High-level Visual Representation via Perceptual Representation Learning

Donghun Lee
Autonomous IoT Research Section
Electronics and Telecommunications Research Institute
Daejeon, South Korea
donghun@etri.re.kr

Samyeul Noh
Autonomous IoT Research Section
Electronics and Telecommunications Research Institute
Daejeon, South Korea
samuel@etri.re.kr

Ingook Jang
Autonomous IoT Research Section
Electronics and Telecommunications Research Institute
Daejeon, South Korea
ingook@etri.re.kr

Seonghyun Kim
Autonomous IoT Research Section
Electronics and Telecommunications Research Institute
Daejeon, South Korea
kim-sh@etri.re.kr

Heechul Bae
Autonomous IoT Research Section
Electronics and Telecommunications Research Institute
Daejeon, South Korea
hessed@etri.re.kr

***Abstract*— Recent advancements in the field of representation learning and video prediction have demonstrated the potential for enhancing manipulation and control strategies across various applications through precise anticipation of future states. Nevertheless, the intricate dynamic nature inherent in real-world data poses a formidable challenge in acquiring these representations. Autoregressive models, which employ the generated future frame as input for the subsequent frame prediction, suffer from issues such as compounding errors, memory overload, and extended training times due to the need for reconstructing the state from the latent vector in each iteration. To address these limitations, recent studies have introduced the concept of State Space Models (SSMs) to forecast from the latent space, offering the advantage of predicting distant future states. However, these methodologies often exhibit restricted capabilities in extracting object-centric representations. More recent object-centric approaches concentrate on closely associated features from the input data, yet their ability to capture higher-level representations remains constrained. In this paper, we propose integrating a perceptual network into the slot attention mechanism to facilitate the extraction and segregation of high-level representations. Leveraging a pre-trained perceptual network, we derive elevated object-oriented representations for each perceptual layer, aligning them with corresponding slots. This elevated representation, rich in object-centric information, holds the potential to enhance comprehension of the present state and provide valuable guidance for accurate future state prediction.**

Keywords*— *Representation Learning, Object-centric, Generalization, Image Analysis

I. Introduction

Recent progress in representation learning and video prediction has underscored the potential of precise future state prediction to enhance manipulation and control strategies across various domains, including edge devices [1], multiagent [2], autonomous driving [3], navigation [4], manipulation [6], drones [7] and simulators [16]. However, the complex and dynamic nature inherent in real-world data presents significant challenges in acquiring effective representations.

Deterministic methods, such as RNNs, exhibit limitations in handling dynamic environments. Autoregressive models like SAVP [8], SVG [9], and SV2P [10] employ generated future frames as inputs for predicting subsequent frames. Nonetheless, these techniques repeatedly reconstruct the state from the latent vector, rendering them susceptible to compounded errors, memory issues, and prolonged training times [11].

Recent exploration of the State Space Model (SSM) paradigm, exemplified by PlaNet [14], seeks to leverage the latent space for predicting future states, offering the advantage of forecasting outcomes in the distant future. However, these approaches often encounter challenges in effectively capturing object-centric representations.

Object-centric strategies, such as slot attention, have recently exhibited promising outcomes in representation learning, particularly in handling novel compositions. While these approaches employ attention mechanisms [12] for image and video predictions [13], they primarily concentrate on closely related input features, resulting in limited efficacy in extracting high-level representations.

This study introduces the integration of a perceptual network with slot attention, aiming to extract and disentangle high-level representations. Leveraging a pre-trained perceptual network, we obtain elevated object-oriented representations for each perceptual layer, aligning them with the respective slots. This advanced object-oriented representation holds potential for enhancing the comprehension of the present state and providing valuable guidance for accurate future state prediction.

II. Method

In this paper, we introduce the Perceptual Slot Attention module, a novel extension of the Slot Attention module, designed for object-oriented perceptual representation learning. This new module comprises four essential components: the Input Module, the Perceptual Network Module, the Attention Module, and the Output Module.

