HHS Public Access

Author manuscript

IEEE Int Conf Data Min Workshops. Author manuscript; available in PMC 2024 July 01.

Published in final edited form as:

IEEE Int Conf Data Min Workshops. 2023 December ; 2023: 329–337. doi:10.1109/icdmw60847.2023.00048.

Deep Attention Q-Network for Personalized Treatment Recommendation

Simin Ma,

Junghwan Lee,

Nicoleta Serban,

Shihao Yang

Georgia Institute of Technology

Abstract

Tailoring treatment for severely ill patients is crucial yet challenging to achieve optimal healthcare outcomes. Recent advances in reinforcement learning offer promising personalized treatment recommendations. However, they often rely solely on a patient's current physiological state, which may not accurately represent the true health status of the patient. This limitation hampers policy learning and evaluation, undermining the effectiveness of the treatment. In this study, we propose Deep Attention Q-Network for personalized treatment recommendation, leveraging the Transformer architecture within a deep reinforcement learning framework to efficiently integrate historical observations of patients. We evaluated our proposed method on two real-world datasets: sepsis and acute hypotension patients, demonstrating its superiority over state-of-the-art methods. The source code for our model is available at https://github.com/stevenmsm/RL-ICU-DAQN.

Index Terms—

Deep learning; Reinforcement learning; Transformer; Healthcare; Precision medicine

I. Introduction

Treatment recommendation for the patients in intensive care unit (ICU) is a critical task, as it plays a vital role in the management and care of critically ill patients. Current state-of-the-art methods for providing treatment recommendations primarily involve rule-based protocols and evidence-based clinical guidelines, which are informed by randomized controlled trials (RCTs), systematic reviews, and meta-analyses. However, RCTs may not be available or definitive for many ICU conditions [1], and individual patients may respond differently to the same treatment strategy [2]. Therefore, more personalized and effective treatment plans that take into account the dynamic nature of patients' conditions and the potential presence of multiple comorbidities are needed in the ICU setting to benefit critically ill patients. Recent developments in artificial intelligence (AI) have demonstrated various successful applications in the healthcare domain, such as diagnosis [3, 4], treatment [5, 6], and resource

sma318@gatech.edu .

management [7]. Reinforcement learning (RL), in particular, is well-suited for learning optimal individual treatment interventions. RL involves sequential decision-making in an environment with evaluative feedback, with the goal of maximizing an expected reward [8]. RL shares the same goal as clinicians: making therapeutic decisions to maximize a patient's probability of a good outcome. Therefore, RL has many desirable properties and has already shown its success in providing sequential treatment suggestions in various ICU settings, such as optimal dosing of medication [9, 10, 11, 12, 13, 14, 15, 1], optimal timing of intervention [16, 17], optimal choice of medication [18], and optimal individual target lab value [19], among others. The findings from these studies all suggest that if physicians followed the RL policy, the estimated hospital mortality could be improved.

However, the patient-clinician interactions in the aforementioned studies are all modeled as Markov decision processes (MDPs), while in practice, the pathology is often complex, and the "true" underlying states of the patients are latent and can only be observed through emitted signals (observations) with some uncertainty. The challenge is that the ICU setting might not be a fully observable environment for RL agents; this could be due to a variety of factors such as noisy measurements, omission of relevant factors, and the incongruity of the frequencies and time-lags among the considered measurements [20]. To alleviate the issue of a partially observable environment, RL agents may need to remember (some or possibly all) previous observations. As a result, RL methods typically add some sort of memory component, allowing them to store or refer back to recent observations to make more informed decisions. For example, recurrent neural networks (RNNs) have been used to encode histories [21, 22, 11] or belief transitions [23, 20]. However, this creates further issues: RNNs can be subject to gradient exploding/vanishing and can be difficult to train. Recent advancements in natural language processing have led to the development of RL studies that employ the powerful Transformer architecture [24]. For instance, [25] takes a step further from [22]'s deep RL approach by replacing the RNNs with Transformer. Similarly, [26] abstracts the RL problem as a sequential modeling problem, and solve it by a variant of Transformer architecture. While these methods have demonstrated improved performance, they lack interpretability between the states and actions, which is a crucial factor in the healthcare settings. This interpretability issue arises due to the black-box nature of the models, which makes it challenging to understand how the decisions are made. Thus, it is important to develop interpretable models that can provide insights into the reasoning behind the decision-making process in healthcare settings.

In this study, we propose a novel data-driven reinforcement learning approach capable of dynamically suggesting optimal personalized clinical treatments for ICU patients. By efficiently memorizing past patient states and actions, our proposed deep reinforcement learning method can identify suitable upcoming actions and interpret the importance of relationships between actions and past observations. Compared to generic approaches for RL [25, 26], the proposed method's Transformer structure is tailor-made for healthcare data, offering improved interpretability. We demonstrate the robustness and efficiency of our proposed method on two ICU patient disease cohorts, sepsis and acute hypotension, and compare the performances against simpler and alternative benchmark approaches. The evaluation results illustrate that our proposed algorithm's learned optimal policy is able to outperform competing policies, with the help of the attention mechanism. Additionally, we

observe that the optimal policy can focus on different past observations by visualizing the attention mechanism, further providing interpretability in our proposed approach.

II. Related Work

A. Reinforcement learning for personalized treatment recommendation

Reinforcement learning (RL) has garnered considerable attention in determining optimal dosages and treatments for patients in the intensive care unit (ICU). Various studies have investigated its application to different medical scenarios, including propofol dosing for surgical patients [15, 14, 17], heparin dosing for patients with cardiovascular diseases [1, 27, 28], intravenous (IV) fluids and vasopressors dosing for sepsis patients [9, 12, 10, 29, 11, 13], and morphine dosing for patients with at least one pain intensity score [30]. Sepsis, a leading cause of hospital deaths, is a disease that is costly to treat [31]. In addition to antibiotics and source control, the use of IV fluids to correct hypovolemia and vasopressors to counteract sepsis-induced vasodilation presents significant challenges. [9] initially formulated the personalized optimal dosage for IV fluids and vasopressors as an RL problem with the goal of improving patients' outcomes, solving it with discrete state and action-based value iteration. [12] extended the model to continuous state space and discrete action and proposed solving the problem using Deep Q-learning [32]. Subsequent research studies proposed various re-formulations of the problem or explored different RL algorithms, including model-based RL algorithms [29], a combination of model-free deep RL approach and model-based kernel RL approach [11], and extension to continuous action space solved via policy-gradient algorithms [13]. To the best of our knowledge, our study is the first to consider modeling patient-clinician interactions as a partially observable environment. Inspired by the Transformer architecture [24], we represent all prior patient observations and actions as the patient's state space in our proposed deep RL approach.

