Learning Smooth Pooling Regions for Visual Recognition
Mateusz Malinowski and Mario Fritz

Motivation
- State-of-the-art object recognition algorithms are based on histograms of feature representations
- Spatial Pooling, in order to preserve some spatial information, aggregates statistics locally
- Current Spatial Pooling schemes are hand-crafted (e.g. SPM)

- Are such spatial regions optimal?
- Can we train jointly both the classifier and spatial regions?
- What assumptions on the Spatial Pooling scheme are needed to achieve best performance?

Our method
- Parameterized pooling operator
- Joint training of classifier and pooling regions
- Efficient and parallel approximation training
- Logistic regression as a classifier
- Our optimization problem:

\[ \text{minimize } \mathbf{W}, \mathbf{\Theta} \quad J(\mathbf{\Theta}) + \frac{\alpha_1}{2} \| \mathbf{\Theta} \|_2^2 + \frac{\alpha_2}{2} \| \mathbf{W} \|_2^2 + \frac{\alpha_3}{2} \left( \| \nabla_x \mathbf{W} \|_2^2 + \| \nabla_y \mathbf{W} \|_2^2 \right) \]

subject to \( \mathbf{W} \in [0,1]^{K \times M \times L} \)

Where \( J(\mathbf{\Theta}) := -\frac{1}{D} \sum_{d=1}^{D} \sum_{i=1}^{L} \mathbf{1}\{y^{(i)} = j\} \log p(y^{(i)} = j | \mathbf{\alpha}^{(1,2)}, \mathbf{\Theta}) \) and \( \mathbf{\alpha}^l := \left\{ \sum_{j=1}^{M} \mathbf{w}^l_j \circ \mathbf{u}_j \right\} \)

Results
- Evaluation on Object and Event recognition tasks
- Hand-crafted Spatial Pooling as a baseline\(^4\,5\)
- Strong improvement over hand-crafted Spatial Pooling\(^4\,5\)
  - 3% on Event and up to 10% on Object recognition
- State-of-the-Art on CIFAR-100 given SPM

Conclusion
- Importance of learnt Spatial Pooling regions
- Scalable algorithm for larger dictionaries
- Discovery of new pooling schemes
- Importance of Spatial Smoothness prior
- Applicable to sum- and max-pooling

---