Crossmodal Attentive Skill Learner

Shayegan Omidshafiei
Laboratory for Information
and Decision Systems
MIT
Cambridge, MA 02139
shayegan@mit.edu

Dong-Ki Kim
Laboratory for Information
and Decision Systems
MIT
Cambridge, MA 02139
dkkim93@mit.edu

Jason Pazis
Amazon Alexa*
Cambridge, MA 02142
pazisj@amazon.com

Jonathan P. How
Laboratory for Information
and Decision Systems
MIT
Cambridge, MA 02139
jhow@mit.edu

Abstract

This paper presents the Crossmodal Attentive Skill Learner (CASL), integrated with the recently-introduced Asynchronous Advantage Option-Critic (A2OC) architecture [Harb et al., 2017] to enable hierarchical reinforcement learning across *multiple* sensory inputs. We provide concrete examples where the approach not only improves performance in a single task, but accelerates transfer to new tasks. We demonstrate the attention mechanism anticipates and identifies useful latent features, while filtering irrelevant sensor modalities during execution. We modify the Arcade Learning Environment [Bellemare et al., 2013] to support audio queries, and conduct evaluations of crossmodal learning in the Atari 2600 game Amidar. Finally, building on the recent work of Babaeizadeh et al. [2017], we open-source a fast hybrid CPU-GPU implementation of CASL.[1]

1 Introduction

Intelligent agents should be capable of disambiguating local sensory streams to realize long-term goals. In recent years, the combined progress of computational capabilities and algorithmic innovations has afforded reinforcement learning (RL) [Sutton and Barto, 1998] approaches the ability to achieve this desiderata in impressive domains, exceeding expert-level human performance in tasks such as Atari and Go [Mnih et al., 2015, Silver et al., 2017]. Nonetheless, many of these algorithms thrive primarily in well-defined mission scenarios learned in isolation from one another; such monolithic approaches are not sufficiently scalable for missions where goals may be less clearly defined, and sensory inputs found salient in one domain may be less relevant in another.

How should agents learn effectively in domains of high dimensionality, where tasks are durative, agents receive sparse feedback, and sensors compete for limited computational resources? One promising avenue is hierarchical reinforcement learning (HRL), focusing on problem decomposition for learning transferable skills. Temporal abstraction enables exploitation of domain regularities to provide the agent hierarchical guidance in the form of options or sub-goals [Sutton et al., 1999, Kulkarni et al., 2016]. Options help agents improve learning by mitigating scalability issues in long-duration

[*] Work was done prior to Amazon involvement of author, and does not reflect views of the Amazon company.
[1] Code: https://github.com/shayegano/CASL

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

missions, by reducing the effective number of decision epochs. In the parallel field of supervised learning, temporal dependencies have been captured proficiently using attention mechanisms applied to encoder-decoder based sequence-to-sequence models [Bahdanau et al., 2014, Luong et al., 2015]. *Attention* empowers the learner to focus on the most pertinent stimuli and capture longer-term correlations in its encoded state, for instance to conduct neural machine translation or video captioning [Yeung et al., 2015, Yang et al., 2016]. Recent works also show benefits of spatio-temporal attention in RL [Mnih et al., 2014, Sorokin et al., 2015].

One can interpret the above approaches as conducting dimensionality reduction, where the target dimension is *time*. In view of this insight, this paper proposes an RL paradigm exploiting hierarchies in the dimensions of *time* and *sensor modalities*. Our aim is to learn rich skills that attend to and exploit pertinent crossmodal (multi-sensor) signals at the appropriate moments. The introduced crossmodal skill learning approach largely benefits an agent learning in a high-dimensional domain (e.g., a robot equipped with many sensors). Instead of the expensive operation of processing and/or storing data from all sensors, we demonstrate that our approach enables such an agent to focus on important sensors; this, in turn, leads to more efficient use of the agent's limited computational and storage resources (e.g., its finite-sized memory).

In this paper, we focus on combining two sensor modalities: audio and video. While these modalities have been previously used for supervised learning [Ngiam et al., 2011], to our knowledge they have yet to be exploited for crossmodal skill learning. We provide concrete examples where the proposed HRL approach not only improves performance in a single task, but accelerates transfer to new tasks. We demonstrate the attention mechanism anticipates and identifies useful latent features, while filtering irrelevant sensor modalities during execution. We also show preliminary results in the Arcade Learning Environment [Bellemare et al., 2013], which we modified to support audio queries. In addition, we provide insight into how our model functions internally by analyzing the interactions of attention and memory. Building on the recent work of Babaeizadeh et al. [2017], we open-source a fast hybrid CPU-GPU implementation of our framework. Finally, note that despite this paper's focus on audio-video sensors, the framework presented is general and readily applicable to additional sensory inputs.

