C. Anticipated issues

a) Optimization function: It is likely we will end up tweaking the optimization function over the course of our study, be it by using a linear combination of the functions we already brought up or by using a completely different function.

b) Training set: We intentionnally started out with a fairly small training set in order to be able to put forth some preliminary results faster so that early decisions can be made. It is possible we will not have enough training examples to train very deep architectures and will end up using bigger training set.

c) Evaluation method: The evaluation method detailed above is fairly clumsy in that it is extremely vulnerable to outliers. Various improvements on it could be explored if we notice that we consistently have some outliers that artificially skew the results in favour of one type of intermediate representation.

d) Network hyper-parameters: The above hyper-parameters (number of layers, dimension of each layer, ...) have mostly been chosen arbitrarily. The values chosen seemed reasonnable in light of the pursued goal and the task at hand, but will most likely evolve depending on experiments' results.

APPENDIX

REFERENCES

[1] Antoine Bordes et al. "Joint Learning of Words and Meaning Representations for Open-Text Semantic Parsing." In: *AISTATS*. Vol. 351. 2012, pp. 423–424.

[2] Najim Dehak et al. "Front-end factor analysis for speaker verification". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.4 (2011), pp. 788–798.

[3] Omid Ghahabi and Javier Hernando. "Deep belief networks for i-vector based speaker recognition". In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2014, pp. 1700–1704.

[4] Omid Ghahabi and Javier Hernando. "Deep Learning for Single and Multi-Session i-Vector Speaker Recognition". In: *CoRR* abs/1512.02560 (2015). URL: http://arxiv.org/abs/1512.02560.

[5] G. E. Hinton and R. R. Salakhutdinov. "Reducing the Dimensionality of Data with Neural Networks". In: *Science* 313.5786 (2006), pp. 504–507. ISSN: 0036-8075. DOI: 10.1126/science.1127647. eprint: http://science.sciencemag.org/content/313/5786/504.full.pdf. URL: http://science.sciencemag.org/content/313/5786/504.

[6] Anthony Larcher et al. "ALIZE 3.0-open source toolkit for state-of-the-art speaker recognition." In: *Interspeech*. 2013, pp. 2768–2772.

[7] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[8] Fred Richardson, Douglas Reynolds, and Najim Dehak. "Deep neural network approaches to speaker and language recognition". In: *IEEE Signal Processing Letters* 22.10 (2015), pp. 1671–1675.

[9] George Saon et al. "The IBM 2016 English Conversational Telephone Speech Recognition System". In: *CoRR* abs/1604.08242 (2016). URL: http://arxiv.org/abs/1604.08242.

[10] Mohammed Senoussaoui et al. *An i-vector Extractor Suitable for Speaker Recognition with both Microphone and Telephone Speech*.

[11] Vedran Vukotic, Christian Raymond, and Guillaume Gravier. "Bidirectional Joint Representation Learning with Symmetrical Deep Neural Networks for Multimodal and Crossmodal Applications". In: *ICMR*. ACM. New York, United States, June 2016. URL: https://hal.inria.fr/hal-01314302.