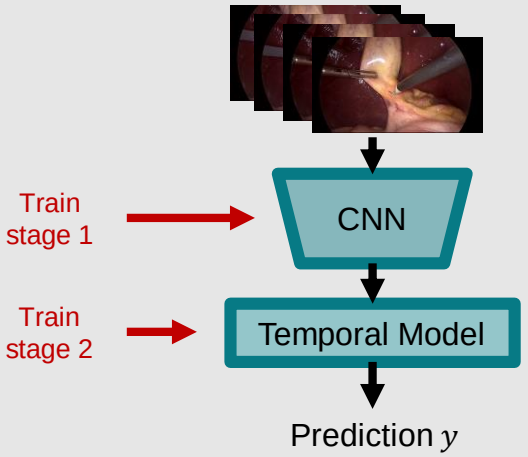
On the Pitfalls of Batch Normalization for End-to-End Video Learning: A Study on Surgical Workflow Analysis

Dominik Rivoir, Isabel Funke, Stefanie Speidel

Surgical Workflow Analysis

Common multi-stage training:



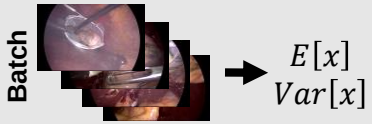
Why is end-to-end training not common?

One reason:
BatchNorm induces pitfalls for training CNNs on sequential video data!

Pitfalls of BatchNorm

Idea of BatchNorm:

1. Approximate feature distributions through training batches



2. Normalize features with approximated stats

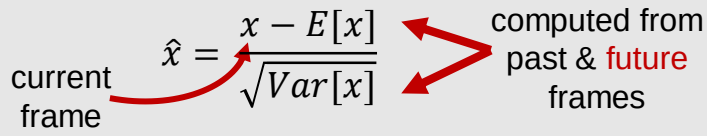
$$\hat{x} = \frac{x - E[x]}{\sqrt{Var[x]}}$$

Pitfall 1:

Video batches are poor estimates due to correlated samples (consecutive video frames)

Pitfall 2:

Leakage of future information through BatchNorm's dependency on other batch samples (including future frames)

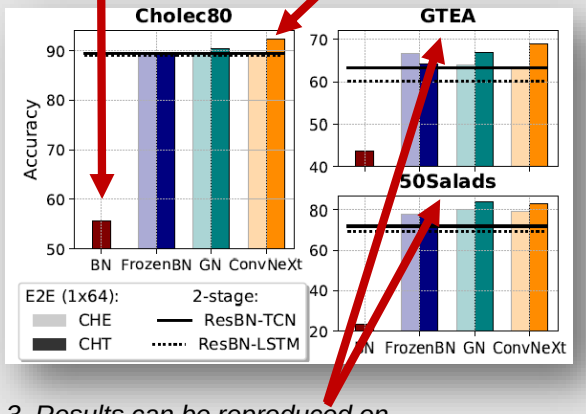


Analysis & Solution

We show:

1. BatchNorm models fail in end-to-end settings

2. Simple, end-to-end CNN-LSTMs w/o BatchNorm beat multi-stage methods and achieve SOTA results in phase recognition and instrument anticipation



3. Results can be reproduced on non-surgical video tasks

CeTI NCT