Visual Attention in Human Action Recognition

Georg Frey
Darmstadt University of Applied Science

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

ORDIX AG

Motivation

Since their inception, major challenges when working with artificial neural networks have been efficacy, efficiency and explainability. All of these problems scale in one way or another with the size of an artificial neural network. Network performance typically increases with network depth, which in turn makes the efficiency a larger concern. Furthermore, networks with a large amount of parameters may also be difficult to diagnose.
These problems are especially evident when working with video data, where network sizes are inherently large due.
A promising concept addressing the efficacy and efficiency of an artificial neural network is attention. Additionally, visual attention may also help understanding and diagnosing a trained model.

Attention

Similar to artificial neural networks themselves, attention is also inspired by nature. As a human would for example pay attention to an object in a picture, attention mechanisms aim to help a model focus on certain parts of the data.
This focus typically happens by highlighting interesting parts of the data or masking background information or noise.
The most fundamental form of attention is a function with three arguments:

- Value - V : The data or features to be attended
- Key - K : Key-pairings for the values
- Query - q : Query data determining what to retrieve

$$A(q, K, V) = \sum_i p(a(k_i, q)) * v_i$$

Here, a is a compatibility function, measuring the response of the key to the query, and p is a scaling function like softmax. As a basic compatibility function the simple dot-product can be used. Importantly, q, K and V do not have to be distinct values and can all stem from the same data (self-attention).

Attention Categories

According to the taxonomy proposed by Chaudhari et al. [1] attention mechanisms can be categorized with regard to four properties:

Category	Type
Number of Sequences	distinctive, co-attention, self
Number of Abstraction Levels	single-level, multi-level
Number of Positions	soft/global, hard, local
Number of Representations	multi-representational, multidimensional

Importantly, these category are not mutually exclusive and attention mechanisms will typically fall into multiple categories. For this purpose of this work, these types are especially relevant:

- Self-attention: The key and query stem from the same data.
- Multi-level-attention: The attention is applied at multiple stages of the artificial neural network, typically as a module.
- Soft attention: The attention is applied globally as a weight over the entirety of the value data (“attention mask”).

Video Classification Architectures

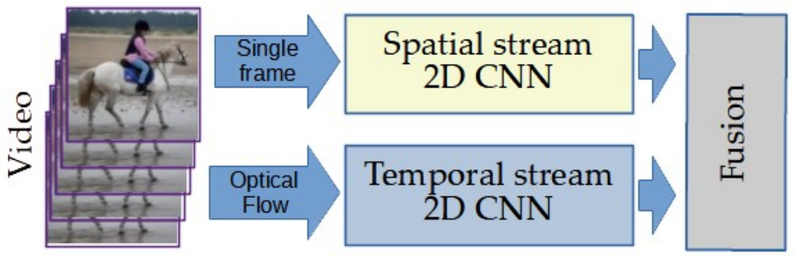
When approaching video classification with artificial neural networks there are a number of naive network architectures:

- 2D-CNN + Fusion (i.e. by averaging class scores)
- 2D-CNN + LSTM
- 3D-CNN

Typically, current state-of-the-art models implement one of the following architecture archetypes:

- Two-Stream Networks
- (2+1)D CNN
- Transformer

For this implementation a two-stream model is adapted for the use with visual attention. This architecture was chosen because its use of a 2D-CNN as backbone architecture makes highly compatible with many visual attention mechanisms.

