

メタ学習としての Generative Query Network

A Meta-Learning Perspective on Generative Query Network

谷口 尚平^{*1}
Shohei Taniguchi岩澤 有祐^{*1}
Yusuke Iwasawa松尾 豊^{*1}
Yutaka Matsuo^{*1}東京大学
The University of Tokyo

Generative Query Network (GQN) is a novel deep generative model for three-dimensional modelling, making it possible to render images from unknown viewpoints. However, properly training GQN requires huge learning costs (including both computational time and hardware), and it is severally sensitive to hyper-parameters. Moreover, the lack of studies on GQN from a probabilistic view makes it difficult to interpret the neural network architecture. In this paper, we formulate the probabilistic model of GQN using meta-learning framework, and based on the formulation, we propose a new method that significantly reduces learning costs, and improves the accuracy of generated images, learning efficiency and robustness against hyper-parameters. We show the effectiveness through experiments using the Shepard Metzler dataset.

1. 研究の背景と目的

Generative Query Network (GQN) [Eslami 18] は、画像と視点座標という2つのセンサー情報のみから環境に関する抽象表現 (シーン表現) を獲得し、未知の視点座標からの観測画像を予測する深層生成モデルであり、Structure from Motion (SfM) などの従来の3次元モデリング手法に代わる新しい手法として注目を集めた。しかし、GQN は学習にかかる時間的・計算リソース的なコストが大きく、また結果がハイパーパラメータに強く依存するため学習が安定しないという課題がある。また、確率モデルとしての検証が不十分であるために、モデルアーキテクチャの解釈性が低く、発展研究の妨げとなっている。本研究では、これらの課題を解決するため、GQN の確率モデルをメタ学習のフレームワークを用いて定式化し、それに基づいて、学習のコストと不安定性を改善する手法の提案を行う。評価実験では Shepard Metzler データセットを用いてその有効性を検証する。

2. 前提知識

2.1 Generative Query Network

Generative Query Network (GQN) [Eslami 18] は、深層生成モデルの1種で、3次元空間上のある視点からの観測画像がいくつか与えられた状況において、別の視点からの観測画像を予測して生成するモデルである。学習においては、多数あるシーンの中から1つをランダムに選び、その中からモデルに事前に与えるコンテキストと予測させるクエリを選択する。モデルは、コンテキストとクエリの視点座標を入力され、クエリに対応する観測画像を予測して生成する。

GQN は、大きく分けて2つのニューラルネットワークで構成される。1つはコンテキストからシーン表現 r を抽出する**表現ネットワーク (representation network)**、もう1つはシーン表現 r とクエリ視点 v^q からクエリ画像 x^q を出力する**生成ネットワーク (generation network)** である。生成ネットワークは、さらに r と v^q から正規分布に従う確率的な潜在変数 z を出力するネットワーク ($\pi(z|r, v^q)$) と r , v^q , z から

x^q を出力するネットワーク ($p(x^q|z, r, v^q)$) の2段階で構成される。

学習においては、潜在変数 z の周辺化を回避するために、別のニューラルネットワークで定義される近似事後分布 $q(z|x^q, v^q, r)$ を用意し、以下の対数尤度の変分下限を目的関数として最大化するように確率的勾配降下法アルゴリズムでパラメータの最適化を行う。

$$\mathbb{E}_{q(z|x^q, r, v^q)} [\log p(x^q|z, r, v^q)] - D_{\text{KL}}[q(z|x^q, r, v^q) || \pi(z|r, v^q)] \quad (1)$$

2.2 メタ学習

メタ学習とは、複数のデータセットが与えられ、それぞれのデータセットごとにドメインやタスクが異なるような状況において、学習対象のドメインやタスクに応じて学習器のバイアスを決定するためのメタ知識を獲得する枠組みを指す [Vilalta 2002]. [Gordon 2018] では、メタ学習を統一的に理解するためのフレームワーク ML-PIP を提案しており、ML-PIP ではメタ学習において仮定する確率モデルを以下のようなグラフィカルモデルで表現する。

