1 Motivations

1. Unsupervised learning from video: learning features, without explicit label, that generalizes to previously unseen tasks.
2. They learn a disentangled representation of frame in a video by factorizing the representation into two components, time-independent and time-varying. They name the two components the content and pose latent representation respectively.
3. Their architecture is generic. What's novel is the loss function and the simplicity of the approach.

2 Key components

2.1 Reconstruction Loss

1. Given a frame from the a video clip, 2 separate encoders are trained to output the content and pose latent representation.
2. They are trained so that the content encoding of the current frame and the pose encoding of future frame can be concatenated and input into a decoder to predict the pixel values of the future frame.
3. They argue that such reconstruction loss is not enough to force the pose representation to carry no content information.
4. They thus introduce an adversarial loss.

2.2 Adversarial Loss

1. They use an adversarial loss that forces the pose encoding of frames from different videos to be indiscriminable.
2. Concretely, a separate discriminator network is trained to predict if a pair of pose encodings comes from frames in the same video clip or different clips using a cross entropy loss.
3. The encoder used for reconstruction loss above is also trained to maximize the uncertainty of the discriminator when the discriminator predicts the pose encoding of frames from the same video clips.

2.3 Similarity Loss

1. They also argue that the content of a video clip changes slowly over time. So they enforce that the content encoding of nearby frame in a video be similar.

3 Experiment

3.1 Forward Prediction

1. Given a frame, they obtain the content and pose encoding of the frame.
2. Given the content and pose encoding, they learn a separate network (LSTM) to predict the pose encoding of future frames with a l2 loss.
3. The content encoding of the current frame and the predicted pose encoding of future frame are used as input into the trained decoder to predict pixel value of future frame. They argue that this works because the decoder was trained for reconstruction of future frame given the future frame pose encoding and the current frame content encoding.

3.2 Classification

1. They argue that the content representation is invariant to local temporal changes, so it is useful to classify the semantic content of an image.
2. They argue that the pose representation of a sequence of frame can be used to predict action sequences for video.

4 Question

1. Regarding the adversarial loss, as they explain, the adversarial loss associated with the encoder only tries to maximize the uncertainty in the prediction of the discriminator when the pair of encoding comes from frames in the same video clip. Wouldn't we want to encourage uncertainty even when the pair comes from frames in different video clip? After all, their motivation is that the pose encoding of frames should not contain information that can be used to tell which video clip the frames were from.

2. In the experiment section, they show that this approach works when there is only one moving object in the scene. Will this work when there are multiple objects and they might interact (i.e. collide)? What do we need to make it work in that scenarios?

5 Other interesting bits

1. They propose useful classifications of unsupervised learning methods:

2. Self-supervised learning, where domain knowledge is used to derive label from the label (predicting patches of objects)

3. Action-conditional predictive model of future frames, which relies on having knowledge of the action sequences.

4. Predictive auto-encoders, which is most general and only predicts the future from the present.