dBaby: Grounded Language Teaching through Games and Efficient Reinforcement Learning

Guntis Barzdins
AiLab at IMCS
University of Latvia
Riga LV-1459, Latvia
guntis.barzdins@lu.lv

Renars Liepins
Innovation Labs
LETA
Riga LV-1050, Latvia
renars.liepins@leta.lv

Paulis F. Barzdins
School of Informatics
University of Edinburgh
Edinburgh EH8 9AB, UK
s1768357@sms.ed.ac.uk

Didzis Gosko
Liquid Studio
Accenture
Riga LV-1039, Latvia
didzis.gosko@accenture.com

Abstract

This paper outlines a project proposal to be submitted to EC H2020 call ICT-29-2018. The purpose of the project is to create a digital Baby (dBaby) - an agent perceiving and interacting with the 3D world and communicating with its Teacher via natural language phrases to achieve the goals set by the Teacher. The novelty of the approach is that neither language nor visual capabilities are hard-coded in dBaby - instead, the Teacher defines a language learning Game grounded in the 3D world, and dBaby learns the language as a byproduct of the reinforcement learning from the raw pixels and character strings while maximizing the rewards in the Game. So far such approach successfully has been demonstrated only in the virtual 3D world with pre-programmed Games where it requires millions of episodes to learn a dozen words. Moving to human Teacher and real 3D environment requires an order-of-magnitude improvement to data-efficiency of the reinforcement learning. A novel Episodic Control based pre-training is demonstrated as a promising approach for bootstrapping the data-efficient reinforcement learning.

1 Introduction

The digital Baby (dBaby) project proposal to EC H2020 call ICT-29-2018 outlined in this paper is a follow-up to the already running ICT-16-2015 project SUMMA.[1] During the SUMMA project, it has become apparent [1] that the current state-of-art NLP approaches to ASR, MT, NER, NEL, AMR, KBP, Summarizing are suitable only for gisting purposes but will never achieve true natural language understanding expected by the project user-partners BBC and DW for the news-monitoring use-case. The current statistical NLP tools that learn from text-only corpora suffer from the symbol grounding problem [2] and thus are constrained by the training corpora inter-annotator agreement ratio. The deep learning methods have contributed vastly to approaching the corpora inter-annotator agreement ratio, but they are helpless to overcome it.

Genuine natural language understanding requires a different approach. Grounded language learning through reinforcement learning (RL) has recently emerged [3] as a promising alternative and is at the core of the dBaby project proposal. The attractiveness of this approach is strengthened by the

[1] http://summa-project.eu

ViGIL at 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

deep reinforcement learning recently demonstrating super-human capabilities both in Atari games [4] and the very challenging game of GO [5]. This progress allows to speculate that someday dBaby might achieve super-human fluency in natural language understanding and generation, although in this project we aim only to demonstrate the basic viability of the approach.