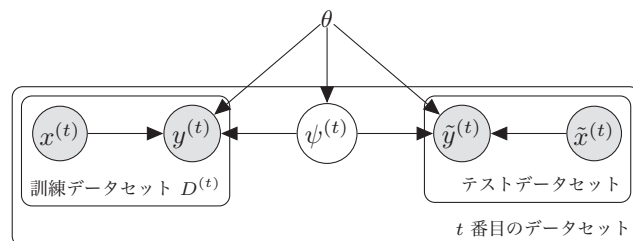


図 1: ML-PIP のグラフィカルモデル

メタ学習で期待されるのは、未知のデータセットに対して、少数の訓練サンプル $D^{(t)}$ からデータセットに固有のパラメータ (もしくは変数) である $\psi^{(t)}$ を推定し、 $\psi^{(t)}$ を用いてテストデータ $\hat{x}^{(t)}, \hat{y}^{(t)}$ に対して正しい予測を行うことである。この $\psi^{(t)}$ の推定を近似事後分布 $q(\psi^{(t)}|D^{(t)}; \phi)$ で行い、以下の尤

連絡先: 谷口 尚平, 東京大学, 〒113-8656 東京都文京区本郷 7-3-1, taniguchi@weblab.t.u-tokyo.ac.jp

度を目的関数として最大化することで、メタ学習の手法を統一的に定式化することができる。

$$\begin{aligned} p(\tilde{y}^{(t)}|\tilde{x}^{(t)}, D^{(t)}; \theta) \\ = \int p(\tilde{y}^{(t)}|\tilde{x}^{(t)}, \psi^{(t)}; \theta) p(\psi^{(t)}|D^{(t)}; \theta) \\ \simeq \int p(\tilde{y}^{(t)}|\tilde{x}^{(t)}, \psi^{(t)}; \theta) q(\psi^{(t)}|D^{(t)}; \phi) \end{aligned} \quad (2)$$

3. メタ学習としての GQN の定式化

GQN において対象とするタスクは、メタ学習のフレームワークに当てはめて解釈することが可能である。つまり、モデルは未知のシーン i に関して、少数のコンテキスト $D_i = \{(v_i^k, x_i^k)\}_{k=1}^M$ からシーンに固有な変数 r_i を推論し、それを用いてクエリの視点座標から対応する観測画像の予測を行う。すると、GQN において、コンテキスト D_i を入力としてシーン表現 r_i を出力する表現ネットワークは、2.2 節における近似事後分布 $q(\psi^{(t)}|D^{(t)}; \phi)$ のモデルとなっていると解釈することができる。すると、GQN の確率モデルは以下のように表現することができる。

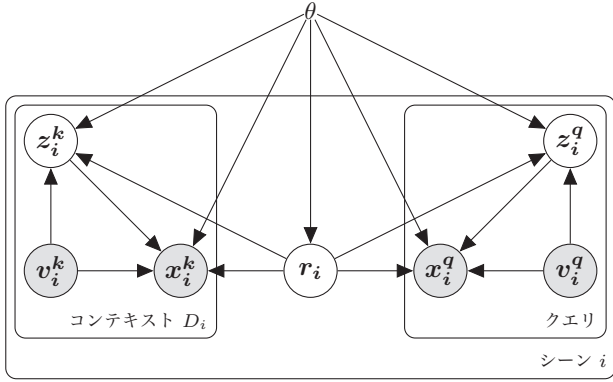


図 2: メタ学習としての GQN のグラフィカルモデル

図 1 と比較するとわかるように、GQN では通常のメタ学習とは異なり、シーン表現 r_i に加えて、個別のデータ (v_i^k, x_i^k) に依存する確率的な潜在変数 z_i^k をさらに定義している。それに伴い、式 (2) の目的関数は、GQN においては変数 r_i, z_i^q を用いて以下のように変化する。

$$\begin{aligned} \log p(x_i^q|v_i^q, D_i; \theta) \\ \simeq \log \int p(x_i^q|v_i^q, r_i; \theta) q'(r_i|D_i; \phi') dr_i \end{aligned} \quad (3)$$

$$= \log p(x_i^q|v_i^q, r_i; \theta) \quad (r_i = f(D_i; \phi')) \quad (4)$$

$$\begin{aligned} &= \log \int p(x_i^q|v_i^q, z_i^q, r_i; \theta) \pi(z_i^q|v_i^q, r_i; \theta) dz_i^q \\ &\geq \mathbb{E}_{z_i^q \sim q(z_i^q|x_i^q, v_i^q, r_i; \phi)} [\log p(x_i^q|v_i^q, z_i^q, r_i; \theta)] - \text{D}_{\text{KL}}(q||\pi) \end{aligned} \quad (5)$$