2 Background

POMDPs This work considers an agent operating in a partially-observable stochastic environment, modeled as a POMDP $\langle \mathbb{S}, \mathbb{A}, \mathbb{O}, \mathcal{T}, \mathcal{O}, \mathcal{R}, \gamma \rangle$ [Kaelbling et al., 1998]. $\mathbb{S}$, $\mathbb{A}$, and $\mathbb{O}$ are, respectively, the state, action, and observation spaces. At timestep t, the agent executes action $a \in \mathbb{A}$ in state $s \in \mathbb{S}$, transitions to state $s' \sim \mathcal{T}(s, a, s')$, receives observation $o \sim \mathcal{O}(o, s', a)$, and reward $r_t = \mathcal{R}(s, a) \in \mathbb{R}$. The value of state s under policy $\pi : Dist(\mathbb{S}) \rightarrow \mathbb{A}$ is the expected return $V_\pi(s) = \mathbb{E}[\sum_{t'=t}^{H} \gamma^{t'-t} r_{t'}]$, given horizon H and discount factor $\gamma \in [0, 1)$. The objective is to learn an optimal policy π^*, which maximizes the value.

Options The framework of *options* provides an RL agent the ability to plan using temporally-extended actions [Sutton et al., 1999]. Option $\omega \in \Omega$ is defined by initiation set $\mathbb{I} \subseteq \mathbb{S}$, intra-option policy $\pi_\omega : \mathbb{S} \rightarrow \text{Dist}(\mathbb{A})$, and termination condition $\beta_\omega : \mathbb{S} \rightarrow [0, 1]$. Initially, a policy over options π_Ω chooses an option among those that satisfy the initiation set. The selected option executes its intra-option policy until termination, upon which a new option is chosen. This process iterates until the goal state is reached. Recently, the Asynchronous Advantage Actor-Critic framework (A3C) [Mnih et al., 2016] has been applied to POMDP learning in a computationally-efficient manner by combining parallel actor-learners and Long Short-Term Memory (LSTM) cells [Hochreiter and Schmidhuber, 1997]. Asynchronous Advantage Option-Critic (A2OC) extends A3C and enables learning option-value functions, intra-option policies, and termination conditions in an end-to-end fashion [Harb et al., 2017]. The option-value function models the value of state $s \in \mathbb{S}$ in option $\omega \in \Omega$,

$$Q_\Omega(s, \omega) = \sum_a \pi_\omega(a|s)\Big(r(s, a) + \gamma \sum_{s'} \mathcal{T}(s'|s, a) U(s', \omega)\Big), \tag{1}$$

where $a \in \mathbb{A}$ is a primitive action and $U(s', \omega)$ represents the option utility function,

$$U(s', \omega) = (1 - \beta_\omega(s')) Q_\Omega(\omega, s') + \beta_\omega(s')(V_\Omega(s') - c). \tag{2}$$

A2OC introduces deliberation cost, c, in the utility function to address the issue of options terminating too frequently. Intuitively, the role of c is to impose an added penalty when options terminate, leading them to terminate less frequently. The value function over options, V_Ω, is defined,

$$V_\Omega(s') = \sum_\omega \pi_\Omega(\omega|s') Q_\Omega(\omega, s'), \tag{3}$$

where π_Ω is the policy over options (e.g., an epsilon-greedy policy over Q_Ω). Assuming use of a differentiable representation, option parameters are learned using gradient descent.

3 Approach

Our goal is to design a mechanism that enables the learner to modulate high-dimensional sensory inputs, focusing on pertinent stimuli that may lead to more efficient skill learning. This section presents motivations behind attentive skill learning, then introduces the proposed framework.

3.1 Attentive Mechanisms

Before presenting the proposed architecture, let us first motivate our interests towards attentive skill learning. One might argue that the combination of deep learning and RL already affords agents the representation learning capabilities necessary for proficient decision-making in high-dimensional domains; i.e., why the need for crossmodal attention?

Our ideas are motivated by the studies in behavioral neuroscience that suggest the interplay of attention and choice bias humans' value of information during learning, playing a key factor in solving tasks with high-dimensional information streams [Leong et al., 2017]. Works studying learning in the brain also suggest a natural pairing of attention and hierarchical learning, where domain regularities are embedded as priors into skills and combined with attention to alleviate the curse of dimensionality [Niv et al., 2015]. Works also suggest attention plays a role in the intrinsic curiosity of agents during learning, through direction of focus to regions predicted to have high reward [Mackintosh, 1975], high uncertainty [Pearce and Hall, 1980], or both [Pearce and Mackintosh, 2010].

In view of these studies, we conjecture that crossmodal attention, in combination with HRL, improves representations of relevant environmental features that lead to superior learning and decision-making. Specifically, using crossmodal attention, agents combine internal beliefs with external stimuli to more effectively exploit multiple modes of input features for learning. As we later demonstrate, our approach captures temporal crossmodal dependencies, and enables faster and more proficient learning of skills in the domains examined.