B. Deep Q-Learning

Reinforcement Learning is concerned with learning control policies for agents interacting with unknown environments. Such environments are often formalized as a Markov Decision Processes (MDPs), described by a 4-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$. At each timestep t, an agent interacting with the MDP observes a state $s_t \in \mathcal{S}$, and chooses an action at $a_t \in \mathcal{A}$ which determines the reward $r_t \sim \mathcal{R}(s_t, a_t)$ (reward distribution) and next state $s_{t+1} \sim \mathcal{P}(s_t, a_t)$ (state transition probability distribution). The goal of the agent is to maximize the expected discounted cumulative reward, $\mathbb{E}\left[\sum_t \gamma^t r_t\right]$, for some discount factor $\gamma \in [0, 1)$.

Q-Learning [33] is a model-free off-policy algorithm for estimating the long-term expected return of executing an action from a given state in an MDP. These estimated returns are known as Q-values. A higher Q-value indicates an action a is judged to yield better long-term results in a state s. Q-values are learned iteratively by updating the current Q-value estimate towards the observed reward plus the max Q-value over all actions a' in the resulting state s':

$$Q(s,a) := Q(s,a) + \alpha\left(r + \gamma \max_{a' \in \mathcal{A}} Q(s',a') - Q(s,a)\right)$$

(1)

In more challenging domains, however, the state-action space of the environment is often too large to be able to learn an exact Q-value for each state-action pair. Instead of learning a tabular Q-function, Deep Q-Networks (DQN) [32] learns an approximate Q-function featuring strong generalization capabilities over similar states and actions, with the help of neural networks. DQN is trained to minimize the Mean Squared Bellman Error:

$$L(\theta) = \mathbb{E}_{s,a,r,s'}\left[\left(r + \gamma \max_{a' \in \mathcal{A}} Q(s', a'; \theta') - Q(s, a; \theta)\right)^2\right]$$

(2)

where transition tuples of states, actions, rewards, and future states (s, a, r, s') are sampled uniformly from a replay buffer, D, of past experiences while training. The target $r + \gamma \max_{a'} Q(s', a'; \theta')$ invokes DQN's target network (parameterized by θ'), which lags behind the main network (parameterized by θ) to produce more stable updates.

III. Methods

A. Problem formulation

When an environment does not emit its full state to the agent, the problem can be modeled as a Partially Observable Markov Decision Process (POMDP), described by 6-tuple ($\mathcal{S}$, $\mathcal{A}$, $\mathcal{P}$, $\mathcal{R}$, Ω, $\mathcal{O}$). The two additional sets, Ω, $\mathcal{O}$, represents the observations set and state-observation distributions, respectively. In particular, at each time step t, after the agent (in state $s_t \in S$) interacts with the environment by taking action $a_t \in \mathcal{A}$ and obtain the reward $r_t \sim \mathcal{R}(s_t, a_t)$, the agent moves into the next state $s_{t+1} \sim \mathcal{P}(s_t, a_t)$, but no longer observes true system state and instead receives an observation o_{t+1} Ω, generated from the underlying system state s_{t+1} according to the probability distribution $o_{t+1} \sim \mathcal{O}(s_{t+1})$. In ICU setting, s_t can be interpreted as the "true" underlying patient's health status, while o_t is the current observable measurements (vital signs, demographics, etc.). Because agents in POMDPs do not have access to the environment's full state information, they must rely on the observations $o_t \in \Omega$. In this case, DQN may not learn a good policy by simply estimating the Q-values from the current patient's observation, o_t, since it may not be representative enough for the "true" underlying patient's health status, s_t. Instead, one often needs to consider some form of all patient's historical observations and actions, for instance $\{(o_0, a_0), (o_1, a_1), \ldots, (o_{t-1}, a_{t-1})\}$, to approximate the true current state, s_t. Because the history grows indefinitely as the agent proceeds in a trajectory, efficient ways of encoding the history is needed, for example using an agent's belief [23], using recurrent neural network and its variants [22, 34], etc. Here, we incorporate the recently developed Transformer's attention mechanisms [24] into the Deep Q-Networks, which is able to incorporate histories into the Q-function and reflect the relative importance between the upcoming actions and the histories.

B. Proposed Method: Deep Attention Q-Network

The transformer architecture [24], originally introduced for sequence to sequence translation in natural language applications, utilizes attention mechanism [35], which is able to "focus" on different portions of the input when translating to outputs. With its strong interpretability and computational efficiency, the transformer architecture, originally formed as an encoder-decoder structure, is now broadly used in various applications, using either the encoder [36], the decoder [37], or reconstructed architectures [38]. Transformer and its attention structures seem like a natural fit to represent the histories in POMDPs, as it encapsulates several inductive biases nicely. Therefore, we propose Deep Attention Q-Network (DAQN), by "inserting" the attention mechanisms to the traditional Deep Q-Networks to learn the approximate Q-function. The high-level overview of the proposed DAQN is shown in Figure 1, and the detailed workflow is presented in the caption.

Similar to the original Transformer's decoder structure [24], each attention-like blocks in DAQN features two main sub-modules: encoder-decoder attention and position-wise feedforward network. First, in the encoder-decoder attention, the fixed start token (serving as a dummy variable) will be projected to queries Q, and positional encoded and embedded observation histories will be projected to keys K and values V, through the learned weight matrices W^Q, W^K, W^V, respectively. Then, via "Scaled Dot-Product Attention" [24], a softmax function is applied on the dot products between queries and keys, which are the attention weights and are used to obtain a weighted sum on the values. On the high-level, the attention weights can be interperated as the "importance" weight of each observation history relative to the the Q-values of state-action pairs. Then the output from the encoder-decoder attention will pass through a fully connected feed-forward network, which is applied to each position separately and identically. After each submodule, that submodule's input and output are combined and followed by layer normalization [39].

We additionally incorporate Dueling Q-network architecture [40] and Double-Deep Q-network architecture [41] into our proposed architecture, and also use Prioritized Experience Replay [42] to accelerate learning, similar to prior RL sepsis studies [10, 12, 11]. More implementation details are presented in the Appendix A1.

IV. Experimental Setup

A. Data

We obtain the data from the "Medical Information Mart for Intensive Care database" (MIMIC-III) version 1.4 [43], which is a publicly available database that consists of de-identified health-related data associated with patients stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The database includes information such as demographics, vital sign measurements made at the bedside, laboratory test results, procedures, medications, etc. In this study, we focus on two ICU patient cohorts: sepsis patient and acutely hypotensive patient.