これは、式 (1) と一致する (ただし、式 (1) では、シーンの表記 i を省略している)。GQN ではシーン表現 r_i を推論する近似分布 $q'(r_i|D_i; \phi')$ を決定論的な表現ネットワークでモデル化しているため、式 (4) において、変数 r_i に関する周辺化が省略されていることに注意されたい。

このように、メタ学習として GQN の確率モデルを解釈すると、以下のような不十分な点が存在することがわかる。

不確実性の考慮が不明瞭

GQN では、観測されない潜在的な変数として、 r_i と z_i^k の 2 つを定義しており、 r_i は表現ネットワークによって決定論的に推論し、 z_i^k は正規分布を仮定して確率的な推論を行なっている。 z_i^k において、確率的な推論を行うのは、コンテキストからシーン固有の知識を推論する際に含まれる不確実性を考慮するためであると考えられる (例えば、コンテキストとして 1 つの視点からの観測のみが与えられた場合、物体の影となっていて観測できない部分の情報はそれのみから得ることはできず、不確実である)。しかし、 z_i^k は各データ点 v_i^k, x_i^k に依存する変数となっているため、ここで考慮される不確実性はシーンに関する知識自体のものではなく、各データ点でのレンダリングにおける不確実性となっている。一方で、シーン表現 r_i の推論は決定論的に行い、ここでは不確実性は考慮されないため、GQN ではシーン固有の知識そのものについての不確実性は明示的にはモデル化されていない。このような不確実性についての不明瞭なモデル化は、モデルの解釈性を下げるだけでなく、観測不可能な変数を増やすことで、多くの学習パラメータが必要となり、学習にかかるコスト増大の原因になっていると考えられる。

冗長な近似

GQN では、式 (3) における $q'(r_i|D_i; \phi')$ による $p(r_i|D_i; \theta)$ の近似に加えて、式 (5) において $q(z_i^q|x_i^q, v_i^q, r_i; \phi)$ による $p(z_i^q|x_i^q, v_i^q, r_i; \theta)$ の近似が発生しており、学習時には 3 つのパラメータ θ, ϕ, ϕ' を同時に最適化することが必要となっている。GQN の学習の不安定性はこのような冗長な近似による複雑な最適化が主な原因になっていると考えられる。

4. 提案手法

前章の議論を踏まえて、GQN の問題点を改善する手法の提案を行う。提案手法では、GQN におけるシーン表現 r_i と潜在変数 z_i^k を正規分布に従う 1 つの確率的な潜在変数 z_i に統合し、冗長な近似分布の排除と潜在変数の確率的な推論を可能にする。提案手法の確率モデルは、以下のようなグラフィカルモデルで表現される。

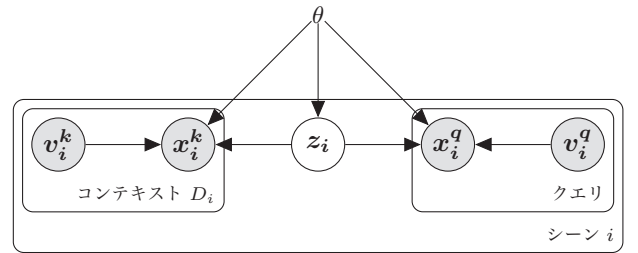


図 3: 提案手法のグラフィカルモデル

これに伴い、最大化する目的関数は以下のような対数尤度の近似下界となる。

$$\begin{aligned} \log p(x_i^q|v_i^q, D_i; \theta) &= \log \int p(x_i^q|v_i^q, z_i; \theta) p(z_i|D_i; \theta) dz_i \\ &\simeq \log \int p(x_i^q|v_i^q, z_i; \theta) q(z_i|D_i; \phi) dz_i \\ &\geq \mathbb{E}_{z_i \sim q(z_i|D_i; \phi)} [\log p(x_i^q|v_i^q, z_i; \theta)] \end{aligned} \quad (6)$$

