社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

Attention 機構を用いた物体中心表現学習

中田　秀基†　　麻生　英樹†

† 産業技術総合研究所　〒 305–8560 茨城県つくば市梅園 1-1-1
E-mail: †{hide-nakada,h.asoh}@aist.go.jp

あらまし　動画像を用いた表現学習では、個々の物体をマスクで分離した上で個別に物体表現を教師なしで学習する手法が広く用いられている。これらの表現学習手法の性能は高く、多くのダウンストリームタスクで高い性能を示しているが、計算量が膨大であるという問題点がある。われわれは、従来の動画表現学習手法である ViMON をベースとし、これに Attention 機構を導入することで、性能を維持しつつ計算量を低減することを試みた。Attention 機構を導入する位置によって 2 つの手法を提案し、それぞれ実装を行い、再構成誤差、実行時間、ダウンストリームタスクの性能で評価を行った。その結果、ベースとなる手法と比較して、より高い性能を示しながら大幅な計算量の低減できることを確認した。

キーワード　表現学習, 教師なし機械学習, Attention 機構

Object-Centric Representation Learning with Attention Mechanism

Hidemoto NAKADA† and Hideki ASOH†

† National Institute of Advanced Industrial Science and Technology
Umezono 1–1–1, Tsukuba, Ibaraki, 305–8560 Japan
E-mail: †{hide-nakada,h.asoh}@aist.go.jp

Abstract　For object-centric representation learning, several slot-based methods, that separate objects using masks and learn the objects separately, are proposed. While these methods are proved to be useful on various downstream tasks, it is known that they require a significant amount of computation for training. We propose the introduction of attention mechanisms into slot-based method to simplify and speed up the computation. We pick ViMON as the base structure and propose two methods, named AttnViMON and SFA. We evaluate them in terms of reconstruction error and computation time, and a downstream task. The proposed methods demonstrate that they achieve significant speed-up while showing even better performance.

Key words　Representation Learning, Unsupervised Learning, Attention Mechanism

1. はじめに

動画像を用いた教師なし表現学習は、VQA を始めとするさまざまなダウンストリームタスクに使用することが可能であり、注目を集めている。人間は、世界を一連の物体の集まりとして認識していると考えられることから、個々の物体を分離した上で、個別に表現を学習する手法が有効であると考えられている。この手法の実装として、個々の物体をマスクで分離し、個別に学習する手法が知られている [1]。静止画に対してマスクを生成するネットワークと、マスクされたあとの物体を再構成するネットワークを組み合わせる構成をベースとし、それを時間軸方向に拡張する方法が数多く試みられている。これらの表現学習手法の性能は高く、多くのダウンストリームタスクで高い性能を示しているが、計算量が膨大であるという問題点がある。われわれは、従来の動画表現学習手法である ViMON [1] をベースとし、これに Attention 機構を導入することで、性能を維持しつつ計算量を低減することを試みた。Attention 機構を導入する位置によって 2 つの手法 AttnViMON と SFA を提案し、それぞれを実装して、再構成エラーと実行時間およびダウンストリームタスクで評価を行った。その結果、ベースとなる手法と比較して、より低い再構成エラーとより高いダウンストリームタスク性能を示しつつ、大幅な計算量の低減を達成することを確認した。

