Laboratório de Instrumentação e Física Experimental de Partículas

Competence Center - Big Data

7th Informal Meeting – 6st July 2018

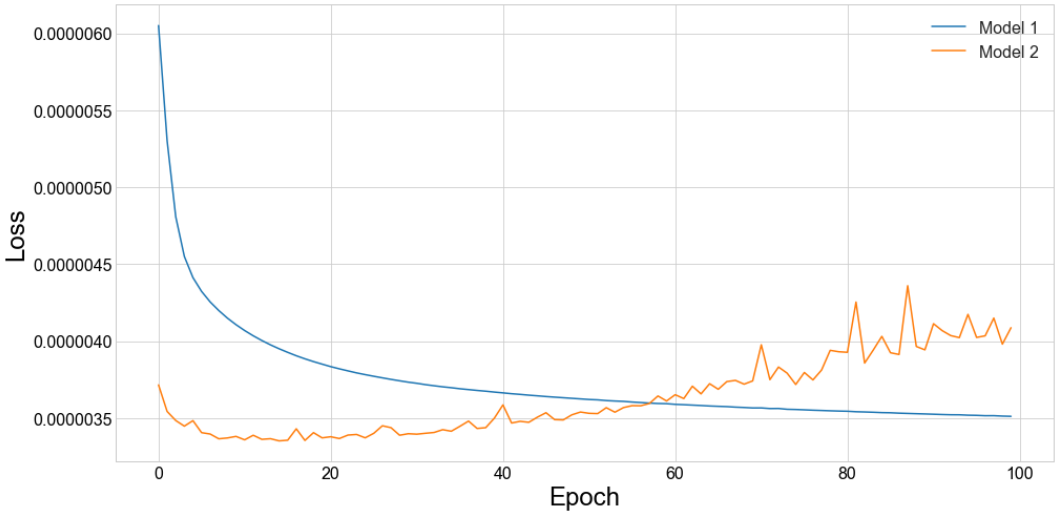
ATTENDEES (via zoom)

Nuno Castro, Giles Strong, Celso Franco, Guilherme Milhano, Liliana Apolinário, Helmut Wolters, Jorge Gomes, João Pedro Marado, Marcin Stolarski, Ricardo Gonçalo