Sepsis Patients.—We extract sepsis patients in MIMIC-III database satisfying the following criteria: (1) patients who were older than 18 years old; (2) patients with the

length of stay over 24 hours (to ensure sufficient data for analysis); (3) patients who were diagnosed with sepsis according to the Sepsis-3 criteria [44]. Also, if a patient had multiple admissions with sepsis, only the first admission was analyzed. After excluding patients with relevant variables missing, we have a total of 6,164 patients. For each patient, we extract the relevant physiological parameters including demographics, lab values, vital signs, and intake/output events. Then, each patient's record is aggregated into windows of 4 hours, with the mean or sum being recorded when several data points were present in one window (same as prior studies [9, 12]). This yielded a feature vector for each patient at each timestep. For each patient $i \in \{1, \ldots, N\}$, at each timestep t, the current feature vector and the previous feature vectors (and previous actions) form the observation $o_t^{(i)}$ in the underlying POMDP.

We focus on the action space defined by two treatment interventions: intravenous fluid (IV fluid) and vasopressor, given their uncertainty in clinical literature [45] and the crucial impact on a patient's eventual outcome. We define a 5 × 5 action space for the medical interventions covering the space of intravenous (IV) fluid and maximum dose of vasopressor in each 4-hour period (4 per-drug quartiles, and a special case of no drug), similar to prior studies' action space discretization [12, 9]. Thus, there are 25 actions in total, where each action represents the intervention as a tuple of Input 4H and Max Vaso at each 4-hour period.

To train an RL agent for sepsis management, we adopted a similar reward function as in [12], which uses the Sequential Organ Failure Assessment scores (SOFA) [46] and the lactate level of the patients. On the high level, higher SOFA scores indicate greater organ dysfunction and is predictive of ICU mortality, while lactate levels measure cell-hypoxia which is higher in septic patients. The rewards function penalizes high SOFA scores and lactate levels at time t, as well as positive changes in these scores. Conversely, positive rewards are given for decreased SOFA scores and lactate levels, indicating improved patient states. Further details of the dataset and reward function are available in Appendix B.

Acutely Hypotension Patients—Following [47], we extract the acutely hypotensive patients in MIMIC-III database that fulfill the following criteria: (1) patients who were older than 18 years old; (2) patients with the length of stay over 24-hours (and select the initial ICU admission only); (3) patients with seven or more mean arterial pressure (MAP) values of 65 mmHg or less, which indicated probable acute hypotension. For each patient, we extract the relevant physiological parameters (Table V), and limit to only using information captured during the initial 48 hours after admission. We have a final cohort consisting of 3910 distinct ICU admissions.

We focused on the action space defined by two treatment interventions: fluid boluses and vasopressor, defining a 4 × 4 action space (3 per-drug quartiles, and a special case of no drug).

We also adopted a similar reward function as [47] for training an RL agent for the management of acute hypotension. The reward at time step t of a patient, is dependent on the Mean Arterial Pressure (MAP) and urine output at time t. Further details of the dataset and reward function are available in Appendix B.

B. Off-policy Evaluation

In this study, we focus on off-policy learning, which means that our RL agent aims to learn an optimal policy (i.e. optimal medication dosages) through data that are already generated by following the clinician policy (see section IV–A). For each cohort, we learn the optimal DAQN policy, and conduct evaluation comparisons against other benchmark policies. This section describe the off-policy evaluation procedure adapted in this study in detail.

Proper quantitative evaluation of learned policy is crucial before deployment, especially in healthcare. Off-policy evaluation (OPE) in the reinforcement learning context is typically used as the performance metric for comparison [48, 49]. Here, we employ weighted doubly-robust method (WDR) to quantify the performance of RL policies [48]. We include the proposed DAQN policy, DRQN (Deep Recurrent Q-Network) policy [22], DQN (Deep Q-Network) policy [12], clinician policy, and random policy for comparison. The DRQN and DQN policy use vanilla LSTM network (long-short-term-memory) [50] and feedforward neural networks [32] for learning the approximate Q-function (see Equation 1), respectively. The clinician policy comprises actions from historical data which clinicians take. For random policy, actions are uniformly sampled from the 0 to the upper bound range. To account for randomness, we performed 50 experiments with a unique train/test set split in each experiment. More details on the benchmark policies and training procedures are presented in the Appendix A2.

V. Results

In this section, we show how the proposed DAQN dynamically suggest optimal personalized healthcare treatments in ICU, as well as the comparisons against other benchmark policies.

A. Sepsis Patients

The results are presented in Table I and the box plot in Figure 4a. The quantitative results demonstrate that our proposed DAQN policy is able to outperform benchmark DRQN policy in both mean and standard deviation, which also incorporates historical patients' observations for state representation, while outperforming traditional DQN policy, clinician policy, and random policy. The value of the clinician's policy is estimated with high confidence, as expected. DAQN and DQN-based policy all bound on the estimated value relatively tight and larger than the value of the clinician's policy (1st to 3rd quartile box small and above the mean for the clinician's policy), in contrast to the random policy's large box extending to well below the clinician's policies. The random policy's values are distributed evenly around zero, which is expected as the reward distributions is also approximately distributed around zero. DRQN policy is also able to outperform clinician's policy and baseline random policy, and exhibit larger mean than DQN policy, but exhibit the largest variance.

In addition, we examine the interpretability in our proposed DAQN policy, by focusing on the attention weights over the input historical patient observations. We observe that the attention weights, produed in the "Encoder-Decoder Attention" block (in Figure 1), is positively associated with patient's SOFA score, change in SOFA score, and lactate level,

which are all important indicator in sepsis patients, with high association with mortality and morbidity [51]. Numerous studies [52, 46, 51, 53] show that the SOFA score is highly sensitive and predictive in the diagnosis of sepsis. For instance, an initial and highest scores of more than 11 or mean scores of more than 5 corresponded to mortality of more than 80%, while change in SOFA score (delta SOFA score) also significantly associated with mortality and patient discharge [53]. Indeed, with the designed reward function that penalizes high SOFA scores, positive changes in SOFA scores, and high lactate levels, the attention weights learn to "focus" on the prior observations and actions that exhibit high SOFA score, change in SOFA score, or lactate level, respectively. Table II shows the correlation coefficient between the average attention weights (across attention heads) in each layer and SOFA score, change in SOFA score, lactate level, respectively. Then, we select three example patients and visualize the average attention weights in each layer with SOFA score, delta SOFA score, and lactate level (the elements that has the highest correlation coefficient with each layer's average attention weight, see Table II), in Figure 2, 5, 6. For example, the attention weights of example patient 1 in 2 exhibit strong trend matching behavior with SOFA score, Delta SOFA score, and lactate level, and learn to "focus" (high attention weights) on past observations that indicates worse patient's health status. This further confirms DAQN's ability to focus on more "important" and severe historical observations when learning the optimal policy. More details on presented in Appendix C.

