

# 深層学習を用いた動画像からの危険認知手法のための基礎的研究

11 月 10 日 (火)

小松 大起

## 1 本実験の目的

人間の認知は時間経過による視覚世界の変化の予測が可能である。従来の深層学習においては静止画の画像処理が中心であったが、本研究では動画像から未来フレームの画像生成へと拡張を行い、危険認知や行動予測などへの予測画像応用を行うための予測画像生成及び、その生成画像の評価方法を明らかにすることである。

## 2 PredNet

PredNet は Deep Recurrent Convolutional Neural Network の 1 種で 神経科学の概念である Predictive Coding を組み込んで作られたモデルである。2016 年に William Lotter, Gabriel Kreiman, David Cox の 3 氏によって公開された。

### 2.1 PredNet の層構造

PredNet の各層には 4 つの素子が存在しておりそれぞれ、Target, Representation, Prediction, Error と呼ぶ。Target は下層からの出力である誤差信号をエンコード、符号化する。Representation は Recurrent unit で、上層からの出力、側方からの誤差信号、1 ステップ前の自分の出力を受け取る。Representation unit は Target の予測をする Prediction unit に投射し、入力 of 予測が出力される。Error は Prediction と Target の誤差であり、Error は上層に送られる。この Error が小さくなるように学習を進めていく。また、層構造の例を図 2 に示す。

更新式は (1) から (4) のように表される。

## 3 使用した動画像

### 3.1 KITTI

ドイツの都市環境を運転している車の屋根に取り付けられたカメラによってキャプチャされたデータセットである。City, Residential, Road のカテゴリに分かれており、それぞれ都市街、住宅地、高速道路というようなシーン分けがされている。それぞれのカテゴリには City には 28 シーン、Residential には 21 シーン、Road には 12 シーン存在している。合計 61 のシーンがあるが、それぞれのカテゴリから 10 フレームがテストデータとしてサンプリングされ予測画像生成に用いられ、57 シーンが訓練に用いられ、4 シーンが検証に用いられる。また、用いられる画像は中央でトリミングされて、 $128 \times 160$  ピクセルの画像となっている。1 フレームは 0.1 秒である。最大 5 フレーム先の画像を生成することが可能。

## 4 先週までの作業

- 授業課題

## 5 今週の作業

- オートエンコーダにおいて白黒画像での画像生成を行なった。カラー画像での画像生成及び精度向上を目指す。それと中間層での特徴量を出力として扱えるようにする。現段階での生成画像の例を図 3 に示す。
- データセットを探してみる。ドライブレコーダの映像を使うことも考える。

## 6 来週以降の作業

- LSTM を作る.

$$A_l^t = \begin{cases} x_t & \text{if } l = 0 \\ \text{MAXPOOL}(\text{ReLU}(\text{CONV}(E_{l-1}^t))) & l > 0 \end{cases} \quad (1)$$

$$\hat{A}_l^t = \text{ReLU}(\text{CONV}(R_l^t)) \quad (2)$$

$$E_l^t = [\text{ReLU}(\hat{A}_l^t - A_l^t); \text{ReLU}(\hat{A}_l^t - A_l^t)] \quad (3)$$

$$R_l^t = \text{CONVLSTM}(E_{l-1}^{t-1}, R_{l-1}^{t-1}, R_{l+1}^t) \quad (4)$$

Fig. 2: PredNet の更新式

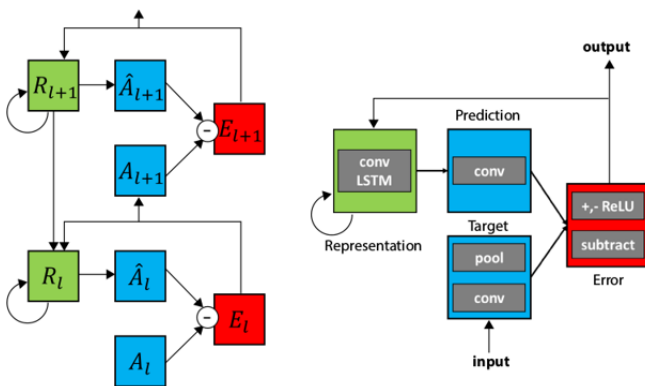


Fig. 1: PredNet の層構造の例

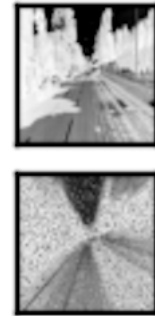


Fig. 3: autoencoder の出力結果