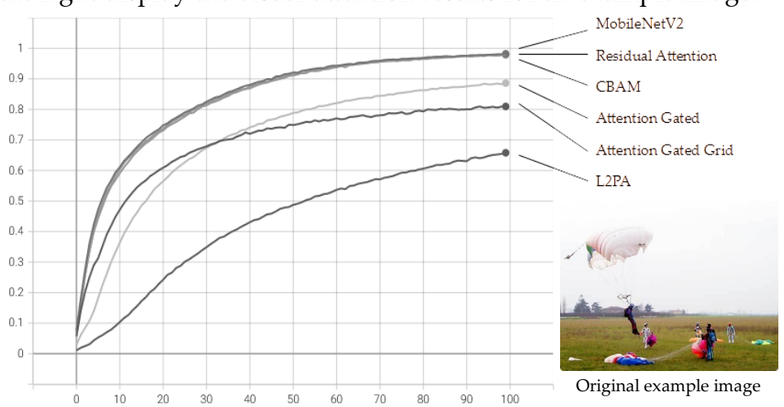


Implementation

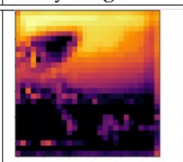
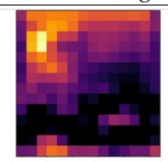
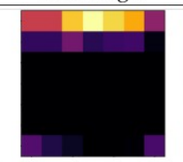
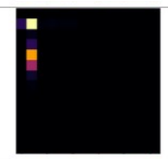
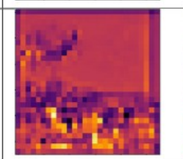
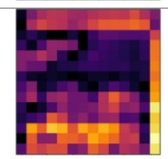
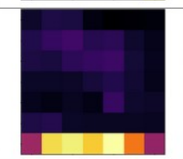
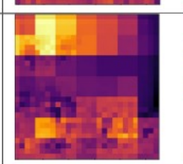
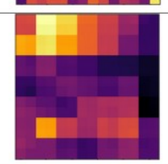
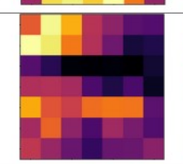
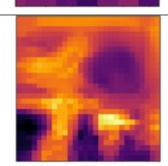
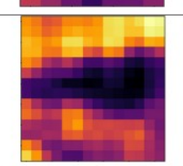
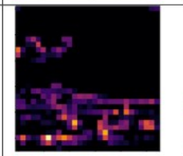
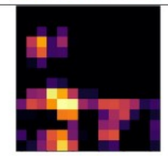
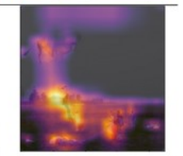
Four attention mechanisms were chosen for implementation.

- Learn to Pay Attention (L2PA) [2]
- Attention-gated Networks [3]
- Residual Attention Networks [4]
- Convolutional Block Attention Module (CBAM) [5]

Each mechanism is modular and was inserted into three different stages of the MobileNetV2 backbone architecture.
With the exception of CBAM, all mechanisms produce a visual attention mask, which can be extracted and visualized as a heatmap. To visualize the behavior of the unmodified network and the CBAM-version, Gradient-weighted Class Activation Mapping (Grad-CAM) [6] was used.
The Following curves show the results when training MobileNetV2 adaptions on the ImageNette dataset. The table to the right display the visual attention results for an example image.



Original example image

Model	Attention: early stage	medium stage	late stage	combined
MobileNetV2 (Grad-CAM)				
MobileNetV2 + L2PA				
MobileNetV2 + Gated Attention				
MobileNetV2 + Gated Grid Attention				
MobileNetV2 + Residual Attention				
MobileNetV2 + CBAM (Grad-CAM)				

Findings & Conclusion

Although two-stream architectures are especially suitable, adapting visual attention mechanisms for video models faces many challenges like compatibility problems with 3-dimensional convolutions, other architecture conflicts and incompatibility with techniques like inflation or general.
Furthermore, post-hoc attention algorithms like Grad-CAM represent an alternative which does not require retraining, have fewer compatibility issues and can more flexibly applied to any layer.

References

[1] Sneha Chaudhari, Varun Mithal, Gungor Polatkan, and Rohan Ramanath. “Anattentive survey of attention models.” In:ACM Transactions on Intelligent Systemsand Technology (TIST)12.5(2021), pp.1–32.
[2] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. “Learn to pay attention.” In: arXiv preprint arXiv:1804.02391 (2018)
[3] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. “Attention gated networks: Learning to leverage salient regions in medical images.” In: Medical image analysis 53 (2019), pp. 197–207.
[4] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. “Residual attention network for image classification.” In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, pp. 3156–3164.
[5] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. “Cbam: Convolutional block attention module.” In: Proceedings of the European conference on computer vision (ECCV). 2018, pp. 3–19.
[6] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. “Grad-cam: Visual explanations from deep networks via gradient-based localization.” In: Proceedings of the IEEE international conference on computer vision. 2017, pp. 618–626.