3.2 Crossmodal Attentive Skill Learner

We propose Crossmodal Attentive Skill Learner (CASL), a novel end-to-end framework for HRL. One may consider many blueprints for integration of multi-sensory attention into the options framework. Our proposed architecture is primarily motivated by the literature that taxonomizes attention into two classes: *exogeneous* and *endogeneous*. The former is an involuntary mechanism triggered automatically by the inherent saliency of the sensory inputs, whereas the latter is driven by the intrinsic and possibly long-term goals, intents, and beliefs of the agent [Carrasco, 2011]. Previous attention-based neural architectures take advantage of both classes, for instance, to solve natural language processing problems [Vinyals et al., 2015]. Our approach follows this schema.

The CASL network architecture is visualized in Fig. 1. Let $M \in \mathbb{N}$ be the number of sensor modalities (e.g., vision, audio, etc.) and x_m denote extracted features from the m-th sensor, where $m \in \{1, \ldots, M\}$. For instance, x_m may correspond to feature outputs of a convolutional neural network given an image input. Given extracted features for all M sensors at timestep t, as well as hidden state h^{t-1}, the proposed crossmodal attention layer learns the relative importance of each

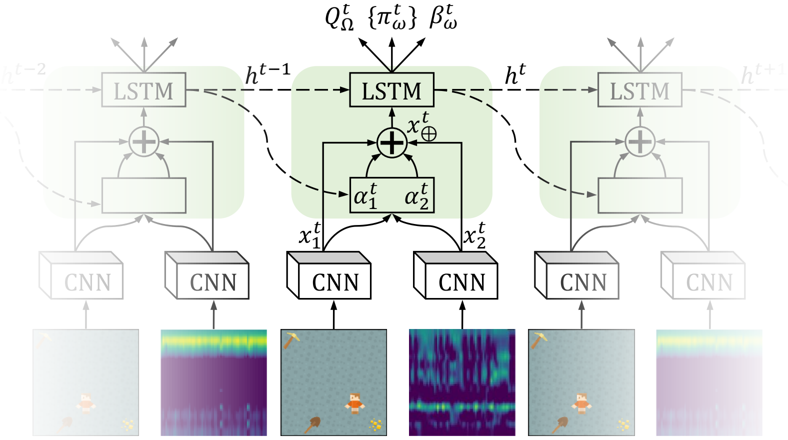


Figure 1: CASL network architecture enables attention-based learning over multi-sensory inputs. Green highlighted region indicates crossmodal attention LSTM cell, trained via backpropagation through time. We provide this attention wrapper as part of our open-source code release.

modality $\alpha^t \in \Delta^{M-1}$, where Δ^{M-1} is the $(M-1)$-simplex:

$$z^t = \tanh\Big(\underbrace{\sum_{m=1}^{M}(W_m^T x_m^t + b_m)}_{\text{Exogeneous attention}} + \underbrace{W_h^T h^{t-1} + b_h}_{\text{Endogeneous attention}}\Big) \tag{4}$$

$$\alpha^t = \text{softmax}\left(W_z^T z^t + b_z\right) \tag{5}$$

$$x_\oplus = \begin{cases} \sum_{m=1}^{M} \alpha_m^t x_m^t & \text{(Summed attention)} \\ \left[(\alpha_1^t x_1^t)^T, \ldots, (\alpha_m^t x_m^t)^T\right]^T & \text{(Concatenated attention)} \end{cases} \tag{6}$$

Weight matrices W_m, W_h, W_z and bias vectors b_m, b_h, b_z are trainable parameters and nonlinearities are applied element-wise.

Both exogeneous attention over sensory features x_m^t and endogeneous attention over LSTM hidden state h^{t-1} are captured in (4). The sensory feature extractor used in experiments consists of 3 convolutional layers, each with 32 filters of size 3 × 3, stride 2, and ReLU activations. Attended features $\alpha_m^t x_m^t$ may be combined via summation or concatenation (per (6)), then fed to an LSTM cell. The LSTM output captures temporal dependencies used to estimate option values, intra-option policies, and termination conditions (Q_Ω, π_ω, β_ω in Fig. 1, respectively),

