Verbal Explanations for Deep Reinforcement Learning Neural Networks with Attention on Extracted Features

Xinzhi Wang[1], Shengcheng Yuan[2], Hui Zhang[3], Michael Lewis[4], Katia Sycara[5]

***Abstract*— In recent years, there has been increasing interest in transparency in Deep Neural Networks. Most of the works on transparency have been done for image classification. In this paper, we report on work in transparency in Deep Reinforcement Learning Networks (DRLNs). Such networks have been extremely successful in learning action control in Atari games. In this paper, we focus on generating verbal (natural language) descriptions and explanations of deep reinforcement learning policies. Successful generation of verbal explanations would allow better understanding by people (e.g., users, debuggers) of the inner workings of DRLNs which could ultimately increase trust in these systems. We present a generation model which consists of three parts: an encoder on feature extraction, an attention structure on selecting features from the output of the encoder, and a decoder on generating the explanation in natural language. Four variants of the attention structure - full attention, global attention, adaptive attention and object attention - are designed and compared. The adaptive attention structure performs the best among all the variants, even though the object attention structure is given additional information on object locations. Additionally, our experiment results showed that the proposed encoder outperforms two baseline encoders (Resnet and VGG) on the capability of distinguishing the game state images.**

I. Introduction

Recent years, increasing attention has been payed to explain Deep Neural Networks(DNN). Most of the works on explainable DNN have been invested to image classification, such as visualizing each layer's activation or feature of classifier [1]. Deep reinforcement learning networks (DRLN), as one kind of DNN, show promising performance in selecting the most valuable action when interacting with environment. Such networks have been extremely successful in learning action control in Atari games. However, the inner decision process of DRLN works like a black box, which needs to be interpreted.

There are different ways to present the interpretation of DRLN. One of them is to interpret with image, such as saliency map [2]. Though it provides information, it explains the AIs behavior through perceptual level (i.e., pixels in the image) rather than conceptual level (i.e., plain text description). When people understand the images, subjectivity and ambiguity will be inevitably involved. Compared with images, verbal explanation, which presents the network behavior in conceptual level, would provide greater accuracy by filling the semantic gap. The gap is defined as 'the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation' [3]. In this paper, we use natural language to explain the reinforcement learning neural network to reduce the ambiguity.

Many DRLN systems are built on specific scenarios which are given by images, for instance, the games with fixed board or fixed map. Such scenarios often lead to high structural similarity in most of their game images, while the important features, which are the key to explain the behavior of the AI, are hidden in the dissimilar parts (the position of the pieces, the position of the main character in the game, etc.). Therefore, interpreting DRLN requires the capability of extracting the features that provide clues for behaviors from similar images.

Existing networks, however, are not good at extracting those features from highly similar images. For example, VGG [4] and Resnet [5], which showed excellent performance on the public datasets such as ImageNet [6], VisualQA, and CIFAR [7], fail to extract features from the images in Atari games. The extracted features are extremely close to each other, since all the state images are in the same background and foreground color. To solve the problem, we introduce a new encoder based on convolutional neural network. Furthermore, since attention has been proven to be effective in language related work [8], we design four attention structures to select the extracted features.

When explaining DRLN, it is a challenge to collect the ground truth of the 'correct explanation'. Lots of existing researches collected these data from humans explanations [9]. However, by doing this, there is the danger that the human's subjective explanation of the DRLN schema could be introduced. The human subjective logic was almost impossible to avoid, and the interpretation model would be trained to guess 'how human interpret DRLN', rather than the targeting networks own logic. In order to prevent the involvement of subjectivity from human, we design a separate constraints-based model to deduce the ground truth explanation from the object saliency map [10], [11] with a series of strong constraints covering DRLN's policies. These constraints are unified and defined by a group of people, which avoid subjective involvement as much as possible

This work has been funded by AFOSR grants FA9550-15-1-0442 and FA9550-18-1-0251

[1] Xinzhi Wang is Assistant Professor, Shanghai University, Shanghai, China, xinzhiw17@gmail.com

[2]Shengcheng Yuan is a Researcher of LazyComposer Inc., Beijing, China, yuansc@lazycomposer.com

[3]Hui Zhang is Professor at Beijing Key Laboratory of City Integrated Emergency Response Science, Tsinghua University, Beijing, China, zhhui@mail.tsinghua.edu.cn

[4]Michael Lewis is Professor at University of Pittsburgh, Pittsburgh, United State, ml@sis.pitt.edu

[5]Katia Sycara is Professor at the Robotics Institute, Carnegie Mellon University, Pittsburgh, United State, katia@cs.cmu.edu

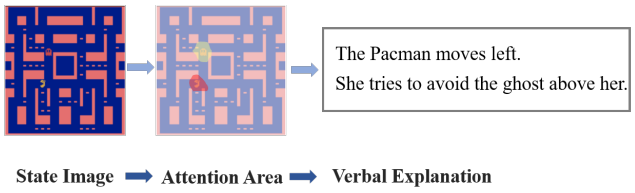


Fig. 1. Attention area of Atari game Ms.Pacman when generating verbal explanation

and are sufficient to train a learning model before being generalized from training scenario to the entire scenario. The object saliency map is only used in this constraints-based model, which assist the group of people designing unified constraints, and is not used in the interpretation model.

