Sports Classification – Temporal vs. Static

Dominique Cheray & Manuel Krämer

Motivation

- Classification is an important task in searching and summarization
- Most classification tasks don't include sports → How do common networks perform on this task?
- What are the networks looking at?
- Sport contains lots of movements → Is temporal information important for classification?

Data Set

- Subset of MPII Human Pose Dataset → 10 sports: Basketball, Horseback riding, Martial Arts, Paddleball, Rock climbing, Rope skipping, Skateboarding, Softball, Tennis, Golf
- 1576 images total, 1266 training images, 310 test images

Samples of the dataset – one row is one class

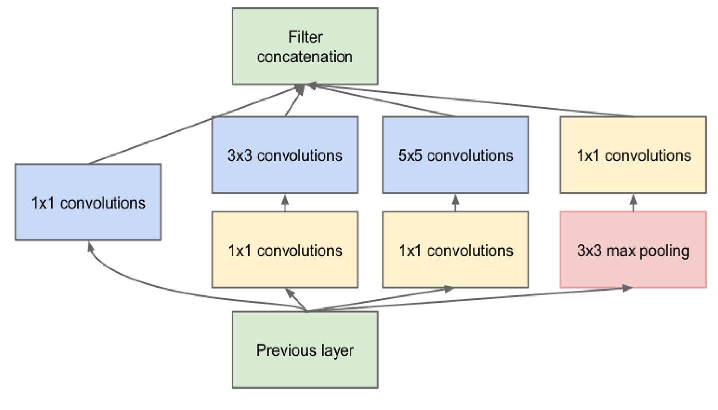
References

- Andriluka, M., Pishchulin, L., Gehler, P. and Schiele, B., 2014. 2d human pose estimation: New benchmark and state of the art analysis. In Proceedings of the IEEE Conference on computer Vision and Pattern Recognition (pp. 3686-3693).
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778)
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. and Torralba, A., 2016. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2921-2929).

Materials and Methods

GoogLeNet

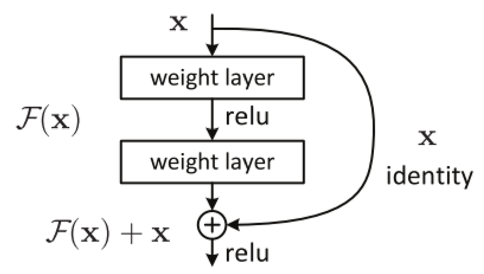
- 27 layers deep network
- 9 Inception modules → reduce the number of parameters, create a deeper and wider topology



Inception module

ResNet

- 34 layers deep network
- Skip connections over residual blocks → more gradient flows backwards



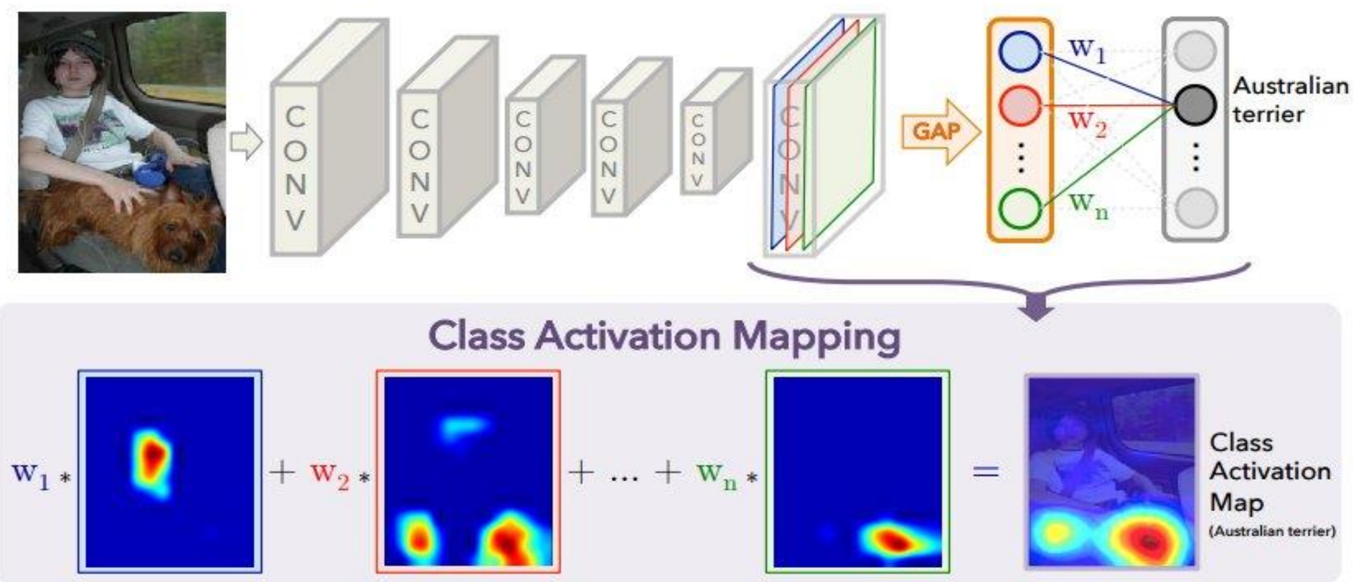
Residual block

Training and Testing

- Softmax Loss as the classifier
- Trained for 200 Epochs using SGD with 0.9 Momentum, 0.001 Learning Rate and a fixed Learning Rate schedule (decrease LR by 4% every 8 epochs)
- Split training images into training and validation set
- Performed data augmentation on the training images
- Final testing was done on the test images