$$Q_\Omega(s,\omega) = W_{Q,\omega} h^t + b_{Q,\omega}, \tag{7}$$

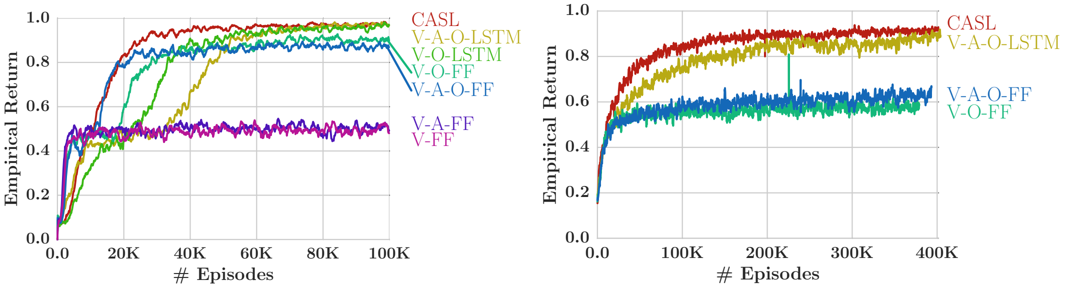
$$\pi_\omega(a|s) = \text{softmax}(W_{\pi,\omega} h^t + b_{\pi,\omega}), \tag{8}$$

$$\beta_\omega(s) = \sigma(W_{\beta,\omega} h^t + b_{\beta,\omega}), \tag{9}$$

where weight matrices $W_{\Omega,\omega}$, $W_{\pi,\omega}$, $W_{\beta,\omega}$ and bias vectors $b_{\Omega,\omega}$, $b_{\pi,\omega}$, $b_{\beta,\omega}$ are trainable parameters for the current option ω, and $\sigma(\cdot)$ is the sigmoid function. Network parameters are updated using gradient descent. Entropy regularization of attention outputs α^t was found to encourage exploration of crossmodal attention behaviors during training.

4 Evaluation

The proposed framework is evaluated on a variety of learning tasks with inherent reward sparsity and transition noise. We evaluate our approach in three domains: a door puzzle domain, a 2D-Minecraft like domain, and the Arcade Learning Environment [Bellemare et al., 2013]. These environments include challenging combinations of reward sparsity and/or complex audio-video sensory input modalities that may not always be useful to the agent. The first objective of our experiments is to analyze performance of CASL in terms of learning rate and transfer learning. The second objective is



(a) Door puzzle domain.

(b) Door puzzle transfer learning.

Figure 2: CASL improves learning rate compared to other networks. Abbreviations: **V**ideo, **A**udio, **O**ptions, **F**eed**F**orward net, **LSTM** net.

to understand relationships between attention and memory mechanisms (as captured in the LSTM cell state). Finally, we modify the Arcade Learning Environment to support audio queries, and evaluate crossmodal learning in the Atari 2600 game Amidar.

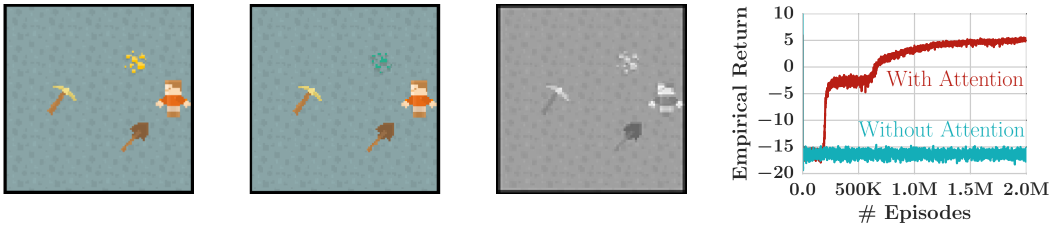
4.1 Crossmodal Learning and Transfer

We first evaluate crossmodal attention in a sequential door puzzle game, where the agent spawns in a 2D world with two locked doors and a key at fixed positions. The key type is randomly generated, and its observable color indicates the associated door. The agent hears a fixed sound (but receives no reward) when adjacent to the key, and hears noise otherwise. The agent must find and pick up the key (which then disappears), then find and open the correct door to receive $+1$ reward (with discount $\gamma = 0.99$). The game terminates upon opening of either door. The agent's sensory inputs x_m^t are vision (grayscale image) and audio spectrogram. This task was designed in such a way that audio is not necessary to achieve the task – the agent can certainly focus on learning a policy mapping from visual features to open the correct door. However, audio provides potentially useful signals that may accelerate learning, making this a domain of interest for analyzing the interplay of attention and sensor modalities.

Attention Improves Learning Rate Figure 2a shows ablative training results for several network architectures. The three LSTM-based skill learners (including CASL) converge to the optimal value. Interestingly, the network that ignores audio inputs (V-O-LSTM) converges faster than its audio-enabled counterpart (V-A-O-LSTM), indicating the latter is overwhelmed by the extra sensory modality. Introduction of crossmodal attention enables CASL to converge faster than all other networks, using roughly half the training data of the others. The feedforward networks all fail to attain optimal value, with the non-option cases (V-A-FF and V-FF) repeatedly opening one door due to lack of memory of key color. Notably, the option-based feedforward nets exploit the option index to implicitly remember the key color, leading to higher value. Interplay between explicit memory mechanisms and use of options as pseudo-memory may be an interesting line of future work.