A verbal generation model with attention mechanism is presented in this paper. Game state images are taken as the model's inputs, and the output is the corresponding verbal explanation which is generated verbatim. The generation model consists of three parts: an encoder on feature extraction, an attention structure on weighting and highlighting the salient features, and a decoder on generating the explanation in natural language. We present the model that can attend to salient part of an image while generating its verbal explanation. We make the following contributions:

- We introduce a verbal generation model for DRLN. The model uses natural language rather than image as the explanation to reduce ambiguity. Additionally, the explanation labels which we collected for the model are generated through a series of unified constraints based on the object saliency map.
- We propose four variants of the attention structure in the verbal generation model to weight salient features. The four variants are named full attention, global attention, adaptive attention and object attention respectively. They imitate human attention on writing description given a specific image, as shown in figure 1.
- We introduce a new convolutional neural network as the image encoder in the verbal generation model to extract distinguishable features from similar images. The encoder has larger kernel size and performs better on our current task than VGG and Resnet.
- Finally, we quantitatively validate the performance of the model on Atari game Ms. Pacman. The results show that the adaptive attention structure performs the best among all the variants, even though the object attention structure is given additional information on object locations. Additionally, the proposed encoder outperforms two baseline encoders (Resnet and VGG) on the capability of distinguishing the game state images.

II. Related Work

In this section, we provide relevant background on previous works on DRLN and its interpretation.

DRLN concerns on how software agents ought to take actions in an environment by maximizing their cumulative reward. A deep Q-network, proposed by Mnih et al. [12], combined Q-learning with a flexible deep neural network. Hasselt et al. [13] improved it to a double deep Q-network to reduce the overestimation by decoupling the target max operation into both action selection and action evaluation. Wang et al. [14] proposed a dueling network architecture to yield better approximation of the state value.

People have long been accustomed to describing human behaviors or emotion using natural language. Graaf et al. [15] introduced theory of how people explain human behavior and sketch it to implement the underlying framework of explanation in Autonomous Intelligent Systems. Narayanan et al. [16] discussed what makes explanations interpretable in the specific context of verification. Thus, natural language plays an important role in the interaction between human and machine. Andreas et al. [17] focused on translating multi-agent communication policies into natural language. Das et al. [18] posed a cooperative game between two agents who communicate in natural language based on an unseen image by employing DRLN to learn the dialog policies. Wang et al. [19], [20] recognize emotions hidden in natural language. Shridhar et al. [21] presented a robot system that follows human natural language instructions to pick and place objects through interactively asking questions to disambiguate referring expressions.

Researchers committed to the interpretation of machine learning systems also in natural language, especially in the area of image classification. Vinyals et al. [22] and Chen et al. [23] presented recurrent neural network to generate natural sentences describing an image. Xu et al. [24] added an attention mechanism to deep recurrent architecture to describe the content of images. Hendricks et al. [25] focused on explaining why the predicted label is appropriate for the image after discriminating properties of visible object. Such works could also support the interpretation of DRLN, since DRLN usually takes images as input data when applied to game environment.

Existing researches on the interpretation of learning algorithm are represented with various forms. For example, Koh et al. [26] designed influence functions to trace a models prediction back to its training data. Du et al. [27] proposed a guided feature inversion framework towards effective interpretation. Zahavy et al. [28] introduced a methodology and tools to analyze Deep Q-networks. Iyers et al. [11] and Li et al. [10] contributed to transparency and explanation in DRLN thourgh saliency maps. Our work focus on verbal explanation, which is different from the above researches. Recently, Ehsan et al. [9] described a rationalization technique that translate internal state-action representations into natural language. Our work achieve more objective language explanation employing object saliency map and verbal generation model.

III. Verbal Generation Model with Attention

The verbal generation model aims to generate natural language explanation for DRLNs actions. It mainly solves two issues: (1) to extract distinguishable features from high

similar input images, and (2) to select salient features for generating verbal explanation verbatim.

The first issue stands out due to the fact that game state screen shots share high similarity with each other. It distinguishes from other public image datasets such as Imagnet, CIFAR-10, COCO, etc. The state of art neural networks, such as VGG, Resnet, and Alexnet, are difficult to extract distinguishable features from these successive game screens shots. To deal with it, we build a self-designed feature extraction network.