B. Acute Hypotension Patients

Managing hypotensive patients in the ICU is a challenging task that lacks standardization due to the high heterogeneity of patients, which often leads to high morbidity and mortality rates [54]. In light of the limited evidence to guide treatment guidelines, RL offers a promising approach to improve strategies for managing these patients [55]. To evaluate the efficacy of RL policies for acutely hypotensive patients, we conducted experiments on the MIMIC-III dataset, and the results are presented in Table I and the box plot in Figure 4b. We used the same hyperparameters for DAQN and the benchmark policies as in the previous section. Our results show that, similar to the sepsis patient cohort, the DAQN policy outperforms the DRQN and DQN policies, with the 3rd quartile box located above the mean of the DRQN and DQN policies. This performance improvement is attributed to the Transformer architecture and attention mechanism employed by the DAQN, which enables it to focus on and efficiently memorize past patient observations and actions as current patient health status representation more robustly than the DRQN, which can suffer from vanishing/exploding gradient problems due to RNNs. However, the DRQN policy shows a higher mean than the DQN policy, which underscores the importance of modeling the ICU setting as a POMDP problem, treating the current patient vitals and static information as observations, and not as the "true" patient underlying health status. The estimated values of the DAQN, DRQN, and DQN policies are relatively tight, and larger than those of the clinician policy and randomized policy. The expected reward of the clinician policy is estimated with high confidence, while the randomized policy's reward distribution is similar to that of the entire cohort. Overall, our results demonstrate the potential of RL policies, especially the DAQN, to improve the management of acutely hypotensive patients in the ICU.

VI. Discussion

Various pioneering studies have explored applying reinforcement learning algorithms to the search for optimal clinical treatment, such as for sepsis patients [10, 11, 12, 13] and for acutely hypotentive patients [47], demonstrating the potential of using RL to improve ICU patient outcomes. However, these studies are limited to a coarse-grained state space that only depends on the current patient's observations for state space definitions. Although this representation of patient's state space is intuitive and straightforward for determining current actions, patients' prior observations and prior actions are also important factors of treatment intervention decisions. Thus, simply modeling the patient-clinician interactions as MDPs is prone to mis-specification of the "true" underlying patient's states, and potentially take sub-optimal actions, leading to sub-optimal results (rewards). In order to make personalized treatment design more clinically meaningful, we proposed to "enrich" the state space, by modeling the ICU setting as a POMDP and letting the RL agent efficiently memorize patients' current and prior observations and actions. By incorporating attention mechanism, our proposed RL algorithm is able to outperform baseline benchmark policies, and provide interpretability, similar to clinician's diagnosis process.

By extending state space with prior observations, we learn the optimal policy using our proposed RL algorithm, which can provide more meaningful and higher-resolution decision support to patients. For quantitative policy evaluation, we compare our optimal policy with other policies learned by alternative benchmark RL algorithms, as well as baseline policies, through off-policy evaluation with WDR estimator [48]. The results from off-policy evaluations shows that the proposed the RL policy is able to perform competitively against alternative benchmark policy that uses simpler memorization techniques (LSTM), and outperform other RL policies that do not incorporate prior observations (and actions). The proposed RL policy can also provide better expected reward compared to clinician policy and random policy baseline. Moreover, similar to clinician decision process and thinking, the proposed RL policy will focus on the prior observations (and actions) that indicate worse patient health in combination of current observations, when making treatment dosage decisions (see Figure 2, 5, 6). With clinician guided reward function and attention mechanism, our proposed RL policy is able to shift focus among previous patient's observations for subsequent treatment decisions. This improvement also makes reinforcement learning-based treatment search closer to real-world deployment.

Potential avenues for future work include a more thorough discussion with clinicians to potentially make the observation and action histories even more representative, and architectural improvements that could provide more detailed interpretation for patient-intervention relation.

As discussed in prior studies [12, 56], evaluating RL policies with off-policy evaluation is challenging, as all the available data are offline sampled (i.e. following clinician's policy). Since the evaluation policies are deterministic, the off-policy evaluation that uses importance weight will only be non-zero if the evaluation policy recommends the same treatment as the clinician's policy. This will results in high-variance estimates of the quality of the evaluation

policy. Future examination from both policy learning and policy evaluation aspects shall be considered.

Another limitation of this study is that the dosages are discretized into per-drug quartiles, with each action representing dosages in a particular range. However, each quartile includes a wide range of dosages , which can be complicated in practice for clinicians to make decisions on the exact dosages of IV fluids and vasopressor to use. [13] proposed a RL-based solution via Deep Deterministic Policy Gradient [57], which is a model-free off-policy algorithm for learning continuous actions. Therefore, it is also important to further investigate policy-gradient algorithms that efficiently memorizes patient's prior observations and actions continuous actions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

This work was supported by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health under Award Number P30DK111024. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1]. Nemati Shamim, Ghassemi Mohammad M, and Clifford Gari D. Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. In 2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC), pages 2978–2981. IEEE, 2016.

[2]. Laffey John Gand Kavanagh Brian P. Negative trials in critical care: why most research is probably wrong. The Lancet Respiratory Medicine, 6(9):659–660, 2018. [PubMed: 30061048]

[3]. Burke Anne E, Thaler Katrina M, Geva Mika, and Adiri Yonatan. Feasibility and acceptability of home use of a smartphone-based urine testing application among women in prenatal care. American Journal of Obstetrics & Gynecology, 221(5):527–528, 2019. [PubMed: 31300161]

[4]. Laserson Jonathan, Lantsman Christine Dan, Cohen-Sfady Michal, Tamir Itamar, Goz Eli, Brestel Chen, Bar Shir, Atar Maya, and Elnekave Eldad. Textray: Mining clinical reports to gain a broad understanding of chest x-rays. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 553–561. Springer, 2018.

[5]. Choi Edward, Bahadori Mohammad Taha, Schuetz Andy, Stewart Walter F, and Sun Jimeng. Doctor ai: Predicting clinical events via recurrent neural networks. In Machine learning for healthcare conference, pages 301–318. PMLR, 2016.

[6]. Fan Jiawei, Wang Jiazhou, Chen Zhi, Hu Chaosu, Zhang Zhen, and Hu Weigang. Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique. Medical physics, 46(1):370–381, 2019. [PubMed: 30383300]

[7]. Watanabe Alyssa T, Lim Vivian, Vu Hoanh X, Chim Richard, Weise Eric, Liu Jenna, Bradley William G, and Comstock Christopher E. Improved cancer detection using artificial intelligence: a retrospective evaluation of missed cancers on mammography. Journal of digital imaging, 32(4):625–637, 2019. [PubMed: 31011956]

[8]. Montague P Read. Reinforcement learning: an introduction, by sutton, rs and barto, ag. Trends in cognitive sciences, 3(9):360, 1999.

