Learnable pooling with Context Gating for video classification

Antoine Miech[1,2] Ivan Laptev[1,2] Josef Sivic[1,2,3]

[1]DI ENS, École Normale Supérieure, PSL Research University [2]Inria [3]CIIRC, CTU in Prague

Ranked 1 at kaggle

Release

LOUPE Tensorflow toolbox: github.com/antoine77340/LOUPE
arXiv: 1706.06905

Goal

- Multi-label video tagging

Groundtruth: Tree- Christmas Tree - Christmas Decoration - Christmas

Challenges

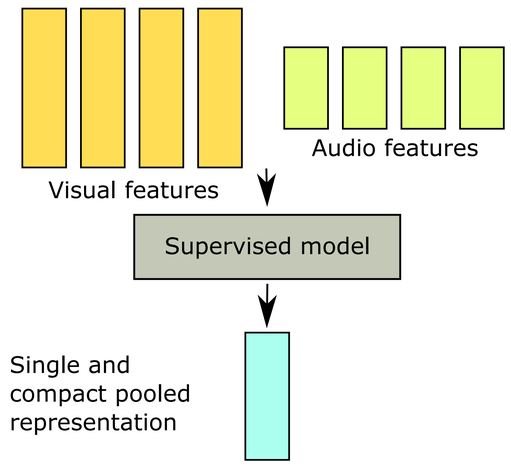
- Large diversity of labels
- Incomplete and noisy annotation
- What is a good representation for video ?

Contributions

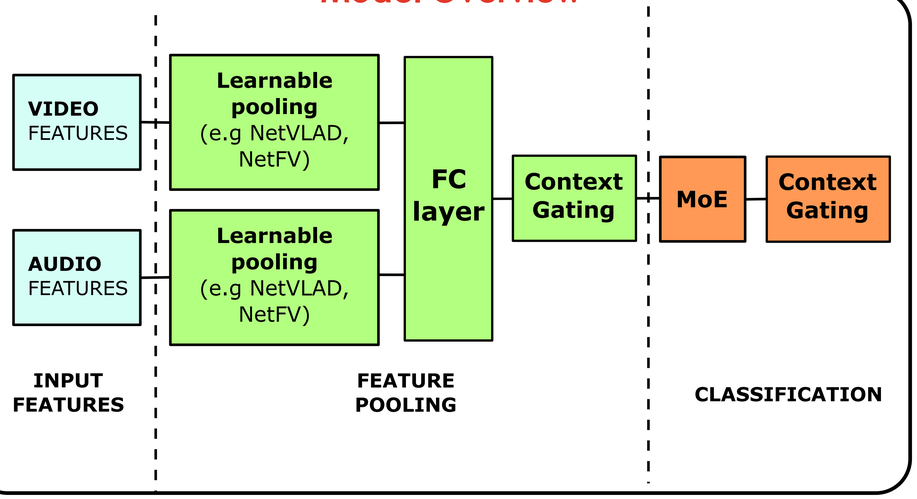
- Learnable pooling for video representations
- Context Gating: A non-linear learnable module for modeling interdependencies between activations

Focus on aggregation

- How to aggregate multi-modal sequences of features into a good compact representation ?



Model Overview



Learnable pooling via Clustering

- **Soft-DBoW**
Soft Bag-of-Words [2] formulated in terms of differentiable operations.

- **NetVLAD [1] / NetRVLAD**
NetVLAD: Approximates a VLAD [4] encoding with differentiable operations.
NetRVLAD: Residual-less NetVLAD.