Class Activation Mapping

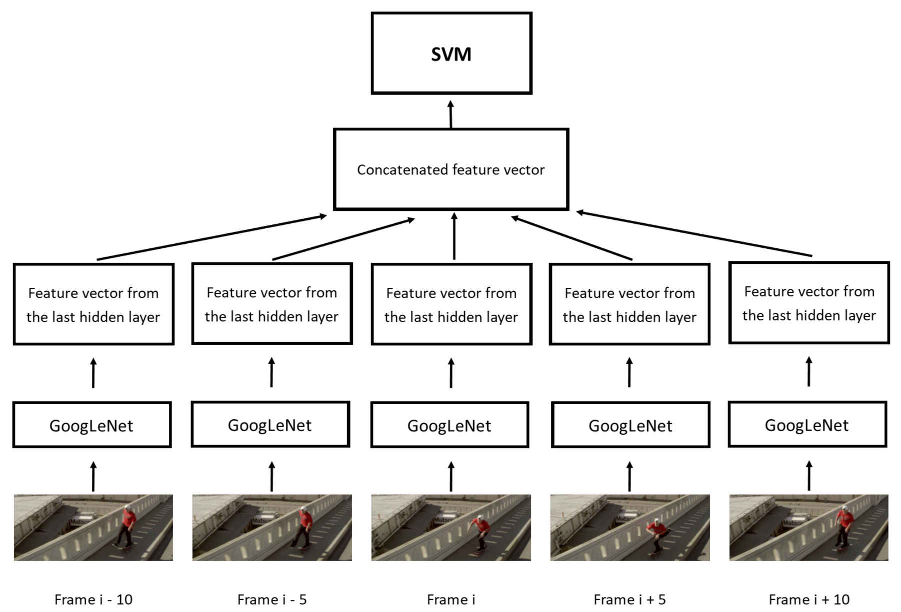
- Indicates the discriminative image regions used by the CNN to identify that class



Class Activation Mapping

Temporal Analysis with SVM

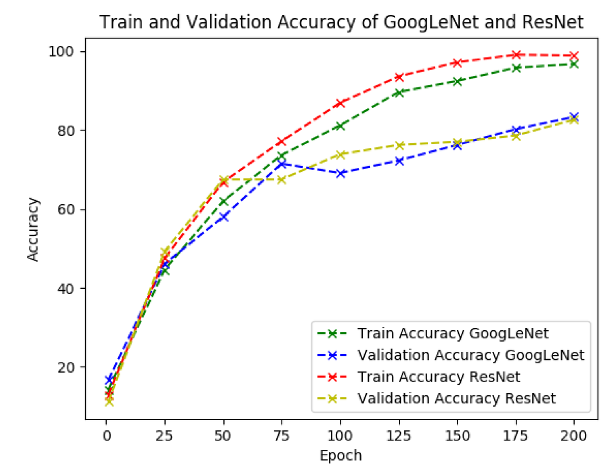
- We use 3, 5, 7 or 9 successive frames with a distance of 5 frames (For 9 frames, 4 frames distance is used)
- Every frame gets processed through the GoogLeNet until the last hidden layer
- The output is a feature vector
- All of them get concatenated and the final array is one instance for the SVM



Results

Classification with the Networks

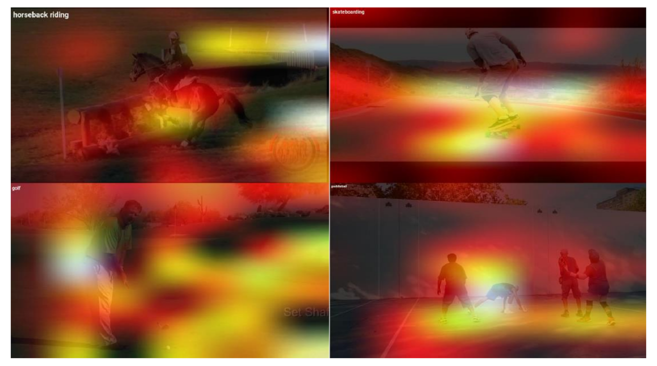
- Accuracy



	GoogLeNet	ResNet
Testing Accuracy	82.3%	76.1%

- Class Activation Mapping

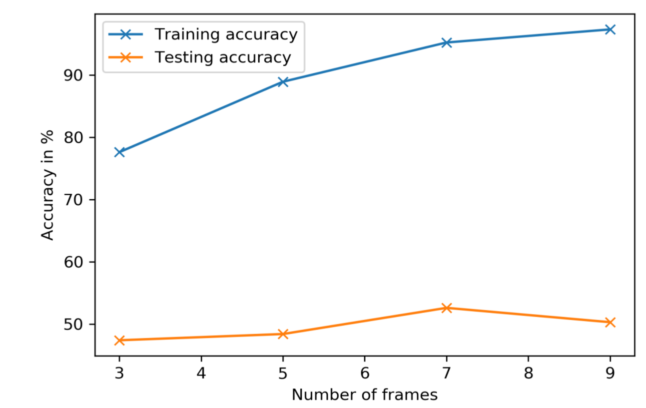
Four CAM examples for GoogLeNet



Four CAM examples for ResNet

Temporal Analysis with SVM

	Training Accuracy	Testing accuracy
3 frames with spacing 5	77.6%	47.4%
5 frames with spacing 5	88.9%	48.4%
7 frames with spacing 5	95.2%	52.6%
9 frames with spacing 4	97.3%	50.3%



- The accuracy is increasing with more temporal information
- Too many frames could involve a cut in the video
  → Worse testing accuracy