[9]. Komorowski Matthieu, Celi Leo A, Badawi Omar, Gordon Anthony C, and Faisal A Aldo. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nature medicine, 24(11):1716–1720, 2018.

[10]. Raghu Aniruddh, Komorowski Matthieu, Celi Leo Anthony, Szolovits Peter, and Ghassemi Marzyeh. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In Machine Learning for Healthcare Conference, pages 147–163. PMLR, 2017.

[11]. Peng Xuefeng, Ding Yi, Wihl David, Gottesman Omer, Komorowski Matthieu, Li-wei H Lehman, Ross Andrew, Faisal Aldo, and Doshi-Velez Finale. Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning. In AMIA Annual Symposium Proceedings, volume 2018, page 887. American Medical Informatics Association, 2018. [PubMed: 30815131]

[12]. Raghu Aniruddh, Komorowski Matthieu, Ahmed Imran, Celi Leo, Szolovits Peter, and Ghassemi Marzyeh. Deep reinforcement learning for sepsis treatment. arXiv preprint arXiv:1711.09602, 2017.

[13]. Huang Yong, Cao Rui, and Rahmani Amir. Reinforcement learning for sepsis treatment: A continuous action space solution. In Machine Learning for Healthcare Conference, pages 1–17. PMLR, 2022.

[14]. Padmanabhan Regina, Meskin Nader, and Haddad Wassim M. Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning. Biomedical Signal Processing and Control, 22:54–64, 2015.

[15]. Borera Eddy C, Moore Brett L, Doufas Anthony G, and Pyeatt Larry D. An adaptive neural network filter for improved patient state estimation in closed-loop anesthesia control. In 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence, pages 41–46. IEEE, 2011.

[16]. Prasad Niranjani, Cheng Li-Fang, Chivers Corey, Draugelis Michael, and Engelhardt Barbara E. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. arXiv preprint arXiv:1704.06300, 2017.

[17]. Yu Chao, Liu Jiming, and Zhao Hongyi. Inverse reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. BMC medical informatics and decision making, 19(2):111–120, 2019. [PubMed: 31196073]

[18]. Wang Lu, Zhang Wei, He Xiaofeng, and Zha Hongyuan. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. In Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pages 2447–2456, 2018.

[19]. Weng Wei-Hung, Gao Mingwu, He Ze, Yan Susu, and Szolovits Peter. Representation and reinforcement learning for personalized glycemic control in septic patients. arXiv preprint arXiv:1712.00654, 2017.

[20]. Li Luchen, Komorowski Matthieu, and Faisal Aldo A. Optimizing sequential medical treatments with auto-encoding heuristic search in pomdps. arXiv preprint arXiv:1905.07465, 2019.

[21]. Wierstra Daan, Foerster Alexander, Peters Jan, and Schmidhuber Juergen. Solving deep memory pomdps with recurrent policy gradients. In International conference on artificial neural networks, pages 697–706. Springer, 2007.

[22]. Hausknecht Matthew and Stone Peter. Deep recurrent q-learning for partially observable mdps. In 2015 aaai fall symposium series, 2015.

[23]. Igl Maximilian, Zintgraf Luisa, Le Tuan Anh, Wood Frank, and Whiteson Shimon. Deep variational reinforcement learning for pomdps. In International Conference on Machine Learning, pages 2117–2126. PMLR, 2018.

[24]. Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N, Kaiser Łukasz, and Polosukhin Illia. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[25]. Esslinger Kevin, Platt Robert, and Amato Christopher. Deep transformer q-networks for partially observable reinforcement learning. arXiv preprint arXiv:2206.01078, 2022.

[26]. Chen Lili, Lu Kevin, Rajeswaran Aravind, Lee Kimin, Grover Aditya, Laskin Misha, Abbeel Pieter, Srinivas Aravind, and Mordatch Igor. Decision transformer: Reinforcement learning via sequence modeling. Advances in neural information processing systems, 34:15084–15097, 2021.

[27]. Ghassemi Mohammad M, Alhanai Tuka, Westover M Brandon, Mark Roger G, and Nemati Shamim. Personalized medication dosing using volatile data streams. In Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[28]. Lin Rongmei, Stanley Matthew D, Ghassemi Mohammad M, and Nemati Shamim. A deep deterministic policy gradient approach to medication dosing and surveillance in the icu. In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 4927–4931. IEEE, 2018.

[29]. Raghu Aniruddh, Komorowski Matthieu, and Singh Sumeetpal. Model-based reinforcement learning for sepsis treatment. arXiv preprint arXiv:1811.09602, 2018.

[30]. Lopez-Martinez Daniel, Eschenfeldt Patrick, Ostvar Sassan, Ingram Myles, Hur Chin, and Picard Rosalind. Deep reinforcement learning for optimal critical care pain management with morphine using dueling double-deep q networks. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 3960–3963. IEEE, 2019.

[31]. Cohen Jonathan, Vincent Jean-Louis, Adhikari Neill KJ, Machado Flavia R, Angus Derek C, Calandra Thierry, Jaton Katia, Giulieri Stefano, Delaloye Julie, Opal Steven, et al. Sepsis: a roadmap for future research. The Lancet infectious diseases, 15(5):581–614, 2015. [PubMed: 25932591]

[32]. Mnih Volodymyr, Kavukcuoglu Koray, Silver David, Rusu Andrei A, Veness Joel, Bellemare Marc G, Graves Alex, Riedmiller Martin, Fidjeland Andreas K, Ostrovski Georg, et al. Human-level control through deep reinforcement learning. nature, 518(7540):529–533, 2015. [PubMed: 25719670]

[33]. Watkins Christopher JCH and Dayan Peter. Q-learning. Machine learning, 8(3):279–292, 1992.

[34]. Zeng Junjie, Ju Rusheng, Qin Long, Hu Yue, Yin Quanjun, and Hu Cong. Navigation in unknown dynamic environments based on deep reinforcement learning. Sensors, 19(18):3837, 2019. [PubMed: 31491927]

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

[36]. Devlin Jacob Chang Ming-Wei Kenton Lee Toutanova Kristina. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, pages 4171–4186, 2019.

[37]. Yang Zhilin, Dai Zihang, Yang Yiming, Carbonell Jaime, Salakhutdinov Russ R, and Le Quoc V. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 32, 2019.

[38]. Wu Haixu, Xu Jiehui, Wang Jianmin, and Long Mingsheng. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. Advances in Neural Information Processing Systems, 34:22419–22430, 2021.

[39]. Ba Jimmy Lei, Kiros Jamie Ryan, and Hinton Geoffrey E. Layer normalization. arXiv preprint arXiv:1607.06450, 2016.

