Chained Predictions Using Convolutional Neural Networks

Georgia Gkioxari — facebook
Alexander Toshev — Google
Navdeep Jaitly — Google

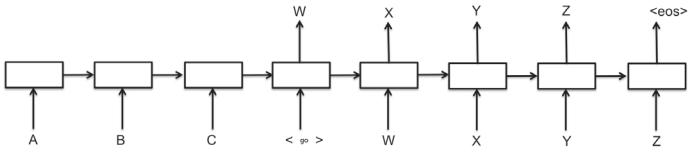
Motivation

- Visual recognition is a structured problem due to object-object, object-scene dependencies
- Given an input x and a set of tasks $Y=\{Y_t\}$, a chain model decomposes tasks sequentially

$$P(Y=y \mid x) = P(Y_0=y_0 \mid x) \cdot P(Y_1=y_1 \mid x, y_0) \cdot P(Y_2=y_2 \mid x, y_0, y_1) \ldots.$$

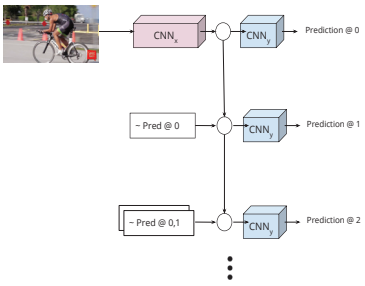
Chain Models in NLP

Seq2seq[1] models use the chain rule to translate a source sentence to a target sentence

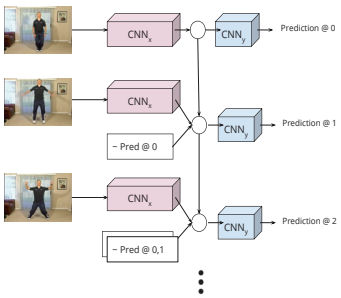


Chain Models in Vision

Single Image



Video



Single Image

- The sequence of tasks is defined by the marginal distribution
- Weights are not tied in order to allow semantic flow of information

Video

- The tasks are decomposed in time
- Weights are tied to enforce recurrence in time

Scheduled Sampling[2] is used to improve learning and bridge the gap between training and testing

[1] Sutskever, Vinyals & Le. Sequence to sequence learning with neural networks, NIPS 2014
[2] Bengio, Vinyals, Jaitly & Shazeer, Scheduled sampling for sequence prediction with recurrent neural networks, NIPS 2015

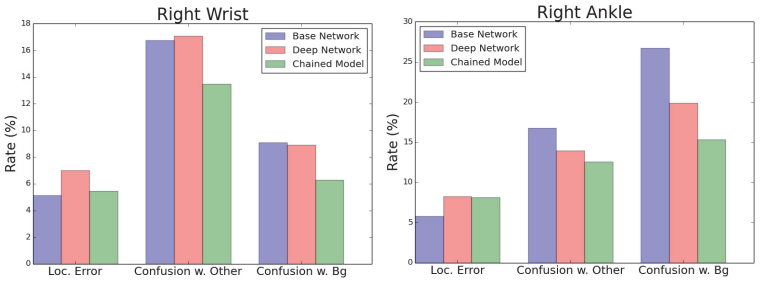
Chain Models for Pose Estimation

Single Image

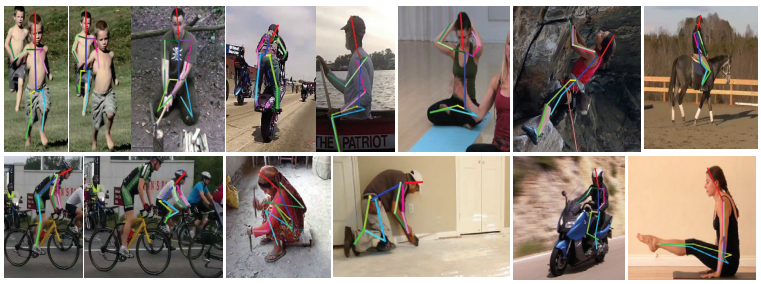
- The task is to localize each joint in an image
- The joints are decoded in descending order of accuracy: from easy to hard
- Dataset: MPII Human Pose dataset

PCKh (%)	Torso	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
Base Model	91.1	90.2	81.0	77.4	77.2	73.7	64.6	81.3
Chain Model	91.7	95.7	85.3	82.2	82.9	80.0	72.4	85.3

Error Analysis



Examples



Video

- The task is to track all joints across video frames
- The joints are decoded in space and time: from easy to hard, from older to newer
- Dataset: Penn Action dataset

PCK (%)	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
Base Model	94.1	90.3	84.2	83.5	88.7	87.2	87.7	87.5
Conv. RNN	95.3	92.5	87.9	87.5	91.1	89.8	90.1	90.1
Chain Model	95.6	93.8	90.4	90.7	91.8	90.8	91.5	91.8