本稿の貢献は、以下の 2 つである。

- ViMON をベースとした 2 つの手法を提案した
- 2 つの手法を評価し、計算量の低減と、性能の向上を確認した。

本論文の構成は以下の通りである。2. で本研究と関連する既

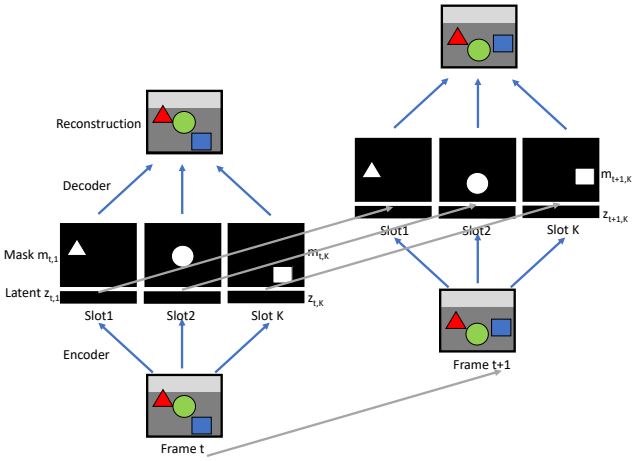

図 1　スロットベースの手法

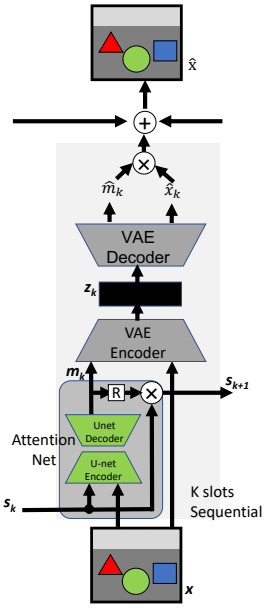

図 2　MONet の概要

存技術を概説し、3. で提案手法について述べる。4. で実験結果を示す。6. では、まとめと今後の課題について述べる。

2.　背　　景

2.1　スロットベースの手法

文献 [1] に従って、スロットを用いた物体中心の動画像教師なし学習手法を概観する。スロットベースの手法は、図 1 に示す基本的な構造を共有している。各タイムフレームの画像は、エンコーダによって複数のスロットと呼ぶマスク $m_{t,k}$ と隠れ変数 $z_{t,k}$ にエンコードされ、その後デコーダによって再構成される。基本的にはこの再構成画像と元の画像の差によって学習を行う。動画像においてはこの作業を各画像ごとに行うのだが、その際に同一スロットがおなじ物体を表すように何らかの方法で時間軸方向で情報を共有する必要がある。また、各スロットが排他的に異なる情報を表現するように、なんらかの方法で情報を共有する必要がある。本稿では、スロット間の情報共有を水平方向で、時間軸方向の情報共有を右上がりの斜め線で表現する。

このような手法には、われわれがベースラインとして用いた ViMON の他に、TBA(Tracking-by-Animation) [2]、OP3(Object-centric Perception, Prediction, and Planning) [3]、SCALOR(SCALable Object-oriented Representation) [4] などがある。

2.2　MONet

本節では、ViMON のベースとなった MONet について説明する。MONet [5] は、セグメンテーションマスクを用いてオブジェクトセグメンテーションを行う教師なし学習モデルである。MONet は、セグメンテーションマスクを生成する注視マスク生成ネットワーク (Attension Network (注1)) を訓練すると同時に、マスクを用いて分離した画像に対して再構成を行う VAE [6] を訓練する。図 2 に概要を示す。この図は、1 つのスロットに対する処理のみを取り出したものとなっている。下から入力した画像 $\mathbf{x}$ は各スロットに分割されて処理され、最後に最上段に $\hat{\mathbf{x}}$ として再構成される。

MONet は入力画像のスロットへの分割を複数ステップの繰り返し処理で行う。各ステップでは、入力画像に対する注視マスク生成ネットワークを訓練し、そのネットワークの出力したマスクと元画像を入力として VAE を訓練する。この VAE は通常の VAE と同様に入力画像を再構成するように訓練されるが、その際に注視マスクの範囲外に関しては無視するよう訓練される。

各ステップの注視マスク訓練ネットワークは、U-Net [7] 構造を持つネットワークである。このネットワークは、入力として元の画像 $\mathbf{x}$ の他に、スコープと呼ばれる入力 $\mathbf{s}_k$ を前段から受け取る。スコープはそのステップ以降で処理する対象を示すマスクである。注視マスク訓練ネットワークは、マスク $\mathbf{m}_k$ と、次のスロットで用いるスコープ $\mathbf{s}_k + 1$ とを出力する。これらのネットワークはすべて 1 つのロスで End-to-End で訓練される。最後のステップでは、注視マスク訓練ネットワークを用いず、前段から受け取ったスコープそのものをマスクとして VAE を訓練する。