[40]. Wang Ziyu, Schaul Tom, Hessel Matteo, Hasselt Hado, Lanctot Marc, and Freitas Nando. Dueling network architectures for deep reinforcement learning. In International conference on machine learning, pages 1995–2003. PMLR, 2016.

[41]. Van Hasselt Hado, Guez Arthur, and Silver David. Deep reinforcement learning with double q-learning. In Proceedings of the AAAI conference on artificial intelligence, volume 30, 2016.

[42]. Schaul Tom, Quan John, Antonoglou Ioannis, and Silver David. Prioritized experience replay. arXiv preprint arXiv:1511.05952, 2015.

[43]. Johnson Alistair EW, Pollard Tom J, Shen Lu, Lehman Li-wei H, Feng Mengling, Ghassemi Mohammad, Moody Benjamin, Szolovits Peter, Celi Leo Anthony, and Mark Roger G. Mimic-iii, a freely accessible critical care database. Scientific data, 3(1):1–9, 2016.

[44]. Singer Mervyn, Deutschman Clifford S, Seymour Christopher Warren, Shankar-Hari Manu, Annane Djillali, , Bauer Michael, Bellomo Rinaldo, Bernard Gordon R, Chiche Jean-Daniel, Coopersmith Craig M, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). Jama, 315(8):801–810, 2016. [PubMed: 26903338]

[45]. Napolitano Lena M. Sepsis 2018: definitions and guideline changes. Surgical infections, 19(2):117–125, 2018. [PubMed: 29447109]

[46]. Vincent J-L, Moreno Rui, Takala Jukka, Willatts Sheila, De Mendonça Arnaldo, Bruining Hajo, Reinhart CK, Suter PeterM, and Thijs Lambertius G. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure, 1996.

[47]. Gottesman Omer, Futoma Joseph, Liu Yao, Parbhoo Sonali, Celi Leo, Brunskill Emma, and Doshi-Velez Finale. Interpretable off-policy evaluation in reinforcement learning by highlighting influential transitions. In International Conference on Machine Learning, pages 3658–3667. PMLR, 2020.

[48]. Thomas Philip and Brunskill Emma. Data-efficient off-policy policy evaluation for reinforcement learning. In International Conference on Machine Learning, pages 2139–2148. PMLR, 2016.

[49]. Jiang Nan and Li Lihong. Doubly robust off-policy value evaluation for reinforcement learning. In International Conference on Machine Learning, pages 652–661. PMLR, 2016.

[50]. Hochreiter Sepp and Schmidhuber Jürgen. Long short-term memory. Neural computation, 9(8):1735–1780, 1997. [PubMed: 9377276]

[51]. Lambden Simon, Laterre Pierre Francois, Levy Mitchell M, and Francois Bruno. The sofa score—development, utility and challenges of accurate assessment in clinical trials. Critical Care, 23(1):1–9, 2019. [PubMed: 30606235]

[52]. Liu Zhiqiang, Meng Zibo, Li Yongfeng, Zhao Jingyuan, Wu Shihong, Gou Shanmiao, and Wu Heshui. Prognostic accuracy of the serum lactate level, the sofa score and the qsofa score for mortality among adults with sepsis. Scandinavian journal of trauma, resuscitation and emergency medicine, 27:1–10, 2019. [PubMed: 30616604]

[53]. Lopes Ferreira Flavio, Peres Bota Daliana, Bross Annette, Mélot Christian, and Vincent Jean-Louis. Serial evaluation of the sofa score to predict outcome in critically ill patients. Jama, 286(14):1754–1758, 2001. [PubMed: 11594901]

[54]. Jones Alan E, Yiannibas Vasilios, Johnson Charles, and Kline Jeffrey A. Emergency department hypotension predicts sudden unexpected in-hospital mortality: a prospective cohort study. Chest, 130(4):941–946, 2006. [PubMed: 17035422]

[55]. de Grooth Harm-Jan, Postema Jonne, Loer Stephan A, Parienti Jean-Jacques, Oudemans-van Straaten Heleen M, and Girbes Armand R. Unexplained mortality differences between septic shock trials: a systematic analysis of population characteristics and control-group mortality rates. Intensive care medicine, 44:311–322, 2018. [PubMed: 29546535]

[56]. Gottesman Omer, Johansson Fredrik, Meier Joshua, Dent Jack, Lee Donghun, Srinivasan Srivatsan, Zhang Linying, Ding Yi, Wihl David, Peng Xuefeng, et al. Evaluating reinforcement learning algorithms in observational health settings. arXiv preprint arXiv:1805.12298, 2018.

[57]. Lillicrap Timothy P, Hunt Jonathan J, Pritzel Alexander, Heess Nicolas, Erez Tom, Tassa Yuval, Silver David, and Wierstra Daan. Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971, 2015.