The second issue is inspired from humans attention behavior. If human look at an image and read its description at the same time, human tend to focus on different part of the image for each word or phrase. The same phenomenon occurs when generating verbal explanation for game state images, since different features may contribute to different words. To select potential features to generating each word, four attention mechanisms are designed.

Our proposed model employs an encoder-attention-decoder structure, whose workflow is shown in figure 2. The encoder mainly focuses on extracting features from the game state image. The attention structure, consisting of four variants, is proposed to select the salient features from the extracted features. The decoder is a language model based on Gated Recurrent Unit (GRU) to generate verbal explanation based on the salient features.

The details of the encoder attention and the decoder are given in the following.

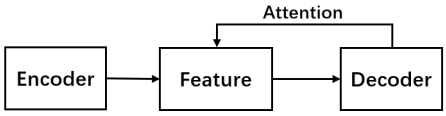


Fig. 2. Generalized AI verbal explanation model

A. Image Encoder

The image encoder aims at extracting features from the game state screenshots. Such extracted features may include location of objects, causal relations among objects, visual relations, etc. In this section, a convolutional neural network (CNN) is designed to capture these features implicitly from the game images.

All the game state images in this paper are 224×224 pixels in size, and contain 15 channels (5 frames with 3 channels in each frame). Let $\mathrm{I} \in \mathbb{R}^{224\times224\times15}$ being the matrix of state image. Let I_e denote the extracted features of image encoder, and function E denote our CNN encoder.

$$\mathrm{I}_e = E(\mathrm{I}|\mathrm{W}_e, \mathrm{B}_e) \tag{1}$$

Where the function E includes three convolution calculation with two large kernels (10×10) whose stride is set to be 2, three ReLU activation operation following the convolution calculation, and two batch normalization. W_e and B_e indicate the parameters in this process. In most pretrained image processing networks, such as VGG, kernel size is set to be (3×3). Kernel size with (10×10) is comparatively large, which is more capable of extracting distinguishable features.

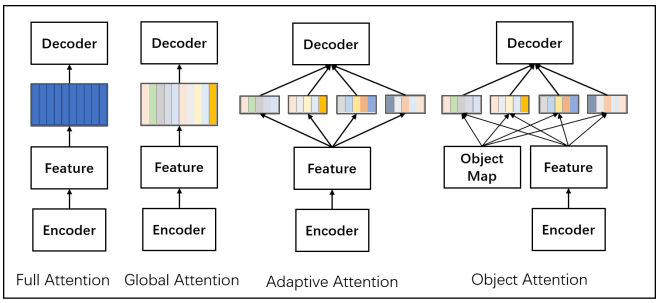


Fig. 3. Four variants of verbal generation model with full attention, global attention, adaptive attention, and object attention

I_e, as the extracted features, has 90 channels of matrix and the size of matrix is 21×21. Each channel represents a kind of image feature. I_e will further work as the input of the decoder.

B. Verbal Explanation Decoder

we introduce some common notations at the beginning of this section. Let $\eta(\cdot)$ denote the softmax operation, $[\cdot]$ denote the concatenate operation, and $L(\cdot)$ denote the linear operation.

Let $\mathrm{V} = \{v_1, ..., v_N\}$ be the vocabulary of the verbal explanations, where N is the size of the vocabulary table. $v_i \in \{0,1\}^N$ denotes the i-th word in the vocabulary with one-hot vector representation. Let $\mathrm{S} = (s_1, ..., s_n)$ denotes the word sequence, where n is the length of the sequence, and $s_t \in \mathrm{V}, \forall t \in \{1, 2, ..., n\}$. Let s_0 denotes the initial vector with the same size with s_t for each sequence.

The decoder employs a GRU structure followed by a softmax operation.

$$\mathrm{O}_t = GRU(\mathrm{A}_{attn}, s_{t-1}) \tag{2}$$

$$s'_t = argmax(\eta(\mathrm{O}_t)) \tag{3}$$

where A_{attn} is the attention on I_e, which will be discussed in the Attention section. O_t is the output of the GRU sequence model with the input A_{attn} and the sequence word s_{t-1}. s'_t is the prediction of word based on the softmax operation of O_t. Then, the negative log likelihood is calculated as follows.

$$l = -\sum_{t}^{n}\sum_{i}^{N} s_{ti}\log(s'_{ti}) \tag{4}$$

The goal of the model is to minimize the loss l between the true value of verbal explanation and the predicted verbal explanation.