ロスは式 1 のように定義される。

$$L = \sum_{t=1}^{T} (L_{\text{recon}} + \beta L_{\text{prior}} + \gamma L_{\text{mask}}) \qquad (1)$$

ここで、L_{recon} は再構成ロスで、時刻 t における入力画像と再構成画像で計算される。L_{prior} は VAE 部の事前分布との KL ダイバージェンスである。L_{mask} はマスクに関する再構成誤差で、注視マスク生成ネットワークの出力するマスクと VAE の出力するマスクの KL ダイバージェンスで計算される。β および γ は各項の重みを決定するハイパーパラメータである。

2.3　ViMON

ViMON(VideoMONet) [1] は、静止画を対象とした MONet を動画像に拡張したものである。図 3 にこの様子を示す。この

(注1)：この Attention は「注視」の意味で、いわゆる Attention 構造を指すものではないことに注意。

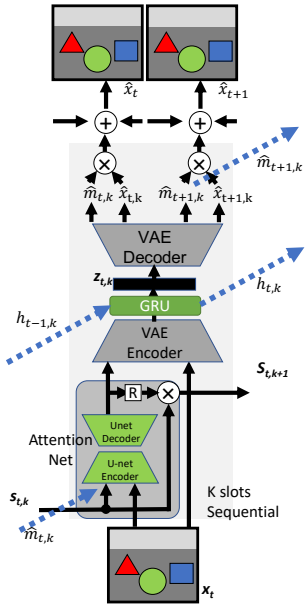

図 3　ViMON の概要

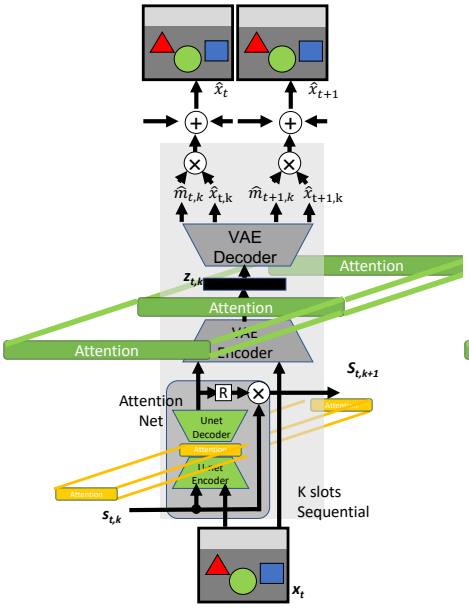

図 4　AttnViMON の概要

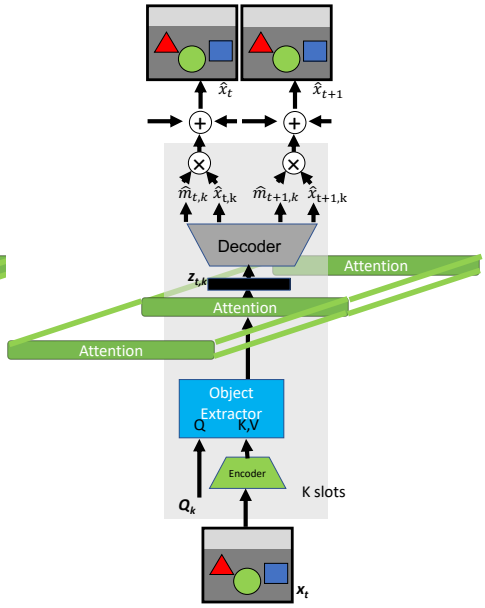

図 5　SFA の概要

図は 1 タイムステップの 1 スロットのみを表した物となっており、実際には多数のスロットと多数のタイムステップに対しておなじ処理を行っていることに注意されたい。