ただし、式 (6) の期待値計算は、分布 $q(\mathbf{z}_i|D_i;\phi)$ からサンプリングした $\mathbf{z}_i$ を用いた生成画像 $\mathbf{x}_i^q$ の尤度の平均でモンテカルロ近似する。

5. 評価実験

提案手法の有効性を検証するため、評価実験によって Generative Query Network と提案手法の比較を行う。

5.1 データセット

実験では、DeepMind 社がインターネット上で公開している Shepard Metzler データセットを用いる [GQN Dataset]. Shepard Metzler データセットでは 7 つのパーツで構成される物体が 3 次元空間上に存在するようなシーンが 200 万用意され、シーンごとに物体のパーツの構成や色が異なる。データセットはそれぞれのシーンで様々な視点から物体を観測した 15 枚の 64×64 ピクセルの画像で構成されている。データセットの概要は図 4 を参照。実験では、このデータセットを 9 : 1 の割合で訓練用と検証用に分けて使用する。訓練時には、まず 200 万のシーンの中から 1 つをランダムにサンプリングする。そして、そのシーン i の中からコンテキストとして用いる画像の枚数を 1 ~ 15 の中からランダムに選び、その枚数 (M) 分の画像をデータセットからランダムに抽出して、 $D_i = \{(\mathbf{v}_i^k, \mathbf{x}_i^k)\}_{k=1}^M$ とする。さらに、残りの画像の中から予測の対象とするものを 1 つランダムに選び、クエリ $(\mathbf{v}_i^q, \mathbf{x}_i^q)$ として用いる。この操作を毎ステップ行い、そのデータを用いてミニバッチ学習を行う。

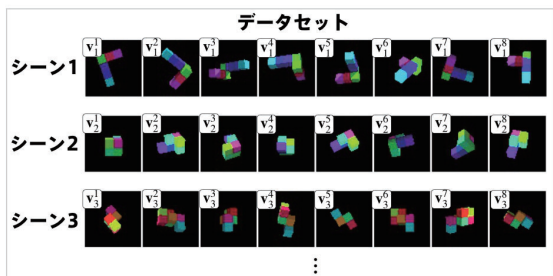


図 4: データセットの概要

5.2 評価指標

今回は、両手法の生成画像の定性的な評価に加えて、検証用データセットからコンテキストが 3 つ与えられた場合におけるクエリの予測画像の精度を定量評価指標として用いる。ここでは、予測精度として、平均をモデルの予測画像のピクセル値、分散を 1 とした正規分布に対する正解画像のピクセル値の負の対数尤度を用いる。これは、ピクセル値の平均二乗誤差と等価である。

また、これに加えて、ハイパーパラメータに対する頑健性を評価するために、潜在変数の次元数 (チャンネル数) を 3 段階 (3, 32, 128) に変化した場合の学習過程における予測誤差の推移を比較する実験も行う。

5.3 結果

5.3.1 定性評価

図 5 は、提案手法による画像生成結果を示している。最上段が予測対象の正解画像で、中段・下段が各学習ステップ数における提案手法の生成画像である。学習が進むにつれてより鮮

