M2CAI Workflow Challenge: Convolutional Neural Networks with Time Smoothing and Hidden Markov Model for Video Frames Classification

Remi Cadene, Thomas RobertNicolas THOME, Matthieu CORD

Sorbonne Universités, UPMC Univ Paris 06, LIP6, Paris, France

Context

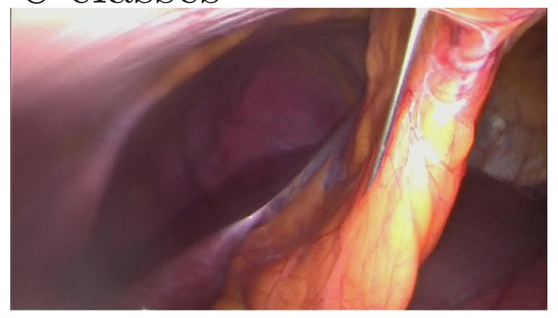
Goal: Surgical video frames classification

▷ Videos of size 1920x1080 Shot at 25 frames per second at IRCAD research center in Strasbourg, France

▷ 27 training videos

▷ 15 testing videos

▷ 8 classes



OK for centered object *KO for "natural" image*

▷ Online prediction: $P(y|x_i, x_{i-1}, x_{i-2}, \ldots)$

▷ Usefull to

▷ Monitor surgeons

▷ Trigger automatic actions

Our approach (usable in production)

▷ Firstly, we train a classifier using Deep Learning

▷ Secondly, we smooth its predictions

▷ Averaging predictions over last 15 frames

▷ Then, we use a Hidden Markov Model (HMM) as a "denoizer"

Frames classifier

Creating validation set by random split:

▷ Training set: 22 videos ▷ Validation set: 5 videos $\{2, 9, 10, 13, 27\}$

Extracting one frame every 25 frames (1 f/s):

▷ Training set: 59,493 images

▷ Validation set: 8,062 images

▷ Testing set: 28,732 images

3 approaches using Deep Learning:

▷ Training CNN From Scratch

▷ Good: learning representations for this task

▷ Bad: maybe not enough data + lot of hyperparameters

▷ Pre-trained CNN as Features Extractor + SVM

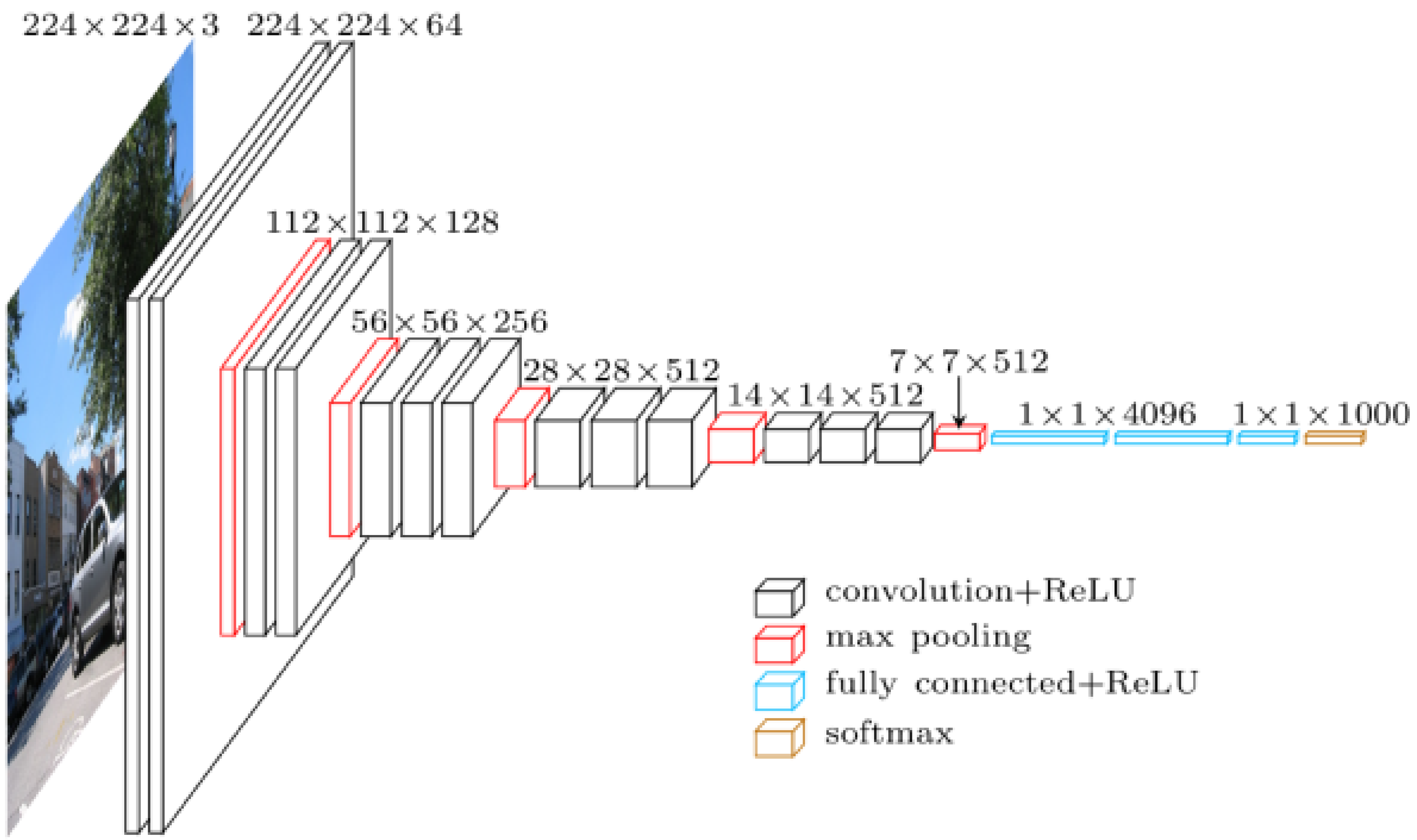
▷ Good: using (good) representations learned on ImageNet