ViMON は、静止画に対する構造は MONet と同一だが、以下の 3 つの方法で時間軸方向の情報を利用するように拡張されている。

- VAE のエンコーダ出力を GRU [8] に与え、前フレームからの情報 $\mathbf{h}_{t-1,k}$ を取り込むと同時に、次のタイムフレームに $\mathbf{h}_{t,k}$ として送る (図 3 中段)。
- VAE で予測した次フレームのマスク $\hat{\mathbf{m}}_{t+1,k}$(図 3 右上の破線矢印) を、次フレームのマスク生成ネットワークの入力の 1 つ $\hat{\mathbf{m}}_{t,k}$ として使用する (図 3 左下の破線矢印)。
- VAE の出力を用いてその時刻のフレーム $\hat{\mathbf{x}}_t$ を予測するだけでなく、次のフレーム $\hat{\mathbf{x}}_{t+1}$ の予測を行い、次フレームの正解画像との差分を訓練に利用する。

AttnViMON のロスは式 2 のように定義される。

$$L = \sum_{t=1}^{T}(L_{\text{recon}} + L_{\text{pred}} + \beta L_{\text{prior}} + \gamma(L_{\text{mask}} + L_{\text{mask\_pred}})) \tag{2}$$

ここで、L_{recon}、L_{prior}、L_{mask} は、ViMON と同じである。L_{pred} は、時刻 t で予測した $t+1$ における値と時刻 $t+1$ の真の値の差分である。$L_{\text{mask\_pred}}$ は、時刻 t で VAE で予測した $t+1$ におけるマスクと、時刻 $t+1$ の注視マスク生成ネットワークの出力の間の KL ダイバージェンスである。

3. 提 案 手 法

3.1 AttnViMON

MONet は、各スロットの処理に依存関係があり、前段のスロットの実行が終了しないと次のスロットに対する処理ができない。ViMON は、このスロット方向の依存関係に加えて、時間軸方向にも GRU やマスクに対する依存関係があるため、並列に実行できない。時間軸方向の依存関係を解消するために、時間軸方向の情報伝達を GRU やマスクの伝搬ではなく、Attention 機構 [9] を導入する方法を考案した。この手法を AttnViMON と呼ぶ。AttnViMON の概要を図 4 に示す。

AttnViMON では、注視マスク生成ネットワーク (図 4 下部) と、VAE の中間層部分 (図 4 中央) の 2 箇所で Attention 機構を用いている。注視マスク生成ネットワークでの Attention 機構は各スロットの時系列に沿った情報の交換に用いる。この Attention 機構はスロットの数だけ存在する。位置情報としては時刻を用いる。一方で VAE 部の Attention 機構は、すべてのスロットとすべてのタイムステップの間の情報交換を可能にする。この Attention 機構は 1 つしか存在しない。位置情報としては、スロットと時刻の両方を用いる。また、Attention 機構では、時系列を逆行する情報の伝播が起こらないように、マスクしている。

3.2 SFA

AttnViMON では、時系列方向の情報伝播を Attention 機構に置き換えた。これに加えてスロット方向の情報伝播も Attention 機構にまとめ、さらに、注視マスク生成と VAE を融合したものが SFA(Slots and Frames Attention) である。図 5 に概要を示す。

SFA では物体抽出ネットワーク (図中 Object Extractor) を用いて、各時刻、各スロットの中間表現を直接生成する。物体抽出ネットワークは、各スロットを表す訓練可能なベクタ $\mathbf{Q}_k$ をクエリとし、コンボリューションネットワークで入力画像 $\mathbf{x}_t$ をエンコードした値を、キーとバリューとするマルチヘッドアテンションネットワークである。$\mathbf{Q}_k$ は、すべての時刻において共通で、ここにオブジェクトの位置不変な情報が学習されることを期待している。