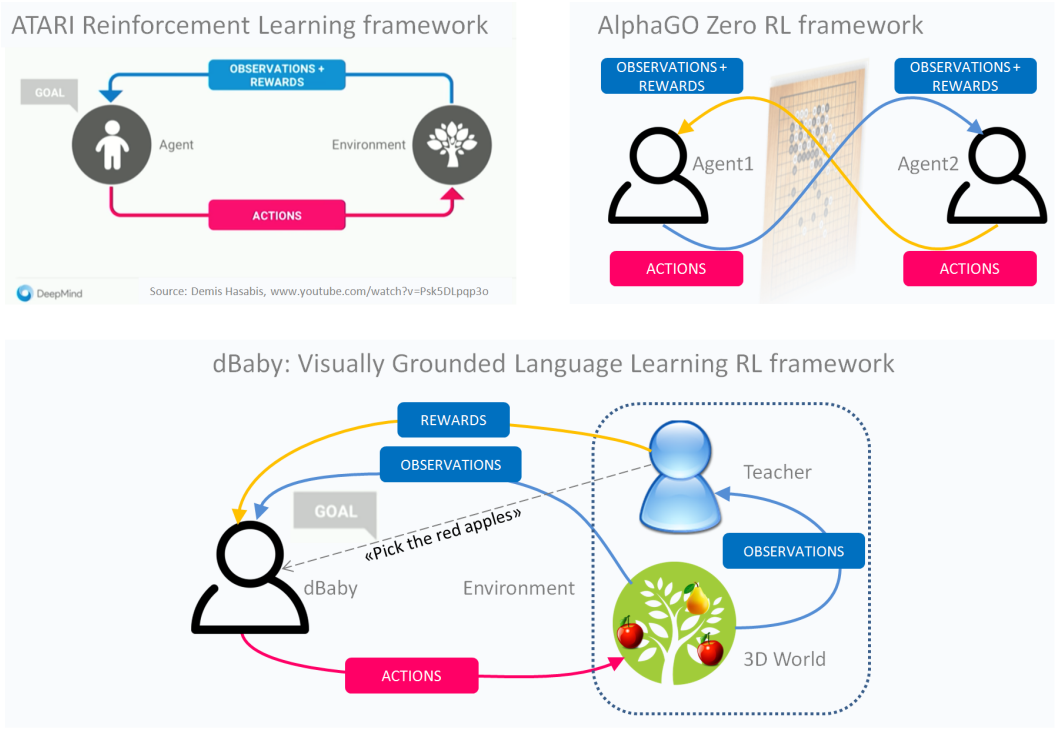

Figure 1: Three reinforcement learning frameworks. (a) Atari, (b) AlphaGO, (c) dBaby.

Figure 1 illustrates how the dBaby approach relates to and differs from other reinforcement learning frameworks. dBaby is a reinforcement learning agent perceiving and interacting with the 3D world and communicating with its Teacher via written natural language instructions to achieve the goals set by the Teacher. The novelty of the approach is that neither language nor visual capabilities are hard-coded in dBaby - instead, the Teacher defines a language learning Game grounded in the 3D world, and dBaby learns the language as a byproduct of the reinforcement learning from raw pixels and character strings while maximizing the rewards in the Game. The dBaby framework differs from the previous approaches in that Environment is split into 3D world and Teacher, where the Teacher is the one defining the goal in natural language and providing the rewards according to the set goal. In this way, the Game to be learned by dBaby is effectively defined by the Teacher, and 3D world merely acts as a "visual language" in which the natural language gets grounded.

The goal of the Game is conveyed by the Teacher to dBaby as a character string (e.g. "Pick the red apples"). dBaby treats this character string as part of the observation from the environment along with the raw pixels observed from the 3D world. The reason why in Figure 1 (c) we separate the goal from the observation is that they come from different sources - from the Teacher and from the 3D world respectively and thus play rather different roles in the overall RL framework. It is interesting to note that the concept of a "goal" separate from "observation" appears already in Figure 1 (a) depicting classic Atari reinforcement learning as presented by D.Hasabis[2] [4], where he describes it as "agent finds itself in some sort of environment and it is trying to achieve a goal in that environment ... Goal is simply to maximize the score". In case of the dBaby framework the "goal" is any natural language phrase spelled by the Teacher, so it can be either "Maximize the score" or it can be a more high-level goal like "Pick the red apples".

[2]http://youtu.be/Psk5DLpqp3o

So far the dBaby approach has successfully been demonstrated only in the virtual 3D world with pre-programmed Games [3] where it requires millions of episodes to learn a dozen words. Section 2 illustrates the dBaby approach and shows that moving to a human Teacher and a real 3D environment or self-play between dBabies requires an order-of-magnitude improvement to data-efficiency of the reinforcement learning. In Section 3 we present Episodic Control (EC) [6] as a promising approach towards bootstrapping the data-efficient reinforcement learning.

2 dBaby Baseline Implementation and Scaling Options

We refer in this paper to [3] as the baseline implementation of the dBaby framework despite that paper not using the terms "dBaby" and "Teacher" for the grounded language learning task equivalent to Figure 1 (c). An interesting aspect of this baseline implementation is that the agent (dBaby) not only listens to the goals expressed as the natural language phrases by Teacher, but it is also able to output the names of objects it sees, eventually leading to the two-way natural language communication. Although this aspect is not essential for the core dBaby framework illustrated in Figure 1 (c), eventually it is of interest for natural language generation and for self-play where one dBaby acts as a Teacher for another dBaby similar to Figure 1 (b).

As was demonstrated in the baseline implementation[3] it is possible to create a pre-programmed Teacher to teach the dBaby the embodied meaning of simple sentences like "pick the blue object" or "pick the red TV next to a green object in the magenta room" in the simulated 3D world. Using a gradual curriculum it would likely be possible to extend the pre-programmed Teacher for teaching more extended tasks like moving objects around or placing objects into some arrangements. However scaling a pre-programmed Teacher to even more varied or random tasks would be difficult, because the Teacher not only has to recognize when a goal has been achieved, but also needs to describe the goal in clear natural language through some language generation module. Language generation might be handled by self-play between the Teacher and dBaby [7] and eventually might lead to super-fluent language skills of Teacher and dBaby, but it would be difficult to force this generated language to be close to natural English.