▷ Bad: large semantic gab between ImageNet and this task

▷ Fine Tuning a pre-trained CNN

▷ Good: adapting the representations learned on ImageNet for this task

Deep learning architecture and methods



▷ Fully connected layer → convolution layer

▷ Sliding window approach / shared features

▷ Spatial aggregation

▷ Object localization prediction

Experiments

Model	Type	Accuracy (%)
InceptionV3	Extraction (repres. of ImageNet)	60.53
InceptionV3	From Scratch (repres. of M2CAI)	69.13
InceptionV3	Fine-tuning (both representations)	79.06
ResNet200	**Fine-tuning (both representations)**	**79.24**

Temporal Method	Accuracy Val (%)	Jaccard Val	Jac
Avg Smoothing	85.97	74.67	
Avg + HMM Online	**88.90**	**81.60**	
Avg + HMM Offline	93.47	87.59	

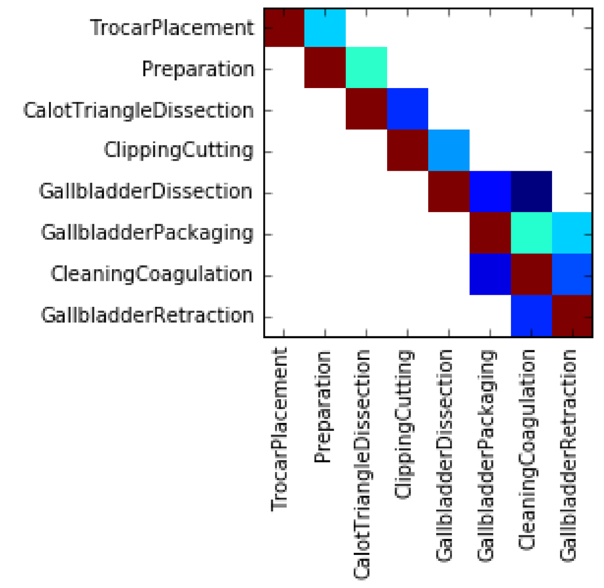
[1] Oquab et al. Is object localization for free? *CVPR*, 2015.
[2] Durand et al. MANTRA. *ICCV*, 2015.
[3] Parizi et al. Automatic discovery of parts. *ICLR*, 2015.
[4] Gong et al. Multi-scale orderless pooling. *ECCV*, 2014.

Hidden Markov Model

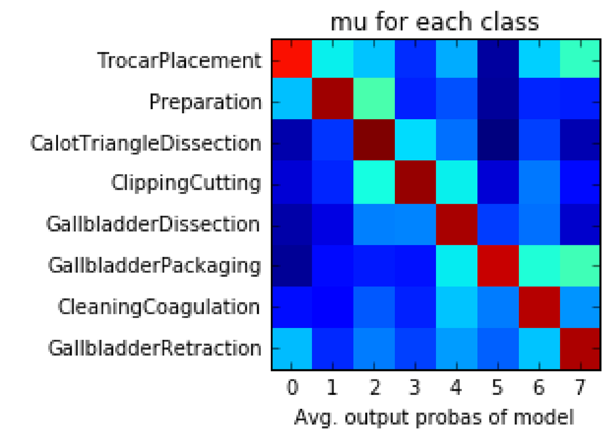
▷ Initial state probabilities

▷ Matrix of probabilities of transition between states

▷ Gaussian parameters for emissions of observations (mean and co-variance matrix)