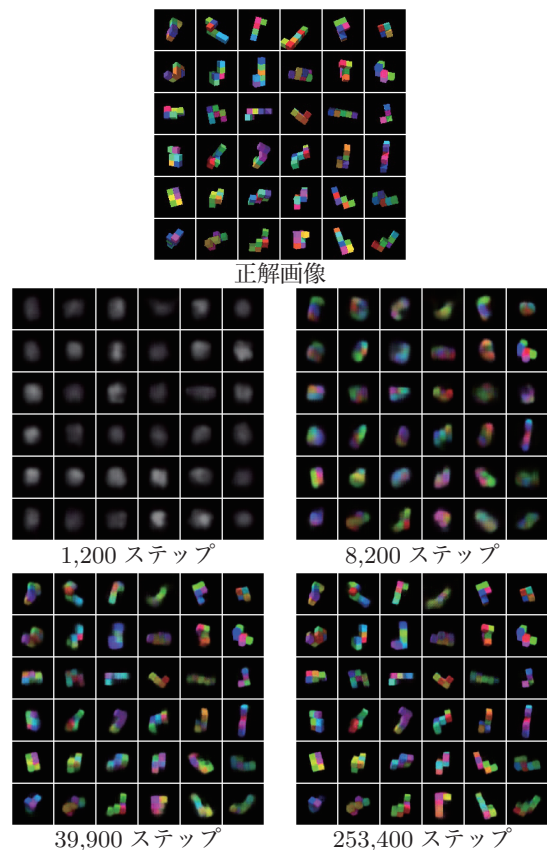


図 5: 提案手法による検証用データに対する画像生成結果

明な画像生成ができていることがわかり、253,400 ステップ後の生成画像は、正解画像と見比べても遜色のない画像が生成できていることが見て取れる。

5.3.2 定量評価 (精度・パラメータ数)

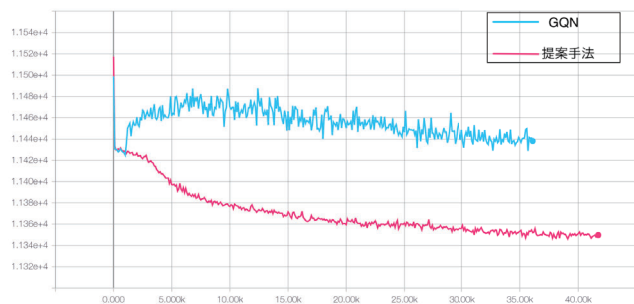


図 6: 検証用データに対する負の対数尤度の推移比較
横軸はステップ数、縦軸は予測誤差 (負の対数尤度) を表す。

図 6 は、提案手法と GQN の検証用データに対する予測誤差 (負の対数尤度) の推移を示している。青線が GQN、赤線が提案手法の値の推移を表している。図からも明らかに、GQN は目的関数に対数尤度に複数の近似を行なった変分下限を用いて、生成分布 $p(\mathbf{x}_i^q|\mathbf{v}_i^q, \mathbf{z}_i^q, \mathbf{r}_i; \theta)$ と 2 つの近似分布 $q(\mathbf{z}_i^q|\mathbf{x}_i^q, \mathbf{v}_i^q, \mathbf{r}_i; \phi)$, $q'(\mathbf{r}_i|D_i; \phi')$ の 3 つパラメータ θ, ϕ, ϕ' を同時に最適化しているために、予測誤差の推移が学習過程において不安定になってしまっている。一方で、提案手法では必要

最低限な推論分布 $q(z_i|D_i; \phi)$ の近似のみを行なっているために、安定して予測誤差が減少していることがわかる。

表 1 では、GQN と提案手法についての定量的な評価指標の値を比較している。負の対数尤度の項は両手法を 35,000 ステップ学習させたときの検証用データに対する値を表している。負の対数尤度は小さいほど予測誤差が小さいことを表すため、この指標でも提案手法は GQN よりも良い結果を残していることがわかる。

また、パラメータ数の項は、両手法の学習において最適化の対象となるニューラルネットワークのパラメータの総数を表している。提案手法のパラメータ数は GQN の 40%未満となっており、負の対数尤度の項と合わせて考えると、少ない計算リソースで GQN を上回る精度を実現できていることがわかる。これは、GQN の潜在変数 z の近似分布 $q(z_i^q|x_i^q, v_i^q, r_i; \phi)$ をモデル化するニューラルネットワークなどがなくなり、モデルの冗長性が排除されたことによる効果である。パラメータ数を減らすことの利点は、最適化が容易になる、計算機リソースが限られるような場合にもモデルの学習を行うことができる、などが挙げられる。特に GQN のような深層生成モデルの場合には、一般に大きなモデルアーキテクチャを使用することが多く、必然的にパラメータ数が増大しやすい傾向にあるため、少ないパラメータで良い精度が得られることは、利点であると言える。多くの場合、パラメータ数を削減することは、モデルの表現力を下げることに相当するため、精度とのトレードオフになるが、提案手法では、パラメータ数を削減した上で、GQN を上回る精度を達成しているため、この結果は提案手法の有効性を裏付けていると言える。

