Paper summary: Learning actionable representations from visual observations

May 6, 2022

1 Idea in few sentances

2 Explanation of the central concept

3 Methodology

4 Initial rambly notes

4.1 Abstract

The latent space used is constructed using time-contrastive networks (TCN). The time-contrastive networks are extended in the following way: they take multiple frames as input - this allows for more accurate position and velocity attribute encoding. The embedding network is trained in a self-supervised approach. The overall training is a two-step process.

4.2 Introduction

The contributions are the following ones:

1. multi-frame TCNs work better than single-frame ones fro statisc and motion attribute classification
2. RL policies learned from pixels using mfTCN outperform learning from scratch or by using positional velocity encoders
3. policies trained in this way are competitive to those who had access to the true state space

4.3 Method

4.3.1 Base network

The base network is a CNN. The mfTCN embeddings are used on top of its representations. In math, the base network encodes the hidden features h_t at

time t from an input image I_t:

$$h_t = \text{CNN}(I_t) \tag{1}$$

4.3.2 Temporal aggregation

The available options to aggregate temporal information are: temporal averaging, temporal convolutions or RNNs. Here 3D temporal convolutions are used:

$$\phi_t = \text{Conv3D}(h_t, h_{t-s\ldots}, h_{t-(n-1)\times s}) \tag{2}$$

4.3.3 Dimensionality reduction

After temporal aggregation, the resulting data is in the form of a 4D convolutional feature map: (time, height, width, channels). This is compressed by using a fully connected layer:

$$\text{mfTCN}_t = \text{FC}(\phi_t) \tag{3}$$

4.3.4 Time-contrastive learning

n-pairs loss from K. Sohn "Improved deep metric learning with multi-class n-pair loss objective". is used as the loss function. Read the paper if you want to know more.

4.4 Other stuff