Therefore a more attractive approach in the near term would be scaling the dBaby framework towards human Teacher and real 3D environment. This would require an order-of-magnitude improvement to the data-efficiency of the reinforcement learning because neither a human Teacher nor robot in a real 3D environment can sustain millions of training episodes.

3 Data-Efficient Reinforcement Learning

dBaby baseline implementation [3] uses A3C approach [8] for deep reinforcement learning. To integrate visual and textual inputs, the A3C neural network is preceded there by the Visual embedding and Word embedding networks, the low-dimensional outputs of which are concatenated before feeding into the A3C network. To improve the data-efficiency of the combined network, the embedding networks there are pre-trained through unsupervised auxiliary tasks. Particularly, the Visual embedding is pre-trained with temporal AutoEncoder (tAE) [9] approach having objective to predict the visual environment state from the previous states and actions taken by the agent.

In this section we propose an additional Episodic Control (EC) [10, 6] pre-training technique which can be used together with tAE to further improve the data-efficiency of the reinforcement learning. The combined pre-training data-efficiency gains in Figure 2 are demonstrated with simplified (policy gradient) version [11] of A3C on the Atari Pong game, under the assumption that similar data-efficiency gains are likely also in the 3D environment of the dBaby framework.

In the first row of Figure 2 two randomly initialized baselines are shown. The graph (a) shows the plain policy gradient reinforcement learning [11] performance on the raw pixels (8400dim) input from the Pong game. The graph (b) shows the performance of the same policy gradient reinforcement learning method when the raw pixel input is replaced by the low-dimensional (80dim) tAE output; this setup corresponds closely to the baseline in [3] and illustrates the exceptional quality of tAE output despite 100x lower dimensionality than the raw pixels.

[3] See video at http://youtu.be/wJjdu1bPJ04

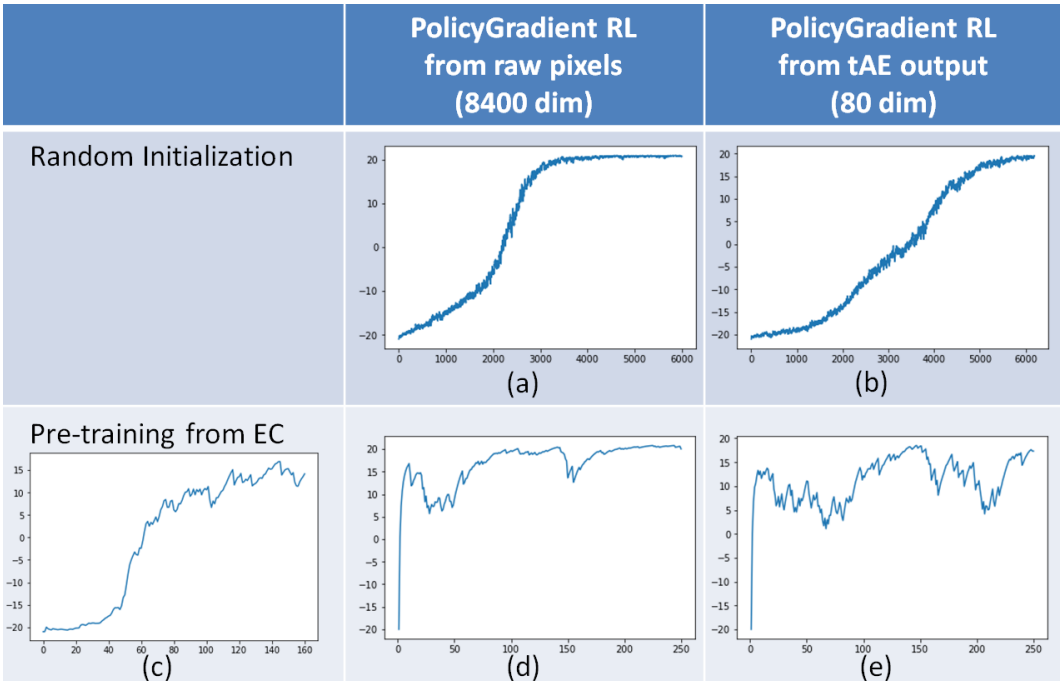

Figure 2: Data-efficiency of reinforcement learning with the combined tAE and EC pre-training on Atari Pong game (episode count on the horizontal axis, score achieved on the vertical axis).