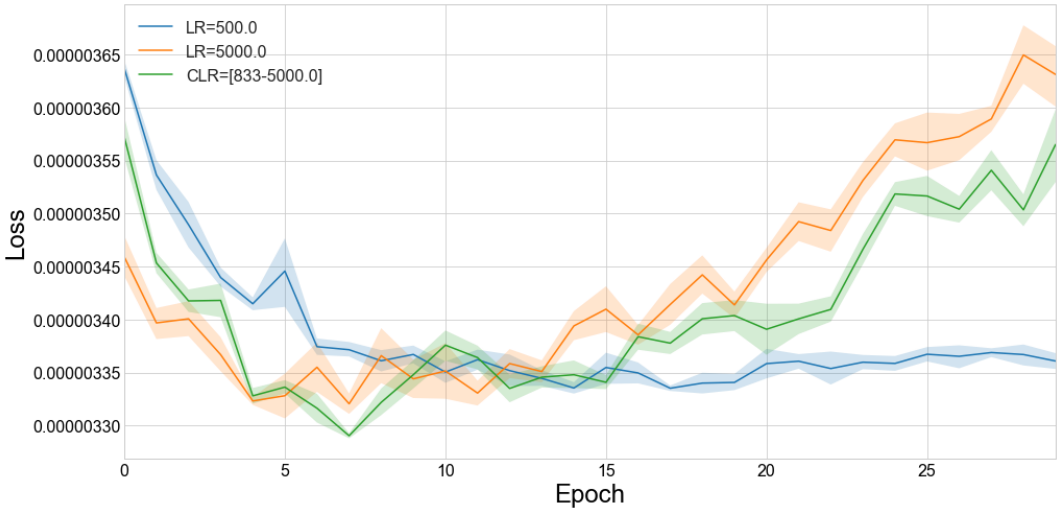
NOTES

- **Agenda:** https://indico.lip.pt/event/454/
- **Introduction**:
    - PTDC project on Big Data has started on July 1st. In this context a researcher position will be opened this month, as announced last meeting.
- **Giles Strong: Journal Club – tuning of hyperparameters in neural networks**
    - Discussion of https://arxiv.org/abs/1803.09820 ("a disciplined approach to neural network hyper-parameters: learning rate, batch size, momentum, and weight decay" by Leslie N. Smith).
        - Optimization of hyper-parameters, regularization, and network architecture is a critical aspect of deep neural networks.
    - Tutorial illustrating the technique in https://github.com/GilesStrong/Smith_HyperParams1_Demo
    - Docker image available (instructions available in the meeting agenda).

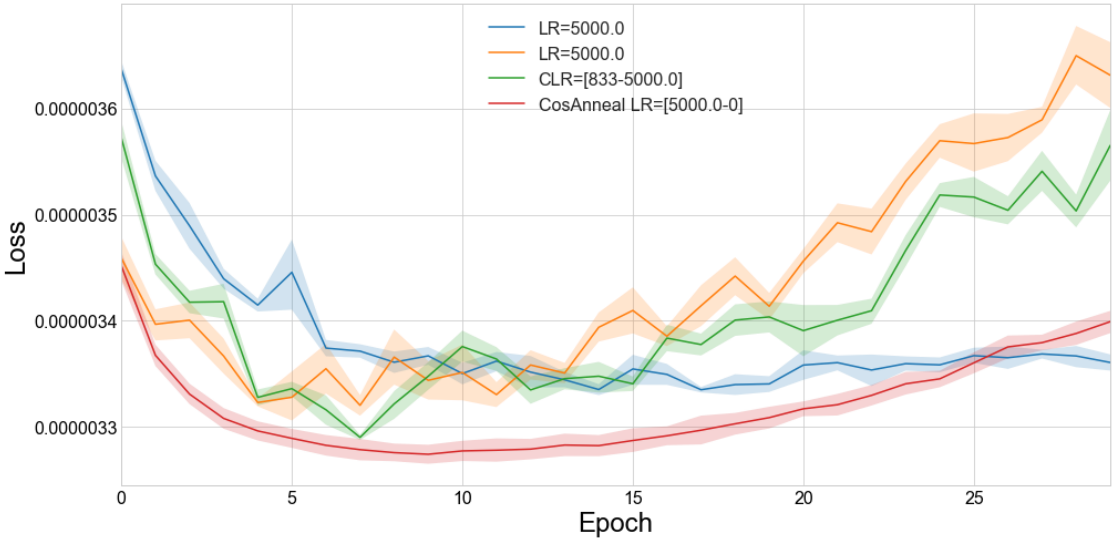
- Learning rates
  - Example of the choice the point of optimal complexity, before we reach overtraining:



- Cyclical learning rates

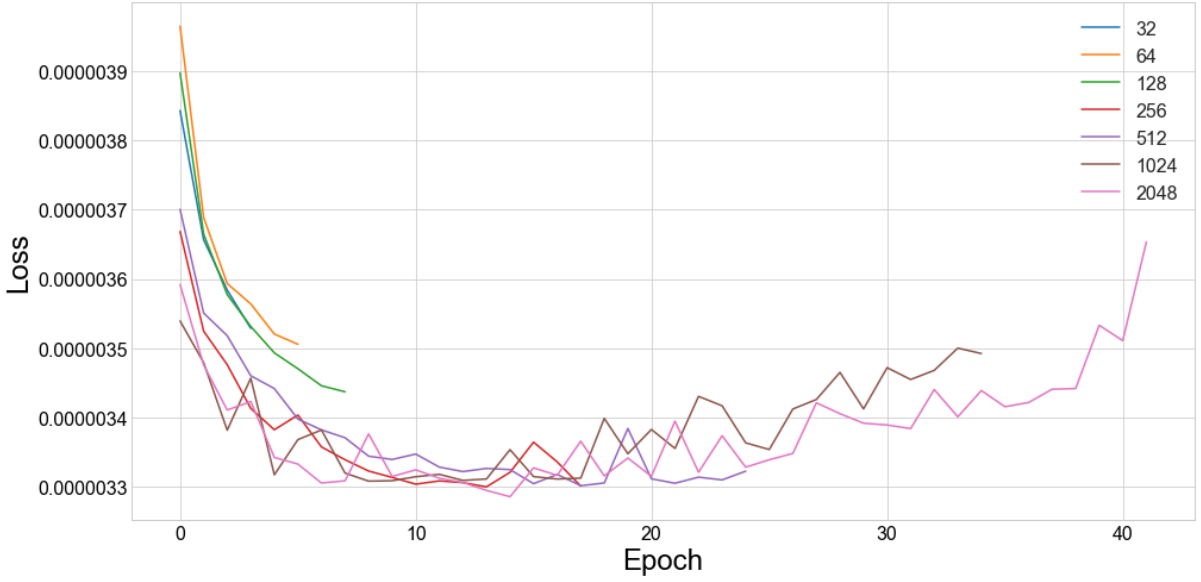


  - Cosine annealing with restarts: annealing the LR according to a cosine function, and then jumping up to the maximum LR at 0. Decaying the learning rate, allows one to make use of large initial learning-rates for quick convergence, and smaller learning-rates to arrive at the minima.



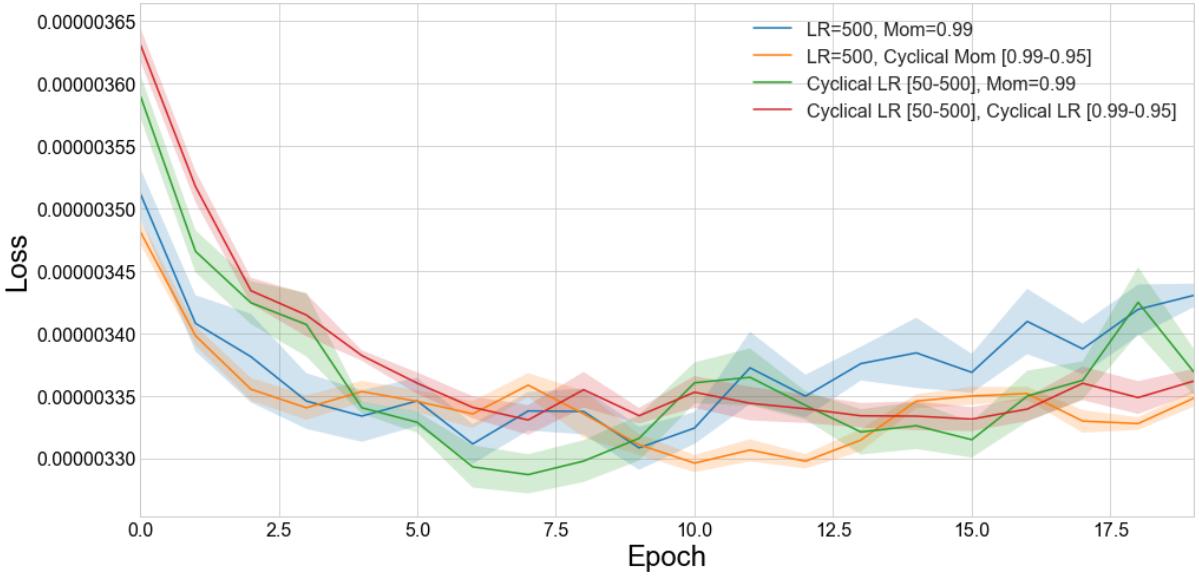
- Using cosine annealing with restarts (red), we're able to reach a slightly better loss than the linear cycle performance (green), in about the same time. However, the validation loss is much more stable (smoother line), which perhaps indicates that it is able to more easily find wider minima.