Attention Accelerates Transfer We also evaluate crossmodal attention for transfer learning (Fig. 2b), using the more promising option-based networks. The door puzzle domain is modified to randomize the key position, with pre-trained options from the fixed-position variant used for initialization. All networks benefit from an empirical return jumpstart of 0.2 at the beginning of training, due to skill transfer. Once again, CASL converges fastest, indicating more effective use of the available audio-video data. While the asymptotic performance of CASL is only slightly higher than the V-A-O-LSTM network, the reduction in number of samples needed to achieve a high score (e.g, after 100K episodes) makes it advantageous for domains with high sampling cost.

Attention Necessary to Learn in Some Domains Temporal behaviors of the attention mechanism are also evaluated in a 2D Minecraft-like domain, where the agent must pick an appropriate tool (pickaxe or shovel) to mine either gold or iron ore (Figs. 3a to 3c). Critically, the agent observes identical images for both ore types, but unique audio features when near the ore, making long-term audio storage necessary for selection of the correct tool. The agent receives $+10$ reward for correct



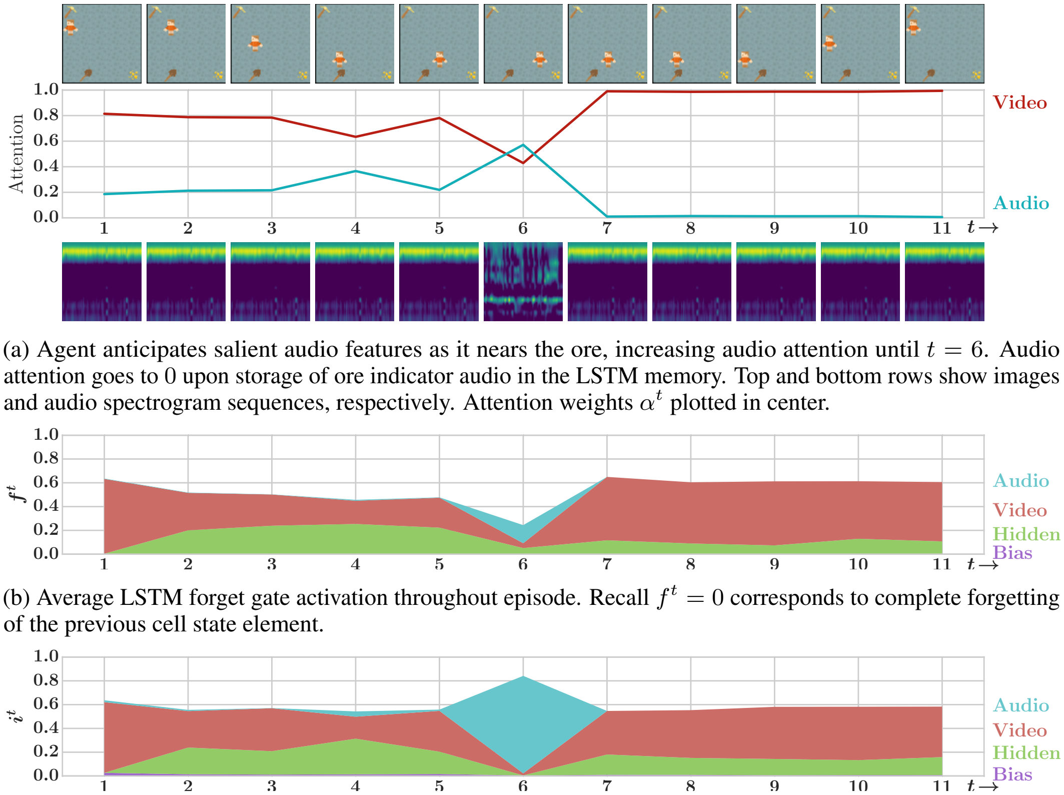
(a) Gold ore should be mined with the pickaxe.

(b) Iron ore should be mined with the shovel.

(c) Ore type indistinguishable by agent's visual input.

(d) Learning in mining domain.

Figure 3: Mining domain. Ore type is indistinguishable by grayscale visual input to agent's network.



(a) Agent anticipates salient audio features as it nears the ore, increasing audio attention until $t = 6$. Audio attention goes to 0 upon storage of ore indicator audio in the LSTM memory. Top and bottom rows show images and audio spectrogram sequences, respectively. Attention weights α^t plotted in center.

(b) Average LSTM forget gate activation throughout episode. Recall $f^t = 0$ corresponds to complete forgetting of the previous cell state element.