C. Attention

As mentioned above, A_{attn} is the attention on I_e. Specifically, $\mathrm{A}_{attn} \in \{\mathrm{A}_f, \mathrm{A}_g, \mathrm{A}_a, \mathrm{A}_o\}$ with A_f being full attention, A_g being global attention, A_a being adaptive attention, and A_o being object attention. The four attentions employ

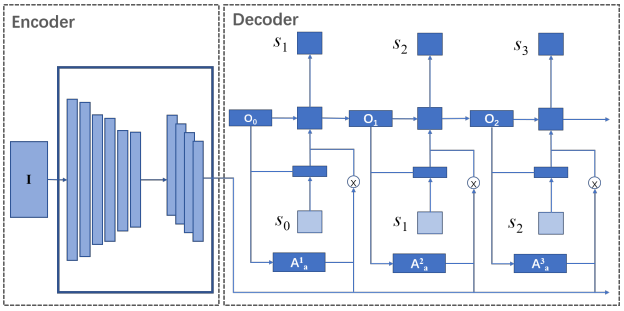


Fig. 4. Verbal generation model employing adaptive attention

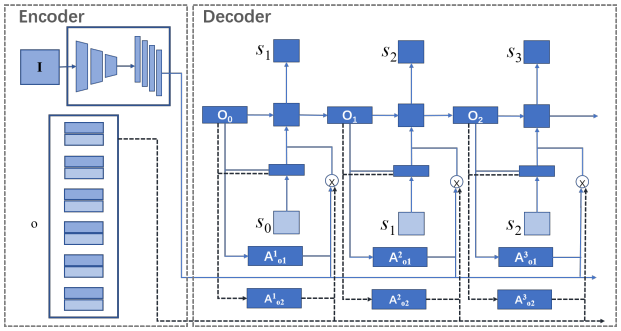


Fig. 5. Verbal generation model employing object attention.

different mechanism to select potential features, as is indicated in figure 3. Details are illustrated as follows.

1) Full attention: The full attention decoder assumes that all the features extracted from the encoder are considered equally as the input of the decoder. The calculation process is shown below.

$$\mathrm{A}_f = L(\mathrm{I}_e|\mathrm{W}_f, \mathrm{B}_f) \tag{5}$$

where A_f is the full attention result. W_f and B_f are the parameters for a linear function L.

2) Global attention: The global attention assumes that the extracted features are weighted only based on themselves (I_e), i.e., the self-attention mechanism. The attention is fixed once the decoder begins generating words. The calculation process is shown in the following.

$$\mathrm{A}_g = f_g(\mathrm{I}_e|\mathrm{W}_g, \mathrm{B}_g) \tag{6}$$

where f_g is the global attention operation. W_g and B_g are the weight and bias parameters for global attention calculation, where $\mathrm{W}_g = \{\mathrm{W}_{g_1}, \mathrm{W}_{g_2}\}$ and $\mathrm{B}_g = \{\mathrm{B}_{g_1}, \mathrm{B}_{g_2}\}$. Specifically, f_g can be divided into the following two steps.

$$i_e = \eta(conv_g(\mathrm{I}_e|\mathrm{W}_{g_1}, \mathrm{B}_{g_1})) \tag{7}$$
$$\mathrm{A}_g = L(i_e \cdot \mathrm{I}_e|\mathrm{W}_{g_2}, \mathrm{B}_{g_2}) \tag{8}$$

where $conv_g$ is a convolutional operation on I_e. There are three convolutional layers in $conv_g$ with parameters W_{g_1} and B_{g_1}. i_e is the self-attention distribution also based on I_e. Then, a linear operation on the multiplication result of i_e and I_e is conducted with parameters W_{g_2} and B_{g_2}. Finally, A_g works as part of input to the decoder.

3) Adaptive attention: Different from the full attention and the global attention, the adaptive attention has a dynamic assumption. For the t-th word in the verbal explanation, the extracted features are weighted only based on the previous word that was just generated by the decoder, i.e., the decoders previous state O_{t-1} and s_{t-1}. The calculation is shown in the following steps.

$$\mathrm{A}^t_a = f_a(\mathrm{I}_e, \mathrm{O}_{t-1}, s_{t-1}|\mathrm{W}_a, \mathrm{B}_a) \tag{9}$$

where A^t_a is the attention on predicting the t-th generated word and f_a is the adaptive attention function. The attentions are paid to encoder output I_e based on word s_{t-1} and output O_{t-1}. W_a and B_a are the weight and bias parameters, where $\mathrm{W}_a = \{\mathrm{W}_{a_1}, \mathrm{W}_{a_2}\}$ and $\mathrm{B}_a = \{\mathrm{B}_{a_1}, \mathrm{B}_{a_2}\}$. Specifically, f_a can be divided into following steps.