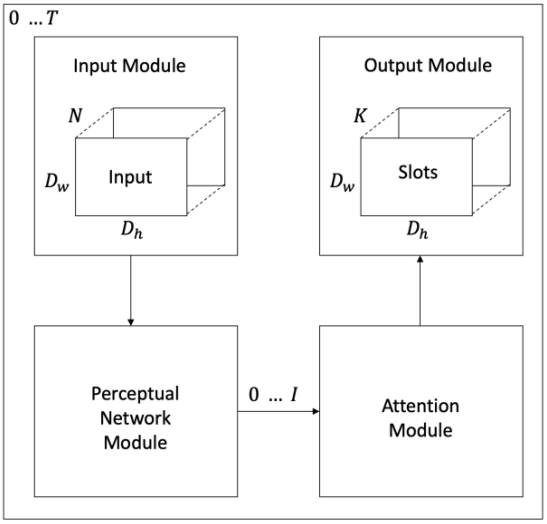

Fig. 1. The Perceptual Slot Attention Module consists with four component which are Input Module, Perceptual Network Module, Attention Module and Output Module. It demonstrates the process of mapping *N* vector inputs to *K* slots across each perceptual layer *i* of the input. In this way, high dimentinal input resource is presented in low dimentional abstracted representation.

Perceptual Slot Attention

The objective of the Perceptual Slot Attention module is to map a set of N input vectors to K slots for each perceptual layer *i* of the input, as visually shown in figure 1. The comprehensive methodology is detailed in algorithm 1, presented in pseudo-code.

The Input module receives N input vectors, each consisting of D dimensions. This paper specifically employs static images as the input source. The Perceptual Network module generates elevated representations for each perceptual layer *i* of the input image. To achieve this, a pre-trained

VGG16 network [15], trained on the ImageNet dataset, is harnessed to extract high-level representations for each perceptual layer *i*.

The attention module adheres to the established slot attention convention. For every perceptual layer *i*, the corresponding slot is initialized following a Gaussian distribution characterized by shared parameters μ and σ. Layer normalization is applied to the input, slots, and each perceptual layer *i* of the slot. The attention mechanism operates for T iterations on each perceptual layer *i* of the slot. During the Softmax procedure, individual slots[1..*i*] compete to capture nearby features, undergoing iterative updates. The temperature parameter for the Softmax function is set to a fixed value of $\sqrt{N}$. Weighted mean is employed as the update mechanism to enhance stability.

The Output module furnishes *K* slots as an output derived from the original *N* input vectors. This transformation effectively reduces the *N*-dimensional input to *K* dimensions while preserving high-level representations.

Algorithm 1. The Module for Perceptual Slot Attention facilitates the mapping of input, comprising *N* vectors, to *K* slots, each having *D*_inputs dimensions, within every perceptual layer *i*. The slots for each perceptual layer are initialized randomly through a Gaussian distribution. The perceptual layer carries high-level representation of the raw input data. In our experiment, we conduct 3 iterations ($T = 3$) across 4 perceptual layers ($I = 4$).

Input: $images \in \mathbb{R}^{N \times D_{images}}$,
$slots_i \sim N(\mu, diag(\sigma)) \in \mathbb{R}^{K \times D_{images}}$
Layer params: k, q, v: $linear\ projections\ for\ attention$;
i: $perceptual\ layer$; $LayerNormalization$
Images = LayerNormalization (images)
for $t = 0 \ldots T$
for i = 0 … I
$slots\_prev_i = slots_i$
$slots_i$ = LayerNorm(slots)
$attn_i$ = Softmax($\frac{1}{\sqrt{N}} k(images) \cdot q(slots)^T$,
$axis = 'slots_i\ ')$
$updates_i$ = WeightedMean(weights=$attn_i + \epsilon$,
values=$v_i(images)$
$slots_i$ += LayerNormalization($slots_i$)
return slots

III. Experiments

In the experiments of this study, the BAIR Push Dataset was utilized, and the Structural Similarity Index Map (SSIM) [5] was used as the evaluation metric. Taking five past frames as input, predictions were made for ten future frames, and the SSIM values were compared and analyzed against the Ground Truth.

For the experiments of this study, input of size 64, RNN Layer of size 256, dimensions of 128 for 'g', and 10 for 'z' were utilized. Training was carried out with a batch size of 100, using the Adam Optimizer with a momentum of 0.9, over 600 epochs and 300 iterations. Posterior RNN Layer and Predictor RNN Layer were employed in quantities of 1 and 2, respectively.