(c) Average LSTM input gate activation throughout episode. Recall $i^t = 1$ corresponds to complete throughput of the corresponding input element.

Figure 4: Interactions of crossmodal attention and LSTM memory. At $t = 6$, the attended audio input causes forget gate activation to drop, and the input gate activation to increase, indicating major overwriting of memory states. Relative contribution of audio to the forget and input activations drops to zero after the agent hears the necessary audio signal.

tool selection, -10 for incorrect selection, and -1 step cost. Compared to the door puzzle game, the mining domain is posed in such a way that the interplay between audio-video features is emphasized. Specifically, an optimal policy for this task must utilize both audio and video features: visual inputs enable detection of locations of the ore, agent, tools, whereas audio is used to identify the ore type.

Visual occlusion of the ore type, interplay of audio-video features, and sparse positive rewards cause the non-attentive network to fail to learn in the mining domain, as opposed to the attentive case (Fig. 3d). Figure 4a plots a sequence of frames where the agent anticipates salient audio features as it nears the ore at $t = 6$, gradually increasing audio attention, then sharply reducing it to 0 after hearing the signal.

Table 1: Preliminary results for learning in Atari 2600 games. The crossmodal attention learner, even *without* options, achieves high score for non-hierarchical methods. We emphasize these are not direct comparisons due to our method leveraging additional sensory inputs, but are meant to highlight the performance benefits of crossmodal learning.

Algorithm	Hierarchical?	Sensory Inputs	Score
Mnih et al. [2015]	✗	Video	739.5
Mnih et al. [2016]	✗	Video	283.9
Babaeizadeh et al. [2017]	✗	Video	218
Ours (*without* options)	✗	Audio & Video	$\boxed{900}$
Harb et al. [2017]	✓	Video	880.0
Vezhnevets et al. [2017]	✓	Video	$>$**2500**

4.2 Interactions of Attention and Memory

While the anticipatory nature of crossmodal attention in the mining domain is interesting, it also points to additional lines of investigation regarding interactions of attention and updates of the agent's internal belief (as encoded in its LSTM cell state). Specifically, one might wonder whether it is necessary for the agent to place any attention on the non-useful audio signals prior to timestep $t = 6$ in Fig. 4a, and also whether this behavior implies inefficient usage of its finite-size memory state.

Motivated by the above concerns, we conduct more detailed analysis of the interplay between the agent's attention and memory mechanisms as used in the CASL architecture (Fig. 1). Readers are referred to the appendix (Section 6.1) for details on how this analysis was conducted, as well as a brief overview of LSTM units. Given the sequence of audio-video inputs in Fig. 4a, we plot overall activations of the forget and input LSTM gates (averaged across all cell state elements), in Fig. 4b and Fig. 4c, respectively. Critically, these plots also indicate the relative influence of the *forget* and *input* LSTM gates' contributing variables (audio input, video input, hidden state, and bias term) to the overall activation.

Interestingly, prior to timestep $t = 6$, the contribution of audio to the forget gate and input gates is essentially zero, despite the positive attention on audio (in Fig. 4a). Recall a low forget gate activation corresponds to complete forgetting of the previous LSTM cell state element, whereas a high input gate activation corresponds to complete throughput of the corresponding input element. At $t = 6$, the forget gate activation drops, while the input gate experiences a sudden increase, indicating major overwriting of previous memory states with new information. Critically, the plots indicate that the attended audio input is the key contributing factor of both behaviors. In Fig. 4a, after the agent hears the necessary audio signal, it moves attention entirely to video; the contribution of audio to the forget and input activations also drops to zero. These behaviors indicate that the agent attends to audio in anticipation of an upcoming pertinent signal, but *chooses not to embed it into memory until the appropriate moment*. Attention filters irrelevant sensor modalities, given the contextual clues provided by exogeneous and endogeneous input features; it, therefore, enables the LSTM gates to focus on learning when and how to update the agent's internal state.

4.3 Early Experiments in the Arcade Learning Environment

Preliminary evaluation of crossmodal attention was conducted in the Arcade Learning Environment (ALE) [Bellemare et al., 2013]. We modified ALE to support audio queries, as it previously did not have this feature; we plan to add this code to the ALE repository.