$$i^t_a = \eta(L([\mathrm{O}_{t-1}, s_{t-1}]|\mathrm{W}_{a_1}, \mathrm{B}_{a_1})) \tag{10}$$
$$\mathrm{A}^t_a = L(i^t_a \cdot \mathrm{I}_e|\mathrm{W}_{a_2}, \mathrm{B}_{a_2}) \tag{11}$$

where i^t_a is adaptive attention distribution. W_{a_1}, B_{a_1}, W_{a_2}, and B_{a_2} are the parameters for linear operation. The rest of the variables and functions keep their meanings in accordance with former description. Finally, the adaptive attention result contributes as the input of sequence decoder model. The verbal generation model employing adaptive attention is shown in figure 4.

4) Object attention: The object attention has the most complex assumption among all the attention variants. Beside the output of encoder, the object attention also weights on object maps, where the object maps contain additional information to locate the specific objects. For instance, there are six kinds of objects in Atari game MS. Pacman, and each type of the object has a corresponding object map. The calculation process is shown in the following.

$$\mathrm{A}^t_o = f_o(\mathrm{I}_e, \mathrm{o}, \mathrm{O}_{t-1}, s_{t-1}|\mathrm{W}_o, \mathrm{B}_o) \tag{12}$$

where f_o is object attention function. The attentions are paid to both encoder output I_e and objects map o based on the previous word s_{t-1} as well as the corresponding O_{t-1} from the decoder. W_o and B_o are weight and bias parameters for object attention calculation, where $\mathrm{W}_o = \{\mathrm{W}_{o_1}, \mathrm{W'}_{o_1}, \mathrm{W}_{o_2}, \mathrm{W'}_{o_2}\}$ and $\mathrm{B}_o = \{\mathrm{B}_{o_1}, \mathrm{B'}_{o_1}, \mathrm{B}_{o_2}, \mathrm{B'}_{o_2}\}$. Specifically, f_o can be divided into two parts. The first part is designed for object maps.

$$i^t_{o_1} = \eta(L([\mathrm{O}_{t-1}, s_{t-1}]|\mathrm{W}_{o_1}, \mathrm{B}_{o_1})) \tag{13}$$
$$\mathrm{A}^t_{o_1} = conv_{o_1}(i^t_{o_1} \cdot o|\mathrm{W'}_{o_1}, \mathrm{B'}_{o_1}) \tag{14}$$

where $i^t_{o_1}$ and $\mathrm{A}^t_{o_1}$ are the attention distribution and final attention result for object maps. $conv_{o_1}$ includes three convolutional layers with the corresponding parameters $\mathrm{W'}_{o_1}$ and

$\mathrm{B'}_{o_1}$. The meaning of the rest of the variables and functions are in accordance with former description. The second part is designed for encoder output.

$$i^t_{o_2} = \eta(L([\mathrm{O}_{t-1}, s_{t-1}]|\mathrm{W}_{o_2}, \mathrm{B}_{o_2}))) \quad (15)$$

$$\mathrm{A}^t_{o_2} = L(i^t_{o_2} \cdot \mathrm{I}_e|\mathrm{W'}_{o_2}, \mathrm{B'}_{o_2}) \quad (16)$$

where $i^t_{o_2}$ and $\mathrm{A}^t_{o_2}$ are the attention distribution and final attention result for encoder output I_e. The meaning of the rest of the variables and functions are in accordance with former description. Finally, $\mathrm{A}^t_{o_1}$ and $\mathrm{A}^t_{o_2}$ are concatenated together as the attention result A^t_o in object attention calculation.

$$\mathrm{A}^t_o = [\mathrm{A}^t_{o_1}, \mathrm{A}^t_{o_2}] \quad (17)$$

A^t_o works as part of the input to the GRU decoder, similar to the other attention variants. The working process of verbal generation model employing object attention is shown in figure 5.

IV. Experiments

We describe our experimental methodology and quantitative results which validate the effectiveness of our model for verbal explanation generation of reinforcement learning.

A. Data

We report the result on data from Atari game MS. Pacman. 19,419 game state images with verbal explanation are randomly selected as training data. 2,050 game state images with verbal explanation are left as testing data. The verbal explanations in the dataset are generated through strict constraints employing object saliency map [10], [11] from 30 game episodes. The constraints is designed by a group of human, and cover Pacman's action, Pacman's intention of chasing beneficial objects and Pacman's intention of avoiding evil objects.