物体抽出ネットワークの出力は Attention 機構で処理される。

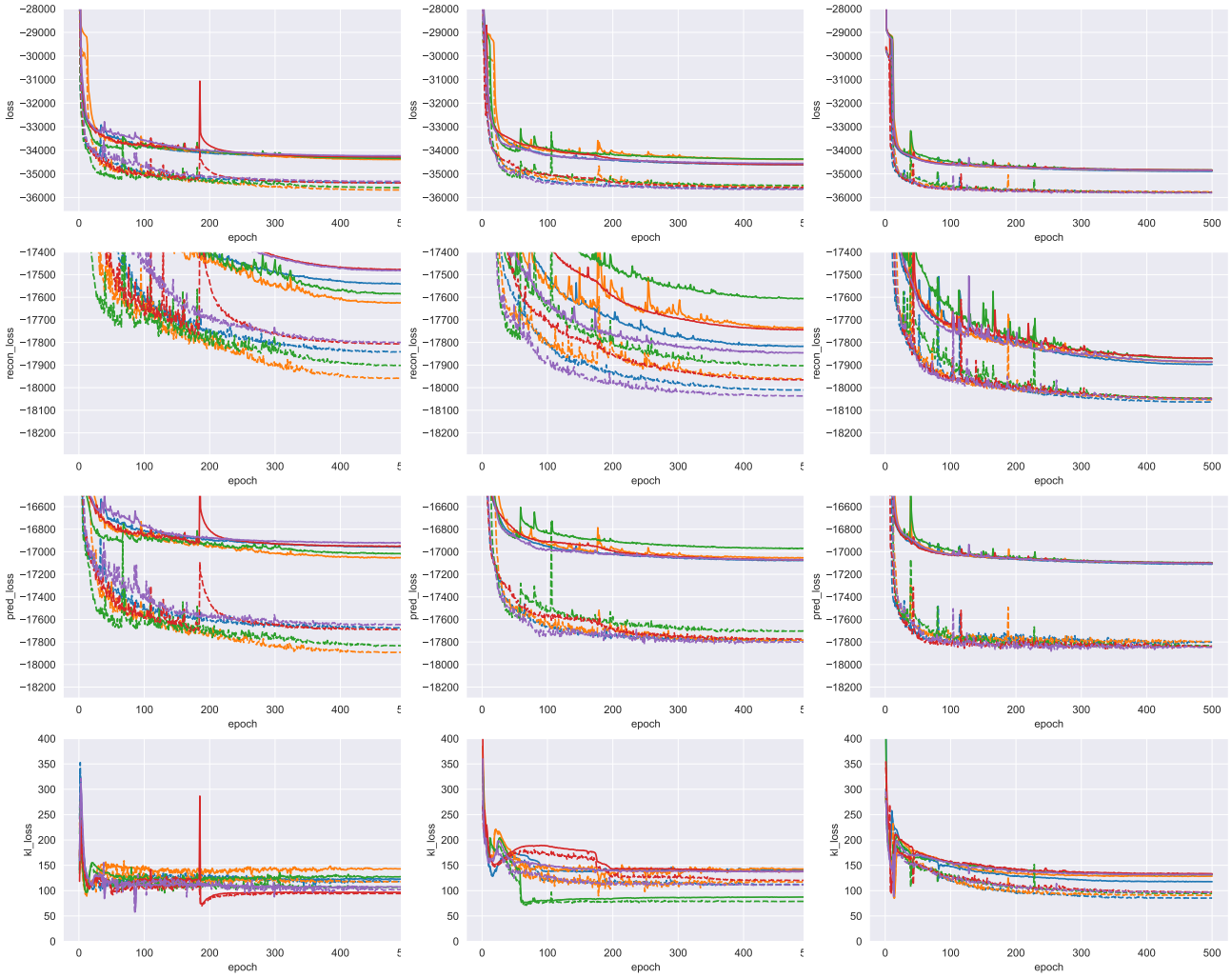

図 6　ViMON のロス

図 7　AttnViMON のロス

図 8　SFA のロス

この Attention 機構は AttnViMON の VAE 部の Attention 機構と同様に、スロット方向と時刻方向の双方に対して情報を交換する。したがって、位置情報にはスロットと時刻の双方を用いる。