Experiments were conducted in the Atari 2600 game Amidar (Table 1). This line of investigation considers impacts of crossmodal attention on Atari agent behavior, even without use of multiple (hierarchical) options; these results use CASL with a single option, hence tagged "non-hierarchical" in the table. Amidar was one of the games in which deep Q-networks failed to exceed human-level performance [Mnih et al., 2015]. The objective in Amidar is to collect rewards in a rectilinear maze while avoiding patrolling enemies. The agent is rewarded for painting segments of the maze, killing enemies at opportune moments, or collecting bonuses. Background audio plays throughout the game, and specific audio signals play when the agent crosses previously-unseen segment vertices. Figure 5

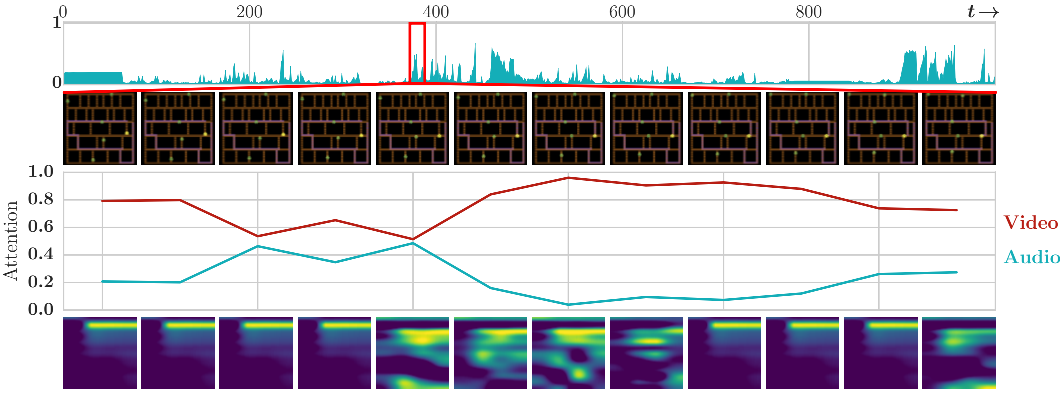


Figure 5: In Amidar, pathway vertices critical for avoiding enemies make an audible sound if not previously crossed. The agent anticipates and increases audio attention when near these vertices. Top row shows audio attention over 1000 frames, with audio/video/attention frames highlighted for zoomed-in region of interest.

reveals that the agent anticipates and increases audio attention when near these critical vertices, which are especially difficult to observe when the agent sprite is overlapping them (e.g., zoom into RGB sequences of Fig. 5).

Our crossmodal attentive agent achieves a mean score of 900 in Amidar, over 30 test runs, outperforming the other non-hierarchical methods. Note agent also beats the score of the hierarchical approach of Harb et al. [2017]. We emphasize these are not direct comparisons due to our method leveraging additional sensory inputs, but are meant to highlight the performance benefits of crossmodal learning. We also note that the state-of-the-art hierarchical approach FeUdal [Vezhnevets et al., 2017] beats our agent's score, and may achieve even higher score with crossmodal inputs.

5 Contribution

This work introduced the Crossmodal Attentive Skill Learner (CASL), integrated with the recently-introduced Asynchronous Advantage Option-Critic (A2OC) architecture [Harb et al., 2017] to enable hierarchical reinforcement learning across *multiple* sensory inputs. We provided concrete examples where CASL not only improves performance in a single task, but accelerates transfer to new tasks. We demonstrated the learned attention mechanism anticipates and identifies useful sensory features, while filtering irrelevant sensor modalities during execution. We modified the Arcade Learning Environment [Bellemare et al., 2013] to support audio queries, and evaluations of crossmodal learning were conducted in the Atari 2600 game Amidar. Finally, building on the recent work of Babaeizadeh et al. [2017], we open-source a fast hybrid CPU-GPU implementation of CASL. This investigation indicates crossmodal skill learning as a promising avenue for future works in HRL that target domains with high-dimensional, multimodal inputs.

Acknowledgement

This work was supported by Boeing Research & Technology, ONR MURI Grant N000141110688, and BRC Grant N000141712072.

References

Mohammad Babaeizadeh, Iuri Frosio, Stephen Tyree, Jason Clemons, and Jan Kautz. Reinforcement learning thorugh asynchronous advantage actor-critic on a GPU. In *ICLR*, 2017.

Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *AAAI*, pages 1726–1734, 2017.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 06 2013.

Marisa Carrasco. Visual attention: The past 25 years. *Vision research*, 51(13):1484–1525, 2011.

Jean Harb, Pierre-Luc Bacon, Martin Klissarov, and Doina Precup. When waiting is not an option: Learning options with a deliberation cost. *arXiv preprint arXiv:1709.04571*, 2017.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural comp.*, 9(8):1735–1780, 1997.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1):99–134, 1998.

George Konidaris and Andrew G Barto. Skill discovery in continuous reinforcement learning domains using skill chaining. In *Advances in neural information processing systems*, pages 1015–1023, 2009.

Tejas D Kulkarni, Karthik Narasimhan, Ardavan Saeedi, and Josh Tenenbaum. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In *Advances in Neural Information Processing Systems*, pages 3675–3683, 2016.