TABLE I
Configuration of six models

Model	Abbr.	Encoder	Attention
Full Attention(Resnet)	FA(R)	Resnet	Full
Full Attention(VGG)	FA(V)	VGG	Full
Full Attention	FA	presented	Full
Global Attention	GA	presented	Global
Adaptive Attention	AA	presented	Adaptive
Object Attention	OA	presented	Object

B. Model Configuration

Six groups of experiments are implemented. They are named as full attention (Resnet), full attention (VGG), full attention, global attention, adaptive attention, and object attention model, whose abbreviations are FA(R), FA(V), FA, GA, AA, and OA respectively. In model FA(R), Resnet network is employed as encoder. In model FA(V), VGG network is employed as encoder. In the other four models, the presented large kernel encoder is conducted. When it comes to the attention mechanism, full attention is applied to model FA(R), model FA(V), and model FA. Global attention, adaptive attention, and object attention are adopted in the other three models (as their names imply). All the six models utilize GRU as the decoder. The configuration of models are shown in table I. All the models are implemented on Linux operating systems with GPU being GTX 1080 . About eight hours are consumed for training each single model.

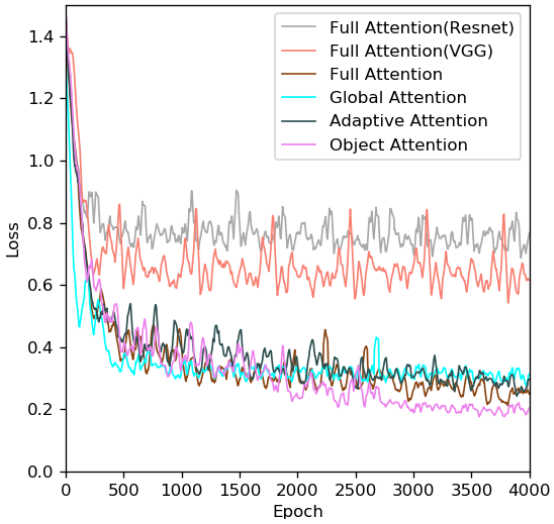
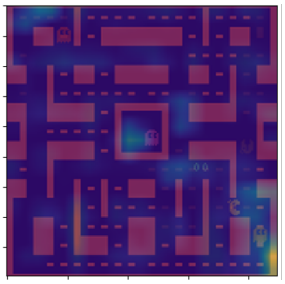


Fig. 6. Training losses of the six models with different encoder or attention



Object ghost and dot have drawn attention of Pacman. The Pacman moves right to eat the dot in the lower right even she is approaching the ghost in the lower right.

Fig. 7. The case of the attention visualization for model AA when generating sentence verbatim. The highlighted area is the attention area when generating corresponding word

C. Evaluation Procedures

There are two groups for comparison. The first group focuses on the choice of convolutional encoder. In our evaluation, we compare the presented encoder (FA) directly with Resnet (FA(R)) and VGG (FA(V)), who work as feature extractor for state image. The second group aims at comparing the four kinds of attention mechanism, i.e, FA, GA, AA, OA.

1) Loss Trend: During the training period, loss function is set to be the negative log likelihood, i.e., equation 4. The loss trend on the training data is shown in figure 6. As shown in the figure, the final losses of both model FA(R) and model FA(V) are around 0.7. Model FA, equipped with the same attention and decoder with FA(R) and FA(V) but different

TABLE II
BLEU SCORE OF GENERATED VERBAL EXPLANATIONS

Training data				
Model	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**
FA(R)	0.536	0.399	0.293	0.209
FA(V)	0.621	0.478	0.382	0.304
FA	0.797	0.735	0.681	0.635
GA	0.487	0.374	0.297	0.236
AA	**0.890**	**0.850**	**0.816**	**0.786**
OA	0.651	0.556	0.478	0.412
Testing data				
Model	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**
FA(R)	0.544	0.414	0.309	0.224
FA(V)	0.626	0.490	0.396	0.319
FA	0.703	0.620	0.548	0.487
GA	0.471	0.357	0.281	0.221
AA	**0.710**	**0.626**	**0.554**	**0.493**
OA	0.628	0.529	0.449	0.382

encoder, reduces its loss to lower than 0.4. It means that the presented encoder can extract more potential features than VGG and Resnet in the game state images. The stable losses of models FA, GA, AA, and OA are all around 0.3. These four models adopt the presented encoder and four kinds of attention. It is hard to say which model is better as their losses are so close to each other. Therefore, the following evaluation methods are designed to assess the validation of models.

