Project Proposal : Deep Mind V2

Randall Balestriero
Ecole Normale Superieure
Paris, France
Email: randallbalestriero@gmail.com

Romain Cosentino
Ecole Normale Superieure
Paris, France
Email: rom.cosentino@gmail.com

***Abstract*—Deepmind has been a breakthrough thanks to its architecture (DQN) combining deep learning techniques and reinforcement learning. This network is now able to play better than expert human in some games (from ATARI database). The overall algorithm combines state-of-the-art techniques used in supervised classification challenges and Q-learning methods. The Q-learning here becomes the trainer/teacher so that everything else can be trained and updated as in a supervised case. It is now possible to play to lots of games by learning how to play it by experience only. Even though the performances are beyond what an human can do for some games, they are not consistent for all kind of games (especially when the complexity increases with more actions or open environments).**

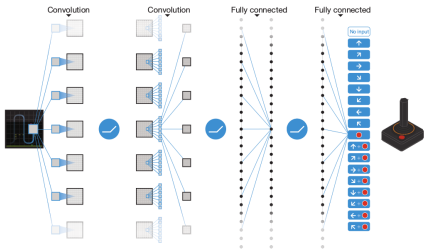

Fig. 1. Global Architecture used by Google [3]

I. What we propose

A. Problem Statement

The main limitations come when dealing with open games or environments where the agent needs to move in addition of other tasks. We can see the current structure as :

- CNN+CNN : find the best representation of the state space (the screen) by learning the right weights to have a better representation of s in this new feature space more tractable (otherwise s would be of infinite dimension).
- MLP+SoftMax Layer : approximate the non linear $Q(s,a)$ function thus defining the policy since in the end the optimal policy should be $\pi(s) = argmax_a Q(s,a)$. The strength here comes from the Kolmogorov theorem saying that in theory such ANN should be able to approximate any kind of non linear function (since the aim is to approximate the Q function).

B. Moving to parallel MLPs and double MDPs

But if we think about it deep mind works really well on games that can be played with one hand (when played by a human player). So the set of action state is small and the agent interactions are really simple (they often have no effects on the overall of the environment). But when dealing with more complex games (FPS-like games or open games), both hands are required for a human to effectively play it. This truly brings a new dimension to the problem and so should be structured as such in the whole architecture. In fact, deep mind performances are barely above the case where the agent takes only random actions all the time (thus there is a clear gape in the scores from game to game).

In fact it is now about taking decision for both hands (double action space) at the same time and on how to combine them to create synergies in the actions to maximize the reward.

We thus propose a parallel architecture with the property that the inputs are the same (analogue to the visual cortex which uses always the same filters but then dispatch the tasks from there). So we first have a sparse coding technique (which could be CNNs,) followed directly by two parallel and independent MLP which correspond to the non parametric estimation of two different Q functions. We will then see the possible changes that could be made (CNN $\rightarrow$ RBM, MLP $\rightarrow$ RNN, parallel and independent MDP $\rightarrow$ linked RNN, double MDP). The use of RNN would mean that $a(t)$ depends not only on $s(t)$ but also on $a(t-1)$ and thus $s(t-1)$.

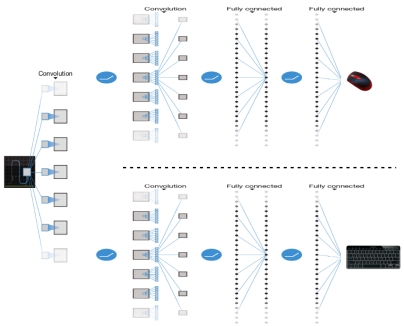

Fig. 2. New proposed architecture

C. Training and Reward Function

In addition the training is a key part. We propose a strategy based on deep mind : we play a complete scenario using the current optimal policy (actual weights) and sometime taking random actions (ϵ-greedy approach) until the end of the game (win/loose, points, ...). Then with the compilation of states and actions that were taken we propagate back the reward with the discount factor γ. Weighting will mean that if we loose, the

worst reward will be for the tuple right before the end of the scenario and so on. This will give us $r(t)\ \forall t$. From this we can apply a standard supervised learning using the Q-learning as the teacher to train the networks. Let's take an example, for scenario i the final reward is R at time T and so the reward for each state t prior to the end would be given by

$$r(t) = \gamma^{T-t} \times R$$

Thus the discount factor γ defines how the reward propagates through time. In fact it is natural to think that the loosing move is closer to the moment you lost rather than far before. It is also analogue to actualisation computation in finance for asset investment and portfolio management.

D. Learning Process and Experience Replay

Beyond the architecture another problem is the learning process. In the presented approach the experience replay was used. A set D was created with tuples $(s(t), r(t), a(t), s(t+1))$. Then a random subset of D was extracted uniformly and used as a mini batch for the network training. If we compare it to a supervised task, the problem is that the classes are not homogeneous for complex games. In fact for simple games (such as Pong) we empirically have actions (up or down here) following a uniform distribution(0.5 in this case)but for more complex games, the probability of doing some actions will be much smaller than other because they will correspond to optimal $Q(s, a)$ case naturally less present in the scenario/situations encountered during training. Thus, there will be a representation problem of the function $Q(s, a)$ which we aim to approximate. The result is that when using mini batches from the set D sampled with a uniform distribution, over-fitting is happening toward the most common scenarios which is not what will allow the best Q-learning.

Thus we would like to first perform some unsupervised clustering on the tuples in D to bring an homogeneous representation of all the situations to avoid this over fitting. So we could use techniques so that we still use experience replay but in a more built fashion to bring back class homogeneity, for example with a GMM prior for the sampling based on the proportional presence of each cluster. This problematic of class representation is crucial in deep learning and should be considered as such.

E. Conclusion

We propose solutions based on biology and studies. We aim to have a more stable and less complex model while being more meaningful. The aim is to be more realistic and less over fitted for 2D games such as pong or tetris. We reach human like thinking and learning process. We solve problems related to Q-learning with neural-nets (decorrelation of the training examples, improving class representation,...). In fact, all kinds of deep learning problematic will be present from unsupervised feature space learning to online supervised non linear function approximation.

REFERENCES

[1] S. Adam, L. Busoniu, and R. Babuska. Experience replay for real-time reinforcement learning control. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 42(2):201–212, March 2012.

[2] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 689–696, New York, NY, USA, 2009. ACM.

[3] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 02 2015.

[4] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *International Conference on Machine Learning (ICML)*, Bellevue, USA, June 2011.

[5] L.Enrique Sucar. Parallel markov decision processes. In Peter Lucas, JosA. Gmez, and Antonio Salmern, editors, *Advances in Probabilistic Graphical Models*, volume 214 of *Studies in Fuzziness and Soft Computing*, pages 295–309. Springer Berlin Heidelberg, 2007.