Yuan Chang Leong, Angela Radulescu, Reka Daniel, Vivian DeWoskin, and Yael Niv. Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron*, 93(2):451–463, 2017.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

Marlos C Machado, Marc G Bellemare, and Michael Bowling. A laplacian framework for option discovery in reinforcement learning. *arXiv preprint arXiv:1703.00956*, 2017.

Nicholas J Mackintosh. A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological review*, 82(4):276, 1975.

Rajbala Makar, Sridhar Mahadevan, and Mohammad Ghavamzadeh. Hierarchical multi-agent reinforcement learning. In *Proceedings of the fifth international conference on Autonomous agents*, pages 246–253. ACM, 2001.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pages 1928–1937, 2016.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

Yael Niv, Reka Daniel, Andra Geana, Samuel J Gershman, Yuan Chang Leong, Angela Radulescu, and Robert C Wilson. Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, 35(21):8145–8157, 2015.

John M Pearce and Geoffrey Hall. A model for pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological review*, 87(6):532, 1980.

John M Pearce and Nicholas J Mackintosh. Two theories of attention: A review and a possible integration. *Attention and associative learning: From brain to behaviour*, pages 11–39, 2010.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

Ivan Sorokin, Alexey Seleznev, Mikhail Pavlov, Aleksandr Fedorov, and Anastasiia Ignateva. Deep attention recurrent Q-network. *arXiv preprint arXiv:1512.01693*, 2015.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Richard S Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. Feudal networks for hierarchical reinforcement learning. *arXiv preprint arXiv:1703.01161*, 2017.

Oriol Vinyals, Łukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. Grammar as a foreign language. In *Advances in Neural Information Processing Systems*, pages 2773–2781, 2015.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J Smola, and Eduard H Hovy. Hierarchical attention networks for document classification. In *HLT-NAACL*, pages 1480–1489, 2016.

Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, pages 1–15, 2015.

6 Appendix

6.1 Analyzing Interactions of Attention and Memory

We provide a brief overview of LSTM networks to enable more rigorous discussion of attention-memory interactions. At timestep t, LSTM cell state C^t encodes the agent's memory given its previous stream of inputs. The cell state is updated as follows,

$$f^t = \sigma\left(W_f[x_\oplus, h^{t-1}] + b_f\right), \tag{10}$$

$$i^t = \sigma\left(W_i[x_\oplus, h^{t-1}] + b_i\right), \tag{11}$$

$$C^t = f^t \odot C^{t-1} + i^t \odot \tanh\left(W_C[x_\oplus, h^{t-1}] + b_C\right), \tag{12}$$

where f^t is the forget gate activation vector, i^t is the input gate activation vector, h^{t-1} is the previous hidden state vector, $x_\oplus$ is attended feature vector, and $\odot$ refers to the Hadamard product. Weights W_f, W_i, W_C and biases b_f, b_i, b_C are trainable parameters. The cell state update in (12) first forgets certain elements (f^t term), and then adds contributions from new inputs (i^t term). Note that a forget gate activation of 0 corresponds to complete forgetting of the previous cell state element, and that an input gate activation of 1 corresponds to complete throughput of the corresponding input element.

Our goal is to not only analyze the overall forget/input activations throughout the gameplay episode, but also to quantify the relative impact of each contributing variable (audio input, video input, hidden state, and bias term) to the overall activations. Many methods may be used for analysis of the contribution of explanatory variables in nonlinear models (i.e., (10) to (12)). We introduce a means of quantifying the correlation of each variable with respect to the corresponding activation function. In the following, we focus on the forget gate activation, but the same analysis applies to the input gate. First, expanding the definition of forget gate activation in (10), assuming use of concatenated attention (per (6)), yields,

$$f^t = \sigma\left([W_{fa}, W_{fv}, W_{fh}, b_f][\alpha_a x_a, \alpha_v x_v, h^{t-1}, I]\right), \tag{13}$$

where x_a and x_v are, respectively, the audio and video input features, and I is the identity matrix. Define $\hat{f}^t_m$ as the forget gate activation if the m-th contributing variable were removed. For example, if audio input x_a were to be removed, then,

$$\hat{f}^t_a = \sigma\left([W_{fv}, W_{fh}, b_f][\alpha_v x_v, h^{t-1}, I]\right). \tag{14}$$

Define the forget gate activation residual as $\tilde{f}^t_m = |f^t - \hat{f}^t_m|$ (i.e., the difference in output resulting from removal of the m-th contributing variable). Then, one can define a 'pseudo correlation' of the m-th contributing variable with respect to the true activation,

$$\rho(\tilde{f}^t_m) = \frac{\tilde{f}^t_m}{\sum_m \tilde{f}^t_m}. \tag{15}$$

This provides an *approximate* quantification of the relative contribution of the m-th variable (audio input, video input, hidden unit, or bias) to the overall activation of the forget and input gates. Armed with this toolset, we can analyze the interplay between attention and LSTM memory.