2) BLEU Score: BLEU[29] is a metric for evaluating the quality of text, which is considered to be the correspondence between original and target text. Here, BLEU scores from 1 to 4 are employed to evaluate the generated explanations. The scores are listed in table II. From the table, it can be concluded that model AA performs the best among the six models, and got the highest score on both training data and testing data on BLEU-1, BLEU-2, BLEU-3, and BLEU-4. Model FA(R) and model FA(V) perform worse compared with model FA, model AA and model OA. The result that model FA performs better than FA(R) and FA(V) proves that our proposed encoder is more suitable in extracting distinguishable features, as only encoder differs. The result that model AA performs better than FA means the adaptive attention is more efficient in generating word sequences when compared with full attention. Model GA shows the worst performance on both training and testing data. It indicates that the global attention, whose attention does not change with the generated words, can not get useful features which can meet all the requirement of all generated words in all time steps. Also, even though additional features are supplemented for model OA, its performance is not the best.

3) Evaluation through Action and Object Prediction: Beside BLEU score, we evaluate the generated verbal explanations through action and object words. The action words include 'up', 'down', 'left', and 'right', indicating the moving action. The object words are 'cherry', 'dot', 'pellet', 'ghost', and 'edible ghost', describing all the game objects except pacman. We collect the recall (labeled as x-r with x being action or object word), precision (labeled as x-p with x being action or object word) on words, and the overall accuracy (labeled as accur) as the quantitative evaluations. The results are shown in table III and IV. Here, only the top three models on BLEU score (FA, AA, and OA) are evaluated.

From the two tables, it can be concluded that model AA performs the best on predicting the action words, with an overall accuracy of 0.8797 on training data and 0.6431 on testing data. Model FA gains the best performance on predicting the object words. Model AA performs competitively with model FA except the recall and precision on object 'cherry'. It means that model AA can not distribute attention on the game object 'cherry' when watching the game state images. The reason may be that 'cherry' appears less frequently and shifts fast in game episode. Model OA also performs well on predicting game objects, as the objects locations are provided as additional information. Table III and IV illustrate part of the interpretation details of model FA, AA and OA from action and object prediction. Table II is the overall evaluation through BLEU score.

4) Adaptive Attention Visualization: We visualize the attention of model AA, which performs well on all the above evaluations comprehensively. The presented encoder (in model FA, GA, AA and OA) encodes the game state images in size of 224×224 into the output in size of $90 \times 21 \times 21$. It means that 90 layers of features are extracted. The layer which gets the highest attention is selected to be visualized. We upsample this layer from size 21×21 to size 224×224 by applying a bilinear filter. Then, we mask the upsampled attention layer on the original game state image.

One case is shown in figure 7. In the case, the ghost in the lower right is highlighted several times when generating words. The explanation increases human trust in DRNL systems by explaining some confusing DRNL actions decisions. In this case, human may distrust the DRNL system when Pacman approaches ghost and by ignoring Pacman's main intention that to eat the dot in the lower-right.

V. CONCLUSION

In this paper, we present a verbal generation model for reinforcement learning employing attention on extracted features, which consists of three parts: encoder on feature extraction, attention structure on selecting salient features from the output of the encoder, and decoder on generating the explanation in natural language. Four variants of the attention structure - full attention, global attention, adaptive attention and object attention - are designed to select distinguishable features. Experiments are implemented on Atari game Ms. Pacman. The adaptive attention structure performs the best among all the variants, even though the object attention structure is given additional information on object locations. Additionally, our experiment results showed that the proposed encoder outperforms two baseline encoders (Resnet and VGG) on the capability of distinguishing the game state images. The model proposed in this paper can be applied on other applications given the input image as training data and pre-trained reinforcement learning models as interpretation target.

TABLE III

EVALUATION OF ACTION PREDICTION

Training data									
Model	**up-r**	**up-p**	**down-r**	**down-p**	**left-r**	**left-p**	**right-r**	**right-p**	**accur**
FA	0.7593	0.7816	0.5841	0.7667	0.7499	0.7131	0.7332	0.6726	0.7198
AA	**0.8702**	**0.9227**	**0.8404**	**0.8480**	**0.9135**	**0.8681**	**0.8718**	**0.8809**	**0.8797**
OA	0.1978	0.2095	0.0832	0.1725	0.2086	0.3412	0.5291	0.3134	0.2878
Testing data									
Model	**up-r**	**up-p**	**down-r**	**down-p**	**left-r**	**left-p**	**right-r**	**right-p**	**accur**
FA	**0.7149**	0.7149	0.4808	**0.6318**	0.6194	**0.6613**	**0.7099**	0.5864	0.6424
AA	0.6748	**0.7632**	**0.6148**	0.5829	**0.6667**	0.6270	0.6125	**0.6188**	**0.6431**
OA	0.4855	0.4589	0.0329	0.1558	0.3933	0.4692	0.6582	0.4095	0.4264