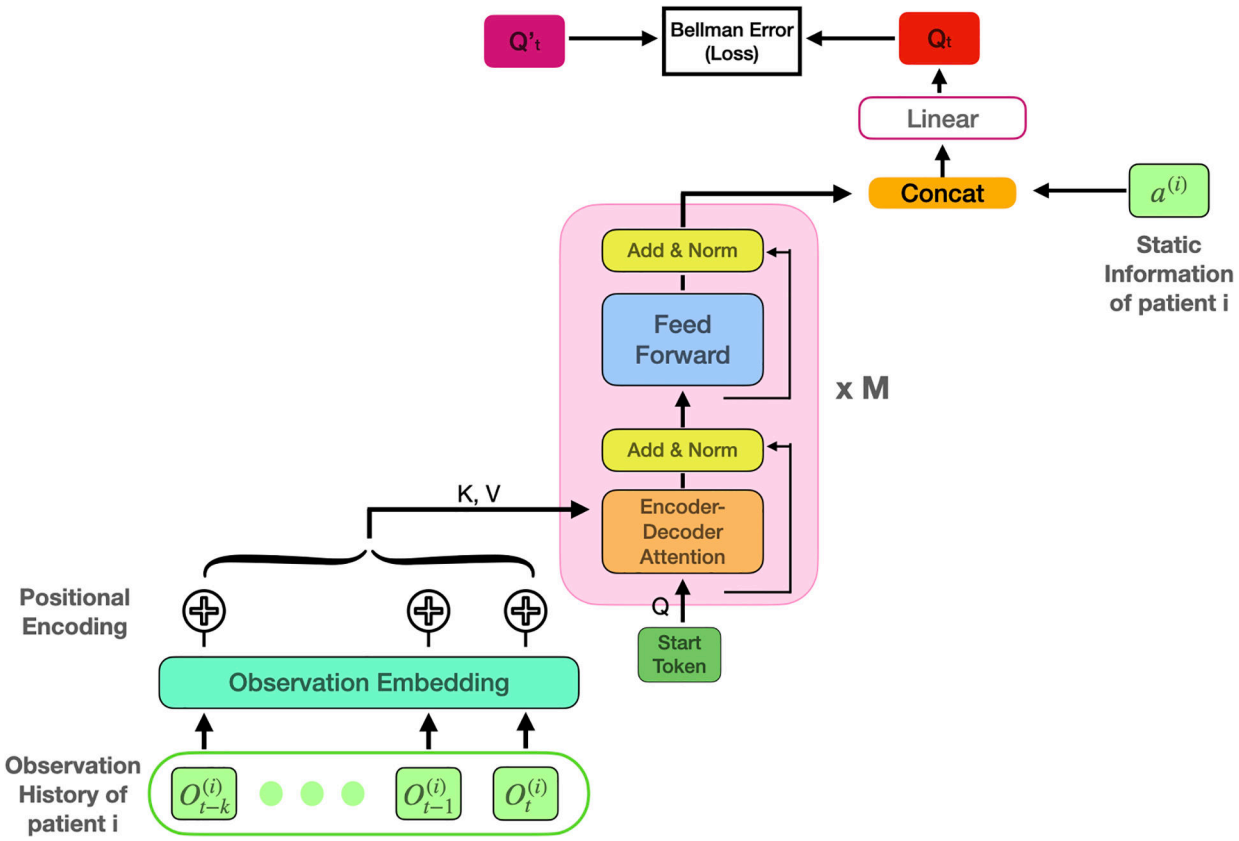
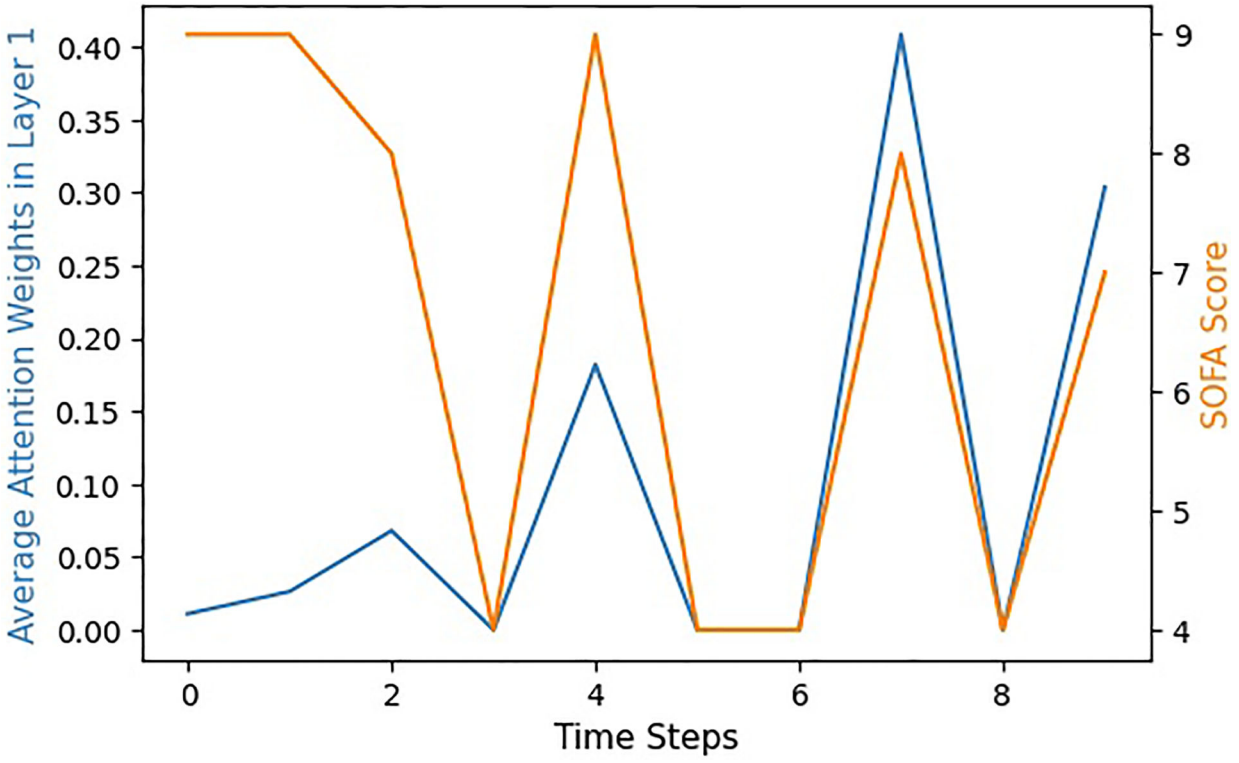
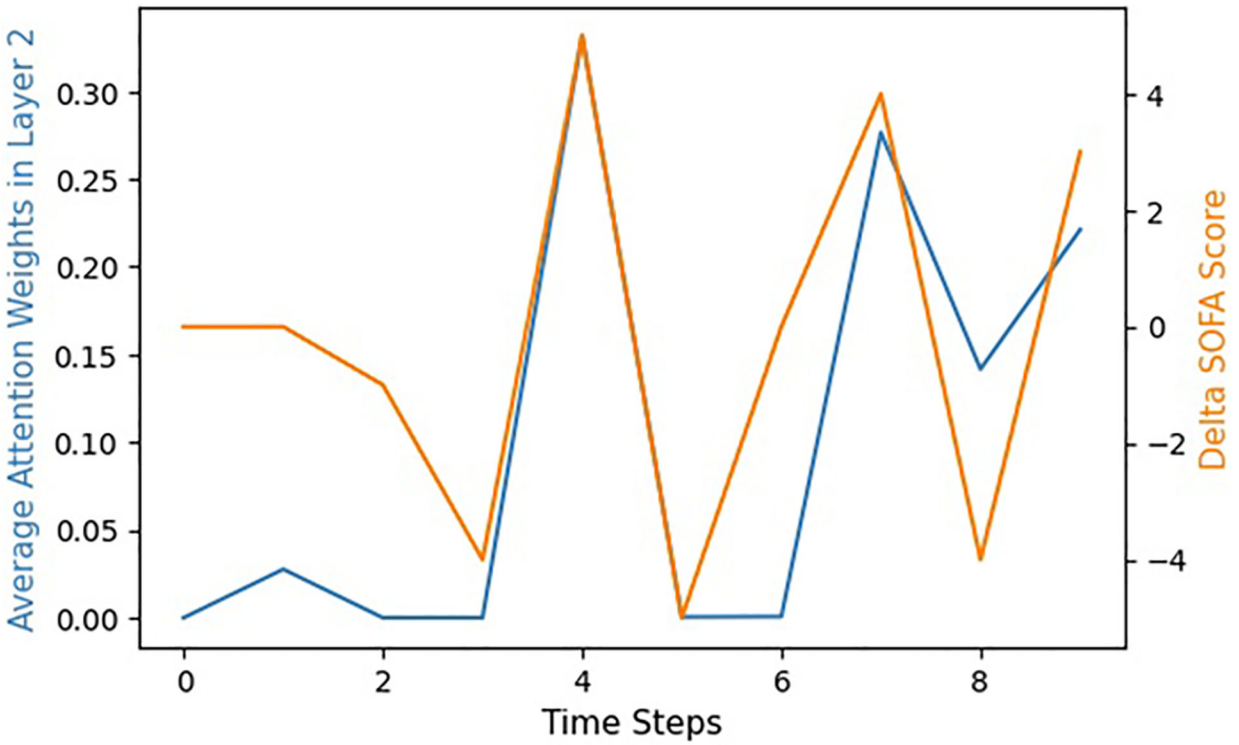