SFA のロスは式 3 のように定義される。

$$L = \sum_{t=1}^{T} (L_{\text{recon}} + L_{\text{pred}} + \beta L_{\text{prior}}) \tag{3}$$

マスク生成ネットワークによるマスク生成を行わないため、マスクに関するロスが省略されている。

4. 評　　価

4.1 評 価 手 法

評価は、ロスの各項目の値と、最適化にかかる時間で行った。データセットには後述する CLEVRER [10] を用いた。スロット数は 8、各スロットの VAE の隠れ変数は 16 とした。入力には連続した 10 フレームを用いた。最適化手法は Adam [11] で、学習率としては、初期値を $4e^{-4}$、500 エポックでゼロとなるコサインアニーリング [12] を用いた。初期値を決める乱数のシードを変えて 5 回試行を行った。

実験には産総研の保有する ABCI [13] の V-node を 1 ノード用いた。V-node には、NVIDIA V100 が 4 機搭載されている。

4.2 ロスの挙動

図 6、図 7、図 8 に、各手法の訓練の際のロスを示す。最上段から、ロスの総計、再構成ロス、予測ロス、VAE の KL ロスである。それぞれ実線が検証セットのロス、破線が訓練セットのロスである。5 本の線はそれぞれの試行に対応する。

ViMON と AttnViMON では試行ごとのばらつきが大きく、初期値に対して鋭敏であることがわかる。これに対して SFA は安定しており、初期値の変動に対して寛容であることが伺える。

4.3 再構成誤差と次フレーム予測誤差

学習終了時の再構成誤差と次フレーム予測誤差をそれぞれ図 9 と図 10 に示す。平均値と最大値最小値を表示している。いずれの場合も SFA が高い性能を示している。また、SFA では値のばらつきが小さいことがここでも確認できる。

4.4 訓 練 時 間

ABCI の V-node を 1 ノード専有し、V100 を 4 機を用いて学習した。この際の訓練時間を表 1 に示す。5 回実行した際の平均値である。ViMON が約 76000 秒 (およそ 21 時間) かかっているのに対して、AttnViMON では約 42500 秒 (およそ 12

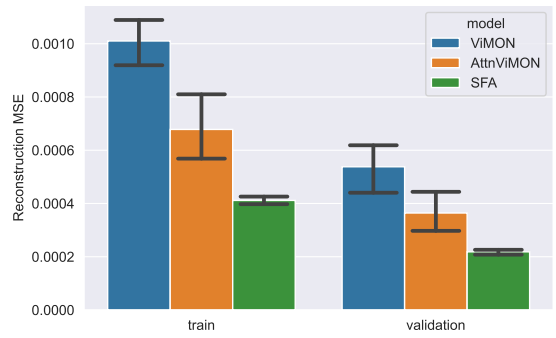

図 9　再構成誤差

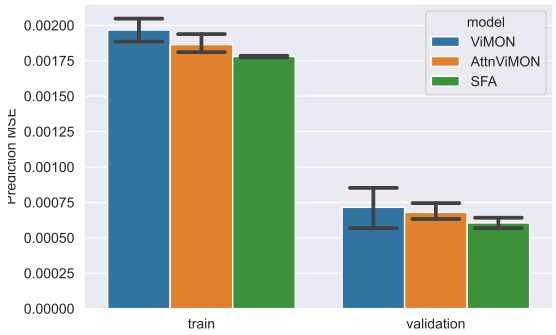

図 10　次フレーム予測誤差

表 1　訓 練 時 間

手法	訓練時間 (秒)
ViMON	76,107
AttnViMON	42,507
SFA	27,116

時間) と高速化された。さらに SFA では約 27000 秒 (およそ 7 時間半) と大幅に高速化されたことがわかる。

高速化の理由については十分な解析ができていないが、GRU による逐次化が排除されたことによる並列化が有効であったと思われる。特に SFA ではスロット間の依存関係も完全に排除され、すべてのスロットと時刻について並列に実行できる。また、SFA ではネットワークそのものが大きく簡素化され、注視マスク生成ネットワークと隠れ変数の学習ネットワークがいわば融合していることで、計算量そのものも大きく低減できている。