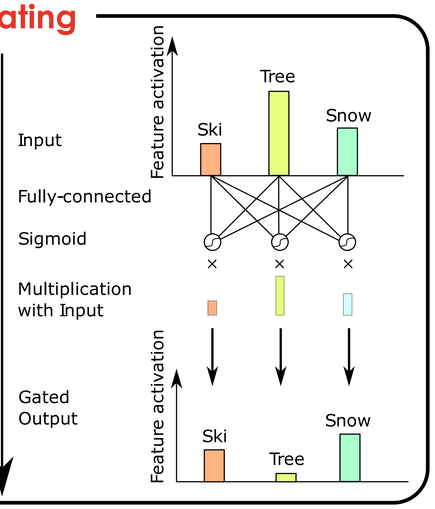
- **NetFV**
A variant of the Fisher Vector representation [5] with an approximation of Gaussian Mixture model by differentiable operations.
This is done through a simple extension of NetVLAD [1].

Context Gating

$$Y = \sigma(WX + b) \circ X$$

X: input layer
Y: output layer
W,b: parameters to learn

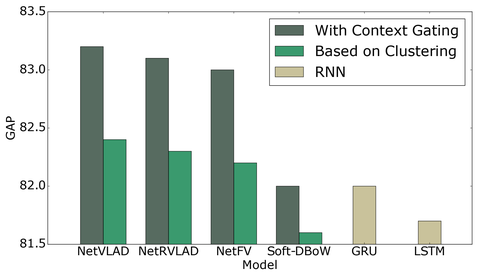
- Models interdependencies between activations with a self-gating mechanism
- Adds quadratic interactions
- Inspired by Gated Linear Unit [3]



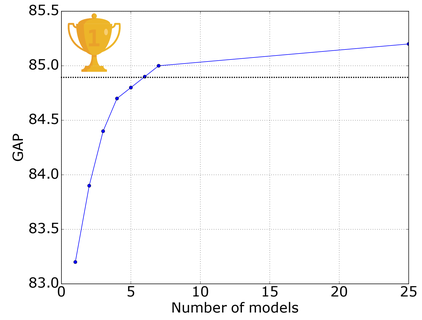
Results on the Youtube-8M dataset

+7 Millions videos	4716 labels	3.4 avg labels/video	450000 video hours

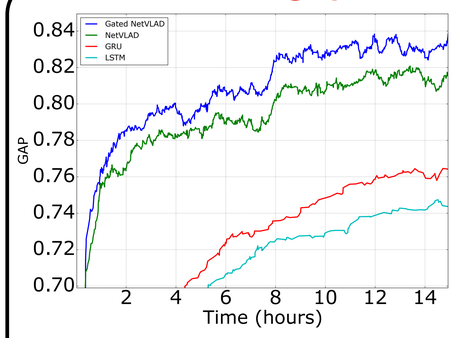
- Comparison of different pooling based methods (with or without Context Gating) and the RNN pooling approaches.



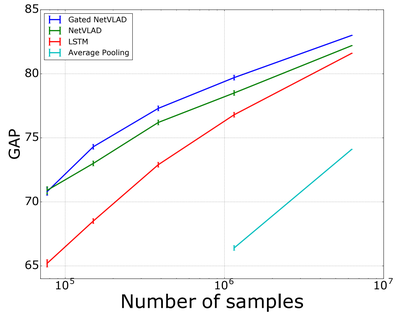
- The effect of ensembling different methods.
A prediction score average of 7 different models can achieve the first place of the kaggle Youtube-8M challenge out of 650 teams.
Models are: Gated NetVLAD, Gated NetRVLAD, Gated NetFV, Gated Soft DBoW, DBoW, GRU and LSTM.



Training speed and Generalization



- Training time vs. performance for LSTM, GRU, NetVLAD and Gated NetVLAD.



- Performance as a function of the training set size.

References

[1] R. Arandjelović, et al. NetVLAD: CNN architecture for weakly supervised place recognition. CVPR 2016
[2] J. Sivic et al. Video Google: a text retrieval approach to object matching in videos. ICCV 2003
[3] Y. N. Dauphin, et al. Language modeling with gated convolutional networks. arXiv:1612.08083
[4] H. Jégou, et al. Aggregating local descriptors into a compact image representation. CVPR 2010
[5] F. Perronnin, et al. Fisher kernels on visual vocabularies for image categorization. CVPR 2007