What have we learned from deep representations for action recognition?

Christoph Feichtenhofer, Axel Pinz, Richard P. Wildes, Andrew Zisserman
CVPR 2018
In summary

• First to visualize **hierarchical features** in deep **motion** network
• **Activity maximization**
• Spatial and temporal **regularization**
• Both specific and generic **representations**
Interpretability timeline

**Erhan 2009**
Introduces activity maximization

**Zeiler 2013**
Uses deconvnets (their invention from 2010) to project the feature activations back to pixel space

**Szegedy 2013**
1. Questions the semantic interp. of single units "grandmother cells", claims it is rather a distributed code
2. First to look at (and coins) adversarial examples

**Agrawal 2014**
Agrees with the distributed code argument and presents more experiments to support this

**Simonyan 2014**
1. Applies act. max. on a supervised convnet model
2. Computes saliency maps using backprop
3. Shows that such gradient-based vis. methods generalize deconvolution reconstructions
## Interpretability timeline

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhou 2015</td>
<td>Introduces class activation mapping (CAM)</td>
</tr>
<tr>
<td>Mahendran 2016</td>
<td>Another questioning of the grandmother cells. This time supported by the notion of the information bottleneck.</td>
</tr>
</tbody>
</table>
| Selvaraju 2017 | Grad-CAM  
More general solution than CAM                               |
| Selvaraju 2018 | Grad-CAM++  
More robust to when there are multiple instances to classify  
Applied to video               |
| Feichtenhofer 2018 | Focuses on spatiotemporal visualizations (for 2-stream models), the approach is without given input (activity maximization)  
- Adds regularization  
+ Have moving video in their pdf paper 🙌  
- Ignores no-GMC papers |
Fig. 4: *Lack of neuron selectivity.* The bottleneck information $r$ is fixed to the one computed during the forward pass $\phi(x)$ through AlexNet and the output of $\phi^+(e, r)$ is computed by choosing $e$ as: the most active neuron (top row), a second neuron at random (middle), or as a positive random mixture of all neurons (bottom row). Results barely differ, particularly for the deeper layers. See figure 1 for the original house input image $x$. Best viewed on screen.
Approach – Activity maximization

\[ x^* = \underset{x}{\arg \max} \frac{1}{\rho^2 \tilde{a}_{t,c}} (a_t(x), e_c) - \lambda_r \mathcal{R}_r(x) \]
Approach – Regularizing local energy

\[ R_B(x) = \begin{cases} N_B(x) & \forall i, j, k : \sqrt{\sum_d x(i, j, k, d)^2} \leq B \\ +\infty & \text{otherwise.} \end{cases} \]  

\[ N_B(x) = \sum_{i,j} \left( \sum_d x(i, j, k, d)^2 \right)^{\frac{\alpha}{2}} \]
Approach – Regularizing local frequency

By varying $\chi$ and $\kappa$ we can have 3 different cases:

1. A purely spatial regularizer ($\kappa > 0; \chi = 0$)
2. An isotropic spatiotemporal regularizer ($\kappa = \chi; \chi > 0$)
3. An anisotropic spatiotemporal regularizer ($\kappa \neq \chi; \kappa, \chi > 0$)
Experiments – “Class-specific” units
Experiments – General units
Experiments – Progressive feature abstraction with depth

**Early layers:**

- Spatial patterns preserved regardless of Chi
- Speed invariance

**Fusion layers:**

Same results as earlier, matching to classes by inspection
Experiments – Progressive feature abstraction with depth

Global layers:

• Matching by inspection, varying Chi

Class output layer:

• Know what the units should correspond to
• Motion is striking, appearance not as specific (faces, barbells)
Claims to discuss

• “Our visual explanations provide qualitative support for the benefits of separating into two pathways when processing spatiotemporal information”

• “At conv5_fusion we see the emergence of both class specific and class agnostic units”

More?