- Batch sizes



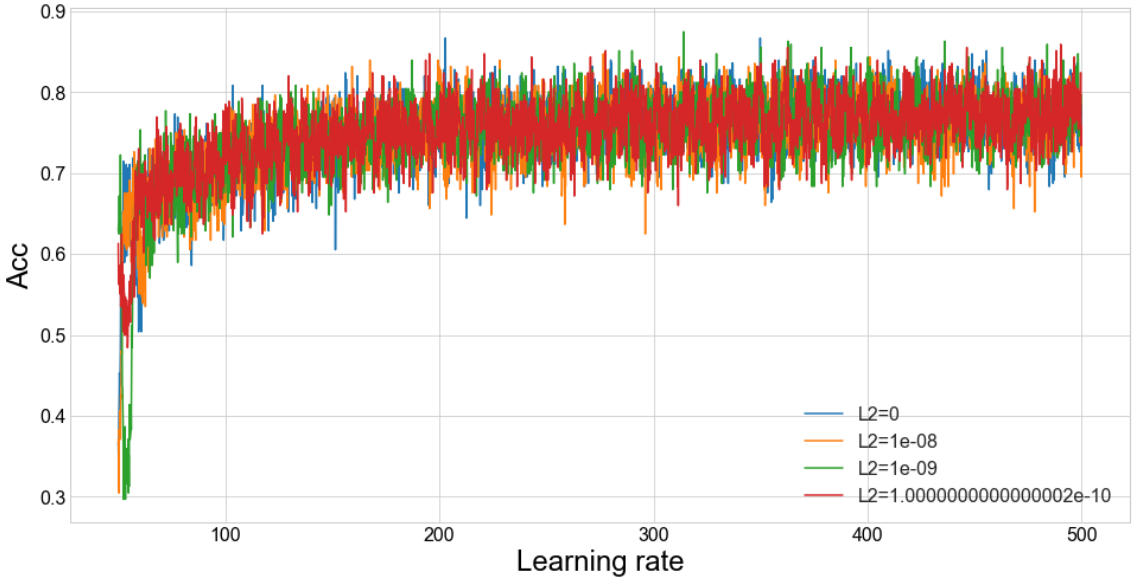
- Low batch sizes run very slowly (can only show a few epochs), and do not allow the use of high learning rates (slow convergence). As the batch size is increased, larger learning rates can be used, but the convergence becomes less stable. For batch sizes of 256 and above, the networks converge to about the same loss in the same number of epochs (although the higher batch sizes will do so quicker time-wise). BS=256 (red) appears to offer the optimal point between stability and convergence time, for this dataset, architecture, and computer.

- Cyclical momentum



  - Comparing the best performing setups from each schedule configuration it seems that of the hyperparameters tested, for this dataset and architecture, a cycled LR with a constant momentum (green) provides the lowest loss, but eventually overfits.
  - Cycling the momentum and keeping the LR constant (orange) reaches almost as good a loss, but after 40% more epochs, and although it later provides less overfitting, it does suffer from regular peaks and troughs due to the cycling.
  - Cycling both the LR and the momentum (red) causes convergence in the same number of epochs as (orange), but at a higher loss. Having reached its minimum, the test loss then remains flat, possibly indicating that with further adjustments of the hyperparameters it might provide superior performance to (orange).

- Weight decay



-

- Although by the end of the learning rate cycle, all four L2 setups converge to about the same point, the architecture, data, learning rate, and momentum we've chosen generally favours lower values for L2, and even at L2=1e-10 (red), the network without L2 (blue) still provides a slightly better loss, therefore it is probably safe to assume that L2 regularisation can safely be switched off for this setup.

NEXT MEETING

Next meeting will be held in September/October - date to be defined.

Volunteers for the journal club are very welcome - please contact us if you would be willing to present in one of the next sessions.