In the second row of Figure 2 the random initialization is replaced by the supervised pre-training using game-play recordings from 160 episodes of the notoriously data-efficient [10] Episodic Control reinforcement learning (c). The resulting graphs (d) and (e) show major data-efficiency gains compared to random initialization in (a) and (b).

Due to space limitations we are omitting here the implementation details for this novel method, and refer to the actual code provided online.[4] All components of the integrated solution are described in the publications referenced above and the novelty is mostly in the way they are combined. The key trick enabling the exceptional performance of the integrated solution is that only the positive examples from the EC game-play (state/action sequences leading to positive reward in the short term) are used for supervised pre-training of RL weights. Another novelty is the use of more efficient Cosine similarity for kNN calculation in the EC implementation along with a sequential recording of all states (including duplicates) in the Episodic Memory thus preventing forgetting of any past experiences.

4 Conclusions

As a project proposal outline this paper clearly is not a finished work - it is rather a perspective on future directions in the grounded language learning. Nevertheless, we hope it being fruitful sharing our perspective already in this early proposal preparation stage.

Acknowledgments

This work was supported in part by the Latvian National research program SOPHIS under grant agreement Nr.10-4/VPP-4/11 and in part by H2020 SUMMA project under grant agreement 688139/H2020-ICT-2015.

[4] https://github.com/LUMII-AILab/dBaby

References

[1] Renars Liepins, Ulrich Germann, Guntis Barzdins, Alexandra Birch, Steve Renals, Susanne Weber, Peggy van der Kreeft, Hervé Bourlard, João Prieto, Ondrej Klejch, Peter Bell, Alexandros Lazaridis, Alfonso Mendes, Sebastian Riedel, Mariana S. C. Almeida, Pedro Balage, Shay B. Cohen, Tomasz Dwojak, Philip N. Garner, Andreas Giefer, Marcin Junczys-Dowmunt, Hina Imran, David Nogueira, Ahmed M. Ali, Sebastião Miranda, Andrei Popescu-Belis, Lesly Miculicich Werlen, Nikos Papasarantopoulos, Abiola Obamuyide, Clive Jones, Fahim Dalvi, Andreas Vlachos, Yang Wang, Sibo Tong, Rico Sennrich, Nikolaos Pappas, Shashi Narayan, Marco Damonte, Nadir Durrani, Sameer Khurana, Ahmed Abdelali, Hassan Sajjad, Stephan Vogel, David Sheppey, Chris Hernon, and Jeff Mitchell. The summa platform prototype. In *EACL*, 2017.

[2] Stevan Harnad. The symbol grounding problem. *arXiv*, cs.AI/9906002, 1990.

[3] Karl Moritz Hermann, Felix Hill, Simon Green, Fumin Wang, Ryan Faulkner, Hubert Soyer, David Szepesvari, Wojciech Czarnecki, Max Jaderberg, Denis Teplyashin, Marcus Wainwright, Chris Apps, Demis Hassabis, and Phil Blunsom. Grounded language learning in a simulated 3d world. *arXiv*, abs/1706.06551, 2017.

[4] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas 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 7540:529–33, 2015.

[5] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas R Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy P. Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550 7676:354–359, 2017.

[6] Alexander Pritzel, Benigno Uria, Sriram Srinivasan, Adrià Puigdomènech Badia, Oriol Vinyals, Demis Hassabis, Daan Wierstra, and Charles Blundell. Neural episodic control. In *ICML*, 2017.

[7] Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. *arXiv*, abs/1703.04908, 2017.

[8] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *ICML*, 2016.

[9] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L. Lewis, and Satinder P. Singh. Action-conditional video prediction using deep networks in atari games. In *NIPS*, 2015.

[10] Charles Blundell, Benigno Uria, Alexander Pritzel, Yazhe Li, Avraham Ruderman, Joel Z. Leibo, Jack Rae, Daan Wierstra, and Demis Hassabis. Model-free episodic control. *arXiv*, abs/1606.04460, 2016.

[11] Andrej Karpathy. Deep reinforcement learning: Pong from pixels, 2017. `http://karpathy.github.io/2016/05/31/rl`.