表 1: 手法ごとの定量評価指標 (パラメータ数, 尤度)

手法	パラメータ数	負の対数尤度
GQN	49,469,343	1.1439×10^4
提案手法	18,713,091	1.1353×10^4

5.3.3 ハイパーパラメータへの頑健性

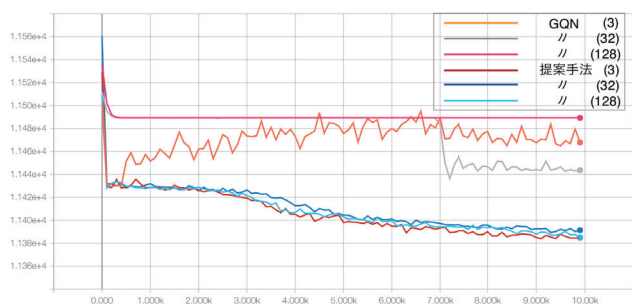


図 7: 潜在変数のチャンネル数を変化させた場合の検証用データに対する負の対数尤度の推移比較

横軸はステップ数、縦軸は予測誤差 (負の対数尤度) を表す。凡例の括弧内の数字は潜在変数のチャンネル数を示している。

最後に、ハイパーパラメータへの頑健性の検証実験の結果が図 7 である。GQN では、潜在変数のチャンネル数によって結果が大きく変化しており、特にチャンネル数が 128 の場合には、学習の初期段階で予測誤差の改善が止まり、学習に失敗していることがわかる。これは、GQN では、潜在変数 z_i^q のサ

ンプリングを近似分布 $q(z_i^q|x_i^q, v_i^q, r_i; \phi)$ から行なっており、パラメータの更新に用いる式 (5) の目的関数では、その近似分布を $p(x_i^q|v_i^q, z_i^q, r_i; \theta)$ に対して近づけるカルバックライバー距離の項が存在するため、 z_i^q の次元数 (チャンネル数) が大きくなった場合に、この項を過剰に最小化する局所解に陥り、学習がうまく進まなくなってしまうことが原因として考えられる。一方で、提案手法の目的関数には、潜在変数 z_i の確率分布間の距離を測る項が存在しないため、潜在変数の次元数の変化に対して頑健であり、実際にどのチャンネル数の場合にもうまく学習が進んでいることがわかる。

6. 考察

評価実験を通して、提案手法が GQN と比較して、予測精度、学習効率、ハイパーパラメータに対する頑健性において、優れた性能を発揮することが示された。これは、GQN の課題であった学習コストの高さや学習の不安定性を大きく改善していることを表している。

本研究では、シミュレート環境のみで実験を行なったが、今後はこれを実環境のデータセットを用いて検証していくことが必要である。また、本手法の 3 次元構造の把握能力を活用して、実ロボットによる複雑な物体の把持タスクの強化学習などに応用することも、今後の研究課題として検討できる。

7. 結論

本研究では、Generative Query Network をメタ学習の観点から検証し、その確率モデルを定式化することで、課題である学習コストの高さや学習の不安定性の原因を解明し、それらを改善する手法の提案を行なった。評価実験では、Shepard Metzler データセットを用いて、GQN と提案手法の予測精度、パラメータの総数、ハイパーパラメータへの頑健性を定量的に比較し、提案手法の有効性を示した。今後は、本手法を実環境のデータセットに対して適用する、実ロボットの強化学習に応用するなどの発展研究が期待される。

8. 謝辞

本研究は株式会社ドワンゴより提供の GPU サーバー「紅莉栖」を使用して行われました。

参考文献

- [Eslami 18] SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, Vol. 360, No. 6394, pp. 1204-1210, 2018.
- [Vilalta 2002] Ricardo Vilalta and Youssef Drissi. A perspective view and survey of meta-learning. *Artificial Intelligence Review*, Vol. 18, No. 2, pp. 77-95, 2002.
- [Gordon 2018] Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard E Turner. Meta-learning probabilistic inference for prediction. *arXiv preprint arXiv:1805.09921*, 2018.
- [GQN Dataset] <https://github.com/deepmind/gqn-datasets>