Figure 2 presents a comparison of the performance between the proposed algorithm and the baseline algorithm SVG-LP [8]. The proposed algorithm demonstrates higher SSIM values from t+3 onwards.

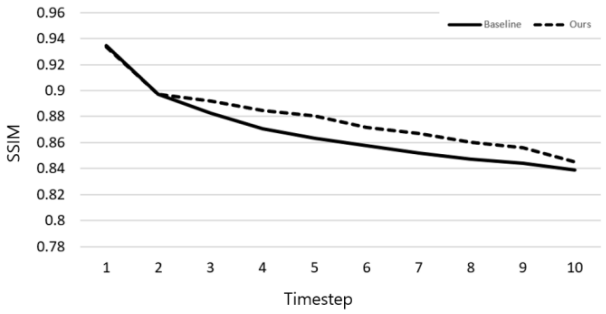

Fig. 2. Figure 2 illustrates the experimental outcomes, offering a comparative analysis of the performance between the proposed algorithm and the baseline algorithm SVG-LP. Notably, the proposed algorithm exhibits superior SSIM values starting from t+3.

IV. Conclusion

This study introduces an approach that employs slot attention on the outputs of perceptual networks to acquire high-level representations. The pre-trained perceptual network generates elevated representations for each perceptual layer, aligning them with corresponding slots. This enriched object-oriented representation holds the potential to enhance comprehension of the present state and provide valuable guidance for precise future state prediction.

Nonetheless, certain limitations exist in this research. To establish the efficacy of the proposed method, validation through testing on both images and videos is imperative, a task slated for future investigations. Moreover, if the approach were to be evaluated on simulation data, it could potentially extend the experiments to encompass physical manipulation tasks as well.

Acknowledgment

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [23ZR1100, A Study of Hyper-Connected Thinking Internet Technology by autonomous connecting, controlling, and evolving ways]

References

[1] Jang, Ingook, et al. "An approach to share self-taught knowledge between home IoT devices at the edge." Sensors 19.4 (2019): 833.

[2] Kim, Hyunseok, et al. "Avoiding collaborative paradox in multi-agent reinforcement learning." ETRI Journal 43.6 (2021): 1004-1012.

[3] Noh, Samyeul, et al. "Co-pilot agent for vehicle/driver cooperative and autonomous driving." ETRI Journal 37.5 (2015): 1032-1043.

[4] Jung, Sunggoo, et al. "Collision-free local planner for unknown subterranean navigation." ETRI Journal 43.4 (2021): 580-593.

[5] Wang, Zhou, et al. "Image quality assessment: from error visibility to structural similarity." IEEE transactions on image processing 13.4 (2004): 600-612.

[6] Lee, Ahyun, Joo-Haeng Lee, and Jaehong Kim. "Data-Driven Kinematic Control for Robotic Spatial Augmented Reality System with Loose Kinematic Specifications." ETRI Journal 38.2 (2016): 337-346.

[7] Jung, Sunggoo, et al. "Collision-free local planner for unknown subterranean navigation." ETRI Journal 43.4 (2021): 580-593.

[8] Lee, Alex X., et al. "Stochastic adversarial video prediction." arXiv preprint arXiv:1804.01523 (2018).

[9] Denton, Emily, and Rob Fergus. "Stochastic video generation with a learned prior." International conference on machine learning. PMLR, 2018.

[10] Babaeizadeh, Mohammad, et al. "Stochastic variational video prediction." arXiv preprint arXiv:1710.11252 (2017).

[11] Oprea, Sergiu, et al. "A review on deep learning techniques for video prediction." IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

[12] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[13] Singh, Gautam, Fei Deng, and Sungjin Ahn. "Illiterate dall-e learns to compose." International Conference on Learning Representations. 2021.

[14] Hafner, Danijar, et al. "Learning latent dynamics for planning from pixels." International conference on machine learning. PMLR, 2019.

[15] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[16] Lee, Donghun, et al. "Learning Control Policy with Previous Experiences from Robot Simulator." 2020 International Conference on Information and Communication Technology Convergence (ICTC). IEEE, 2020.