4.5　Aloe による評価

ダウンストリームタスクとして Aloe [14] を用いて、評価を行った。Aloe は動画像に対する質問応答をするネットワークで、動画像をフレームごとにエンコードした埋め込みと、質問文の単語単位の埋め込みを入力とし、回答の単語もしくは質問の当否を返答するネットワークである。

データセットとしては CLEVRER [10] を用いた。CLEVRER は、それぞれ形状と色彩で区別のつく物体が移動して相互に衝突する様子を示す合成動画データセットで、個々の動画に対していくつかの質問が用意されている。質問は、Descriptive(記述的), Explantory(説明的), Predictive(予測的), Counterfactual(反実仮想的) の 4 つに分類される。Descriptive な質問に対しては単語を返答し、それ以外の質問に対しては質問中に列挙される言明に対してそれぞれ正誤を返答する。ここでは、Descriptive を Desc、それ以外を合わせて MC と呼ぶ。

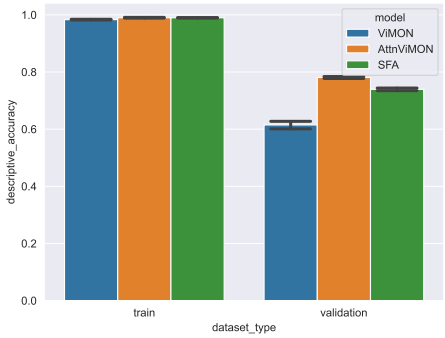

図 11　Desc 正答率

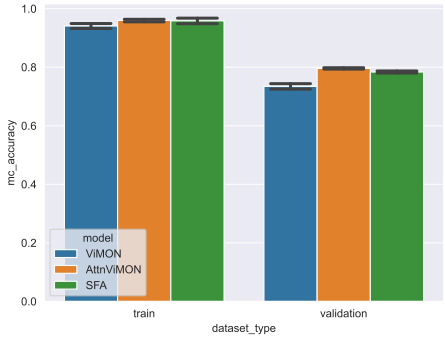

図 12　MC 正答率

ここでは、前節での実験で得られた ViMON, AttnViMON, SFA のそれぞれの手法のネットワークのうち、最も検証ロスが小さいものを用いて CLEVRER データセットをエンコードして Aloe を訓練した。Aloe ネットワークの初期値を決定する乱数シードを変更しつつ 5 回実験を行った。

評価結果を図 12 と図 11 に示す。いずれの場合も AttnViMON では ViMON と比較して性能が大きく向上している。SFA は AttnViMON よりはやや劣るものの、ViMON よりは高い性能を示している。

5.　関 連 研 究

SAVi(Slot Attention for Video) [15] は、時刻 t における各スロットを表す表現 $\mathbf{S}_t = [\mathbf{s}_t^1, .., \mathbf{s}_t^k]$ を、時間軸にそって更新するモデルである。各時刻 t では、われわれのオブジェクト抽出ネットワークに類似する Corrector と呼ばれるマルチヘッドアテンションネットワークを用いて、$\mathbf{S}_t$ と時刻 t における入力画像 $\mathbf{x}_t$ から、その時刻のスロット表現 $\hat{\mathbf{S}}_t$ を導出する。われわれが時間軸方向の情報伝播に Attention 機構を用いているのに対して、SAVi では繰り返し処理を用いている点が異なる。

6.　お わ り に

本稿では、動画を対象とした物体中心表現学習機構に Atten-

tion 機構を導入することで、性能を維持しつつ計算量を削減することを試みた。AttnViMON と SFA の 2 つのネットワークを提案し、CLEVRER を用いて評価した。その結果大幅な速度向上を確認できた。また、ダウンストリームタスクとして質問応答ネットワーク Aloe を用いて評価を行った。AttnViMON、SFA の双方において性能の向上が見られた。