TABLE IV

EVALUATION OF OBJECT PREDICTION

Training data											
Model	**cherry-r**	**cherry-p**	**dot-r**	**dot-p**	**ghost-r**	**ghost-p**	**edible-r**	**edible-p**	**pellet-r**	**pellet-p**	**accur**
FA	0.6519	0.9060	0.9959	0.9969	0.7547	0.8028	0.9099	0.8544	0.8538	0.8242	0.9139
AA	0.0172	0.0058	0.9978	0.9974	0.7576	0.9067	0.9502	0.9425	0.4807	0.7470	0.8196
OA	0.1429	0.0015	0.9913	0.9985	0.5939	0.7988	0.9194	0.5411	0.4999	0.7843	0.7820
Testing data											
Model	**cherry-r**	**cherry-p**	**dot-r**	**dot-p**	**ghost-r**	**ghost-p**	**edible-r**	**edible-p**	**pellet-r**	**pellet-p**	**accur**
FA	0.5152	0.4857	0.9899	0.9939	0.5350	0.5207	0.8050	0.8123	0.3838	0.3775	0.7918
AA	0.0244	0.0286	0.9884	0.9879	0.4736	0.5697	0.8012	0.8128	0.3444	0.6162	0.7151
OA	0.5000	0.0286	0.9870	0.9960	0.4260	0.5770	0.8199	0.5500	0.4046	0.7659	0.7206

REFERENCES

[1] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," *arXiv preprint arXiv:1506.06579*, 2015.

[2] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[3] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 12, pp. 1349–1380, 2000.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 2009, pp. 248–255.

[7] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Citeseer, Tech. Rep., 2009.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[9] U. Ehsan, B. Harrison, L. Chan, and M. O. Riedl, "Rationalization: A neural machine translation approach to generating natural language explanations," *arXiv preprint arXiv:1702.07826*, 2017.

[10] Y. Li, K. Sycara, and R. Iyer, "Object sensitive deep reinforcement learning," in *3rd Global Conference on Artificial Intelligence*. EPiC Series in Computing, 2017.

[11] R. Iyer, Y. Li, H. Li, M. Lewis, R. Sundar, and K. Sycara, "Transparency and explanation in deep reinforcement learning neural networks," *In Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, 2018.

[12] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015. [Online]. Available: http://dx.doi.org/10.1038/nature14236

[13] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," *CoRR*, vol. abs/1509.06461, 2015. [Online]. Available: http://arxiv.org/abs/1509.06461

[14] Z. Wang, T. Schaul, M. Hessel, H. Van Hasselt, M. Lanctot, and N. De Freitas, "Dueling network architectures for deep reinforcement learning," *arXiv preprint arXiv:1511.06581*, 2015.

[15] M. de Graaf and B. F. Malle, "How people explain action (and autonomous intelligent systems should too)," in *AAAI Fall Symposium on Artificial Intelligence for Human-Robot Interaction*, 2017.

[16] M. Narayanan, E. Chen, J. He, B. Kim, S. Gershman, and F. Doshi-Velez, "How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation," *arXiv preprint arXiv:1802.00682*, 2018.

[17] J. Andreas, A. Dragan, and D. Klein, "Translating neuralese," *arXiv preprint arXiv:1704.06960*, 2017.

[18] A. Das, S. Kottur, J. M. Moura, S. Lee, and D. Batra, "Learning cooperative visual dialog agents with deep reinforcement learning," *arXiv preprint arXiv:1703.06585*, 2017.

[19] X. Wang, H. Zhang, S. Yuan, J. Wang, and Y. Zhou, "Sentiment processing of social media information from both wireless and wired network," *Eurasip Journal on Wireless Communications Networking*, vol. 2016, no. 1, pp. 1–13, 2016.

[20] X. Wang, X. Luo, and J. Chen, "Social sentiment detection of event via microblog," in *IEEE International Conference on Computational Science Engineering*, 2013.

[21] M. Shridhar and D. Hsu, "Interactive visual grounding of referring expressions for human-robot interaction," *arXiv preprint arXiv:1806.03831*, 2018.

[22] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE*

conference on computer vision and pattern recognition, 2015, pp. 3156–3164.

[23] X. Chen and C. Lawrence Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2422–2431.

[24] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[25] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *European Conference on Computer Vision*. Springer, 2016, pp. 3–19.

[26] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," *arXiv preprint arXiv:1703.04730*, 2017.

[27] M. Du, N. Liu, Q. Song, and X. Hu, "Towards explanation of dnn-based prediction with guided feature inversion," *arXiv preprint arXiv:1804.00506*, 2018.

[28] T. Zahavy, N. Ben-Zrihem, and S. Mannor, "Graying the black box: Understanding dqns," in *International Conference on Machine Learning*, 2016, pp. 1899–1908.

[29] K. PAPINENI, "Bleu : a method for automatic evaluation of machine translation," in *Proc Meeting of the Association for Computational Linguistics*, 2002.