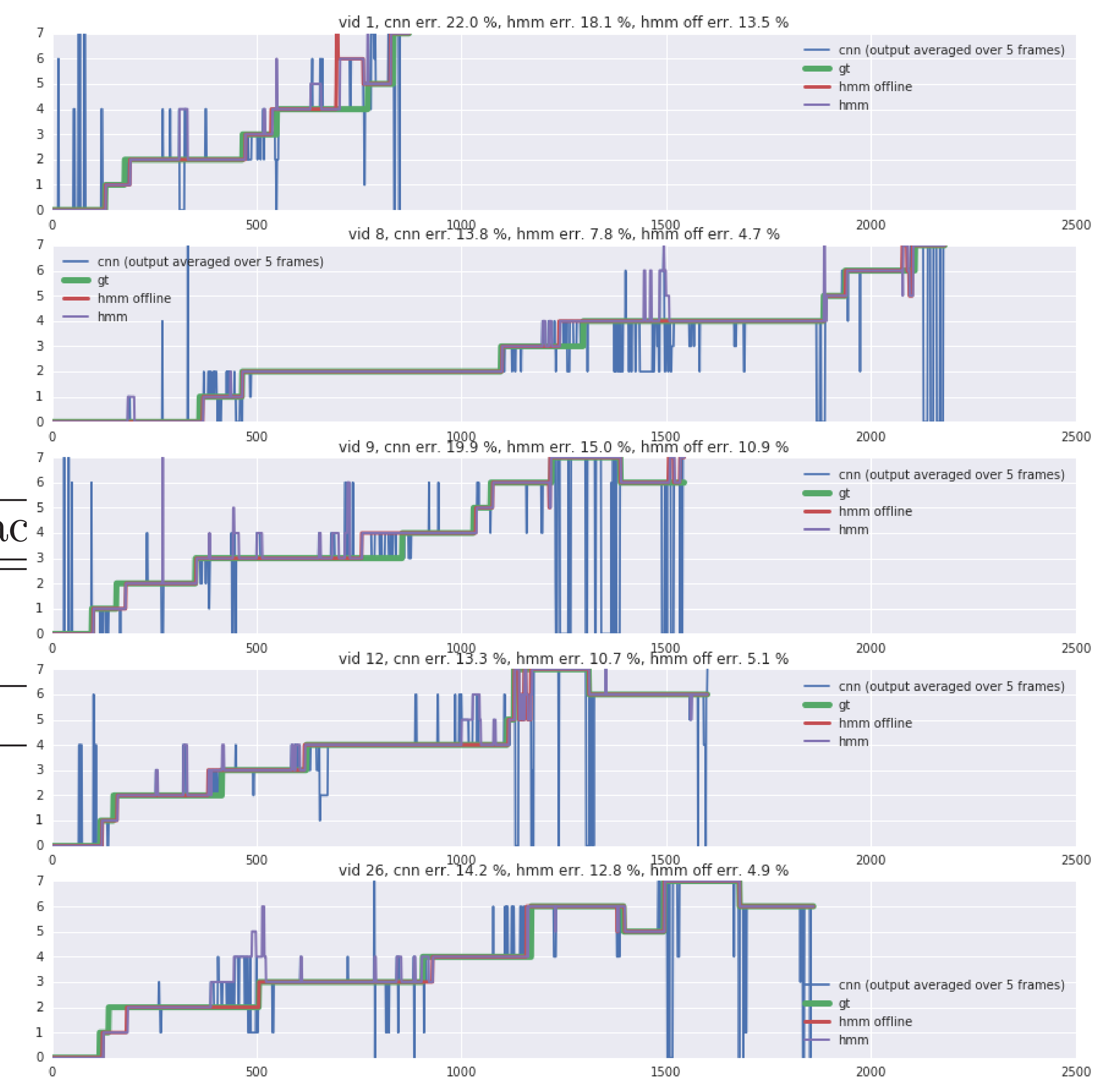


(b)



(c mean)

Visual results



Conclusion

▷ Fine Tuning most accurate approach ▷ HMM is usefull to smooth the predictions ▷ Production ready !

▷ Future Work: ▷ Fine Tuning CNN on full trainset ▷ Ensembling several fine tuned CNNs