補助データを用いた VQ-VAE による運転シーン動画生成の精度向上

ER21092　寄田純平　　指導教授：藤吉 弘亘

1. はじめに

過去の動画から未来の動画を生成する動画生成モデルでは，計算コスト削減のため Vector Quantised-Variational AutoEncoder (VQ-VAE) [1] によって，入力データを離散的な潜在変数に変換する．そして，潜在変数から Transformer などの自己回帰モデルによって次のフレームの潜在変数を予測し，VQ-VAE のデコーダによって潜在変数からデータの再構成を行う．しかし，離散的な潜在変数の表現力には限界があり，再構成後の画像がぼやけたり，物体が背景と同化したりすることがある．そこで，本研究では，車載カメラデータにおける VQ-VAE の再構成の精度向上を目的とする．VQ-VAE の再構成時に物体が存在する座標を示す補助データを追加で入力することで，離散的な潜在変数への変換の際に低下した表現力を補完し，再構成の精度向上を図る．

2. VQ-VAE

VQ-VAE は，データを離散的な潜在変数に変換するオートエンコーダである．VQ-VAE は初めに，エンコーダにより入力データ $\boldsymbol{x}$ を低次元潜在変数 $\boldsymbol{z}_e(\boldsymbol{x})$ へ射影する．次に，潜在変数を事前に定義された離散的潜在変数であるコードブックと比較し，最も近いコードブックベクトル $\boldsymbol{e}$ に置換する．置換した潜在変数 $\boldsymbol{z}_q(x)$ から最後にデコーダにより入力データを再構成する．VQ-VAE の損失関数は以下のようになる．

$$\mathcal{L} = \|\boldsymbol{x} - D(\boldsymbol{e})\|_2^2 + \|sg\left[\boldsymbol{z}_e(\boldsymbol{x})\right] - \boldsymbol{e}\|_2^2 + \beta \left\|sg[\boldsymbol{e}] - \boldsymbol{z}_e(\boldsymbol{x})\right\|_2^2$$

ここで，sg は更新を停止することを示す．これにより，エンコーダとコードブックの更新が互いに干渉し合うことを防ぐ．第 1 項は入力と出力 $D(\boldsymbol{e})$ の再構成誤差を計算する．第 2 項，第 3 項は，潜在変数とそれに対応するコードブックベクトルとの誤差を計算し，それぞれコードブック，エンコーダのパラメータ更新を行う．

3. 提案手法

提案手法では，補助データの潜在変数を車載カメラデータの潜在変数に結合することで，再構成精度の向上を図る．補助データには，ロードマップ，バウンディングボックス，深度情報など，物体の座標を示す計測データを用いる．RADER，LIDAR によって得られた計測データを，カメラ座標系に基づき車載カメラの視点に対応するよう変換し抽出した特徴マップと，特徴マップを RGB で可視化したパースビューの動画を用いる．

車載カメラデータと補助データの潜在変数を結合し，デコーダで車載カメラデータを再構成する．提案手法の学習時の流れを図 1 に示す．学習は 2 段階で行う．1 段階目では補助データのエンコードを目的として，補助データを再構成するように VQ-VAE を学習する．2 段階目では，1 段階目で学習したエンコーダとコードブックを追加した VQ-VAE で，結合した潜在変数から車載カメラデータを復元するように学習する．

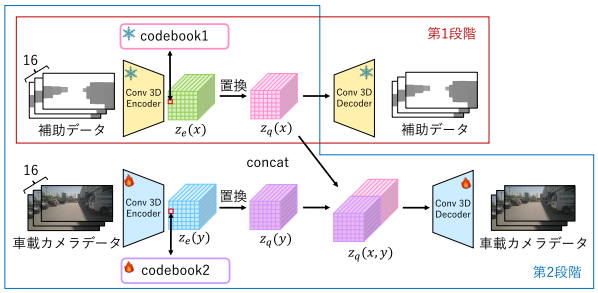

図 1：提案手法の概要

4. 評価実験

提案手法の有効性を示すために，補助データの有無による再構成精度の比較を行う．

4.1. データセット

本実験では車載カメラで撮影された動画とそれに対応する計測データを含む nuScenes データセット [2] を用いる．学習用データは 1,798 本，評価用データは 255 本，検証用データは 255 本とする．

4.2. 実験条件

学習時の設定はバッチサイズを 16，入力サイズを 128 × 128，再構成フレーム数を 16 とする．学習ステップ数は最大 10,000 とし，検証用データに対する loss が最小時のモデルを使用する．評価指標は，元データと生成データの平均二乗誤差 (MSE) と，輝度，構造，コントラストの差を基に算出した類似度 (SSIM) を用いる．

4.3. 定量的評価

提案した手法の再構成精度を定量的に評価する．比較結果を表 1 に示す．表 1 より，従来手法と比較して，パースビューを用いた場合は精度が低下している．一方で，計測データを用いた場合は従来手法よりも精度が向上していることが分かる．これより，計測データを補助データとして用いる提案手法の有効性を確認した．

表 1：定量的評価結果

	MSE ↓	SSIM ↑
従来手法	0.00259	0.870
補助データ：パースビュー	0.00264	0.840
補助データ：計測データ	**0.00251**	**0.871**

4.4. 考察

提案手法によるデータの再構成結果を図 2 に示す．ここで，左から車載カメラデータ，従来手法，パースビュー，計測データを用いた生成例を示す．計測データを用いた場合，視覚的な変化は見られなかったが定量的な精度は向上した．パースビューを用いた場合，変色やノイズが発生し定量的にも精度が低下した．そこで，第 1 段階目の学習をした VQ-VAE で，パースビューの再構成精度を MSE により評価した結果，計測データの再構成と比較して精度が低いことが判明した．パースビューは白色の背景が大半を占めるデータであり，第 1 段階目の学習の際に特徴をうまく捉えられず，適切な学習が行われなかったためであると考えられる．以上より，補助データには計測データを使用することが有効であるといえる．

図 2：定性的評価

5. おわりに

本研究では，VQ-VAE によるデータの再構成において，補助データを新たに入力し潜在変数を結合する手法を提案した．評価実験により，提案手法は従来手法より再構成精度が向上することを確認した．今後は，動画生成モデルに提案手法の VQ-VAE を組み込み，精度向上を目指す．

参考文献

[1] A. van den Oord *et al.*, “Neural Discrete Representation Learning”, NIPS, 2017.

[2] H. Caesar *et al.*, “nuScenes: A multimodal dataset for autonomous driving”, CVPR, 2020.