sequence of AU should also be equal to **100 frames**.

You must split your data into 3 sets: training set (80%), validation set (10%), and a testing set (10%). Your model should be trained using the training set **only**, and the validation process should be done using the validation set. Save your testing set for Lab 3 to evaluate your model.

The **teacher forcing** method is very effective, and you're very encouraged to use it.

Note

You can design your network architecture as you want: type of RNN (GRU or LSTM), number of layers, number of neurons, batch size, and all the hyperparameters can be of your choice. You need to optimize it to reduce the prediction error as much as possible.

Other architectures are welcomed too: you can get creative as much as you want, what is important is to get a good prediction (which will be evaluated in Lab 3) of the 6 AUs based on F0.

Do not hesitate to go online and get inspired by some existing seq2seq architectures (+ reference them in your report)

You must add comments **explaining **every part of your code****

Your report must include :

- An explanation of your network architecture
- A graph illustrating your architecture
- A plot of the training loss Vs. the validation loss
- **1 plot (only)** for each action unit, that shows the predicted action unit Vs. the ground truth

Bonus

- Add an attention mechanism (of your choice) between your encoder and decoder(s) (by adding the attention mechanism, your results will improve, and therefore the evaluation of your model in Lab 3 will produce better results😉)
- Develop a similar model that can predict the 6 action units based on **all** the prosodic features that you have extracted in Lab 1 combined (by adding other encoders to your initial architecture).

References

[1] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, *27*, 3104-3112.

[2] https://blog.keras.io/a-ten-minute-introduction-to-sequence-to-sequence-learning-in-keras.html