今後の課題としては以下が挙げられる。

- 他のダウンストリームタスクでの評価
- SFA の Aloe と End-to-End ファインチューニング

われわれは、Aloe と MONet を End-to-End でファインチューニングすることで、性能向上が得られることを確認している [16]。ViMON のように時間軸に依存関係を持つ複雑な構造のネットワークでは Aloe と End-to-End トレーニングすることは難しいが、SFA のように単純なネットワークであれば可能であると思われる。今後検討を進める予定である。

謝　　辞

実装をお手伝いいただいた井上辰彦氏に感謝します。

文　　献

[1] M.A. Weis, K. Chitta, Y. Sharma, W. Brendel, M. Bethge, A. Geiger, and A.S. Ecker, “Unmasking the Inductive Biases of Unsupervised Object Representations for Video Sequences,” CoRR, vol.abs/2006.07034, , 2020. https://arxiv.org/abs/2006.07034

[2] Z. He, J. Li, D. Liu, H. He, and D. Barber, “Tracking by animation: Unsupervised learning of multi-object attentive trackers,” CoRR, vol.abs/1809.03137, , 2018. http://arxiv.org/abs/1809.03137

[3] R. Veerapaneni, J.D. Co-Reyes, M. Chang, M. Janner, C. Finn, J. Wu, J.B. Tenenbaum, and S. Levine, “Entity abstraction in visual model-based reinforcement learning,” CoRR, vol.abs/1910.12827, , 2019. http://arxiv.org/abs/1910.12827

[4] J. Jiang, S. Janghorbani, G. deMelo, and S. Ahn, “Scalable object-oriented sequential generative models,” CoRR, vol.abs/1910.02384, , 2019. http://arxiv.org/abs/1910.02384

[5] C.P. Burgess, L. Matthey, N. Watters, R. Kabra, I. Higgins, M. Botvinick, and A. Lerchner, “MONet: Unsupervised Scene Decomposition and Representation,” CoRR, vol.abs/1901.11390, , 2019. http://arxiv.org/abs/1901.11390

[6] D.P. Kingma and M. Welling, “Auto-encoding variational bayes,” Proceedings of the 2nd International Conference on Learning Representations, 2014.

[7] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” CoRR, vol.abs/1505.04597, , 2015. http://arxiv.org/abs/1505.04597

[8] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling,” CoRR, vol.abs/1412.3555, , 2014. http://arxiv.org/abs/1412.3555

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in Neural Information Processing Systems, eds. by I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, vol.30, Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[10] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J.B. Tenenbaum, “Clevrer: Collision events for video representation and reasoning,” International Conference on Learning Representations, 2020. https://openreview.net/forum?id=HkxYzANYDB

[11] D.P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” https://arxiv.org/abs/1412.6980, 2014.

[12] S. Correa, “Cosine learning rate decay,” https://scorrea92.medium.com/cosine-learning-rate-decay-e8b50aa455b.

[13] “ABCI AI Bridge Infrastructure: https://abci.ai/”. Accessed: 2023-02-01.

[14] D. Ding, F. Hill, A. Santoro, M. Reynolds, and M.M. Botvinick, “Attention over learned object embeddings enables complex visual reasoning,” Advances in Neural Information Processing Systems 34 pre-proceedings (NeurIPS 2021), 2021.

[15] T. Kipf, G.F. Elsayed, A. Mahendran, A. Stone, S. Sabour, G. Heigold, R. Jonschkowski, A. Dosovitskiy, and K. Greff, “Conditional object-centric learning from video,” CoRR, vol.abs/2111.12594, , 2021. https://arxiv.org/abs/2111.12594

[16] H. Nakada and H. Asoh, “End-to-end training of object segmentation task and video question-answering task,” Proc. of The 17th International Conference on Ubiquitous Information Management and Communication, 2023.