Fig. 1:

Deep Attention Q-Network Architecture. Workflow: (1) input patient i previous k observations, $h_{t-k:t} = \{o^{(i)}_{t-k}, \ldots, o^{(i)}_{t-1}, o^{(i)}_{t}\}$; (2) linearly projects each observation into the model's dimension, and add learned positional encoding to each embedded observation based on its position in the observation history; (3) pass the embedded and positional encoded observations through M attention-like structure blocks, while attempting to decode to a fixed start token; (4) final output from M attention-like structure blocks is concatenated with patient i's static information vector $a^{(i)}$ and fed into a final linear layer to project into the action space dimension, Q_t, where each entry indicates the discounted cumulative reward given the current observation history and action of interest; (5) mean squared Bellman loss is computed between Q_t and Q'_t to learn all network parameters (see Equation 3).





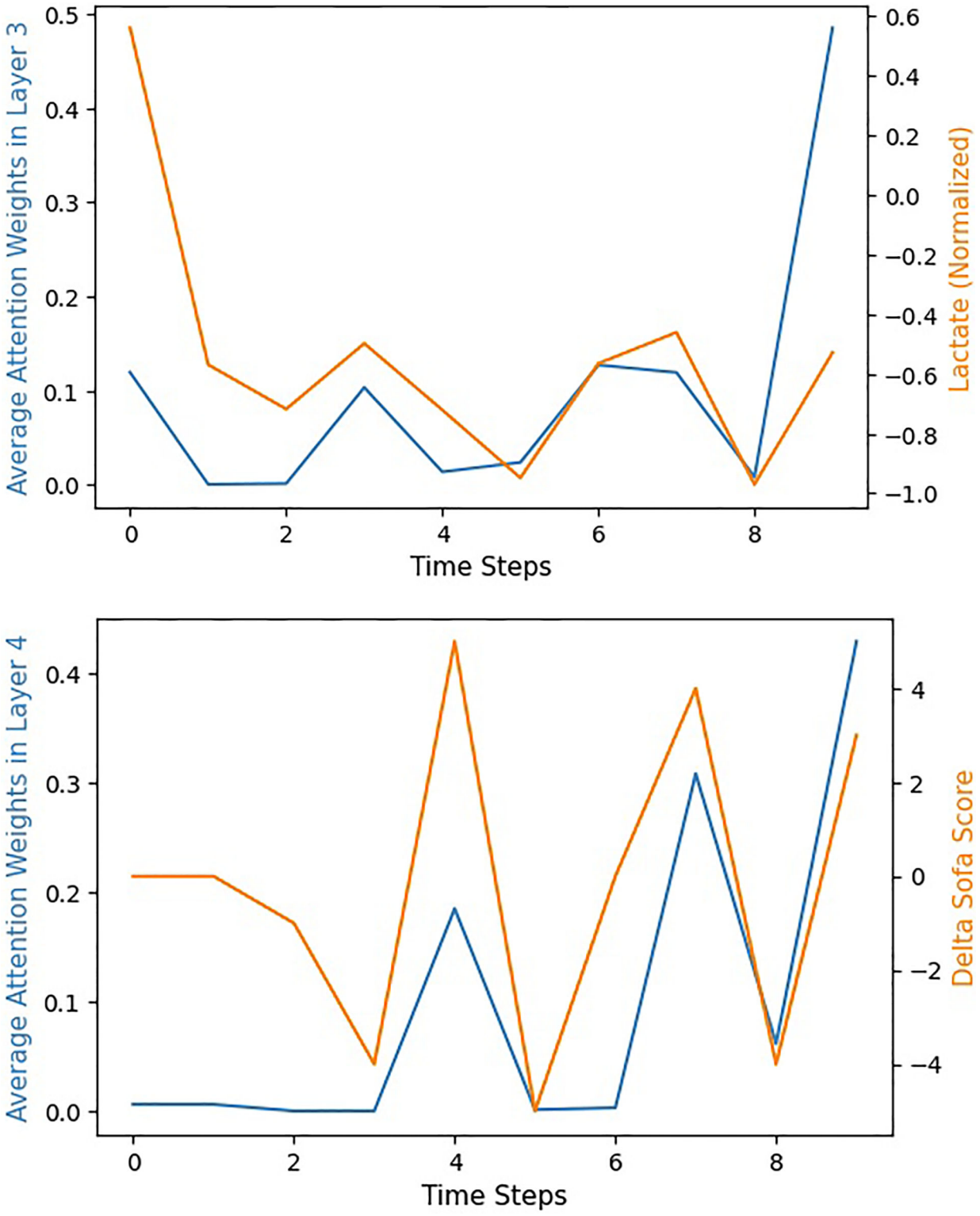


Fig. 2:

Example sepsis patient 1 average attention weights in each layer against SOFA score, delta SOFA score, and lactate level, respectively.

TABLE I:

Mean and standard deviation of off-policy evaluation estimates via WDR estimator [48], across evaluating policies. The best performing policy is boldfaced.

Evaluating Policies	Sepsis	Acute Hypotension
DAQN	**0.34893±0.0632**	**−0.00562±0.0033**
DRQN	0.23683±0.0708	−0.01014±0.0052
DQN	0.16710±0.0646	−0.01422±0.0051
Clinician	0.06772±0.0149	−0.02363±0.0014
Random	−0.03517±0.0617	−0.03857±0.0073

TABLE II:

Correlation coefficient between averaged attention weights (across each head) in each layer and SOFA score, Delta SOFA score, and lactate level.

Attention Layer	SOFA	Delta SOFA	Lactate
Layer 1	0.626	0.368	0.209
Layer 2	0.579	0.723	0.340
Layer 3	0.390	0.120	0.575
Layer 4	0.340	0.383	0.251