Visual Attention in CNNs

Mario Teixeira Parente (m.parente@fz-juelich.de)

ISIS LENS Machine Learning School

February 19, 2021

MLZ is a cooperation between

JÜLICH FORSCHUNGSZENTRUM

Helmholtz-Zentrum Geesthacht Zentrum für Material- und Küstenforschung

Technische Universität München

Idea

- When training a CNN, we would like to be able to focus on "important" parts of the image.
- This can be achieved with *trainable attention* mechanisms.

Trainable attention

Def.: Trainable attention is a set of tools that help a "model-in-training" notice important things more efficiently.

- It is trained *while* the network is trained and is supposed to help the network to focus on key elements of the image.

Post-hoc attention

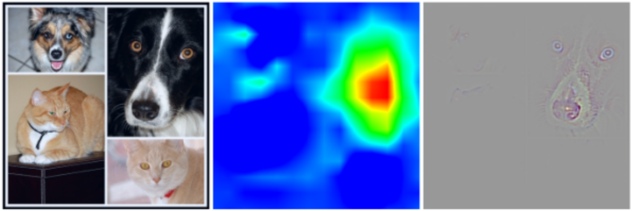
Def.: Post-hoc attention is a group of techniques that help humans visualize what an *already-trained* model thinks is important.

- It is not intended to change the way the model learns, or to change what the model learns.
- Its purpose is to provide insight into the model's decisions.

Attention map

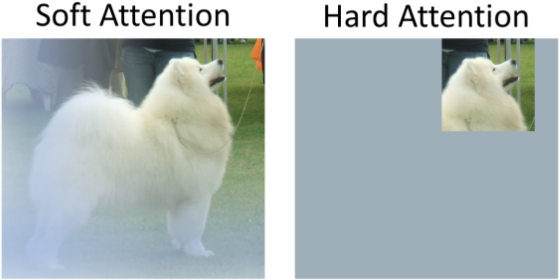
Def.: An **attention map** is a scalar matrix representing the relative importance of layer activations at different 2D spatial locations with respect to the target task.

'border collie' (233)



Soft vs. Hard attention

- A **soft attention** map usually contains decimals between 0 and 1.
- A **hard attention** map contains only 0s and 1s (image cropping).



Outline

- We briefly look at the results from the paper

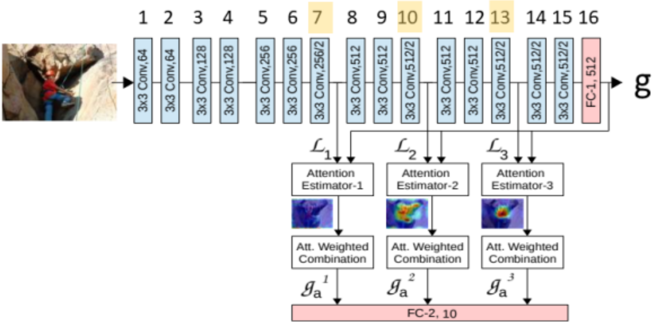
  S. Jetley et al. *Learn to pay attention*. ICLR 2018.

- It discusses *soft trainable attention in a CNN model* for *multiclass classification*.

- The main outcome is that their method improves performance on the CIFAR-100 image data set by 7%.

Model

- The model is based on the VGG CNN (from the **V**isual **G**eometry **G**roup of the University of Oxford).



Notation

$$
\begin{aligned}
s \in \{1, \ldots, 15\} \;&\hat{=}\; \text{index for conv layer} \\
i \in \{1, \ldots, n\} \;&\hat{=}\; \text{index for features of conv layer} \\
g \;&\hat{=}\; \text{“global feature vector”} \\
\ell_i^s \;&\hat{=}\; \text{local features}
\end{aligned}
$$

How the attention works

Step 1: Calculate the **compatibility scores** c_i^s.

- Use local features ℓ and global feature vector g.
- A compatibility score is supposed to have a high value when the image patch described by the local features "contains parts of the dominant image category".
- Two approaches:
  1. **Parameterized compatibility**

$$c_i^s = \langle u, \ell_i^s + g \rangle, \quad i = \{1, \dots, n\} \tag{1}$$

  2. **Dot product**

$$c_i^s = \langle \ell_i^s, g \rangle, \quad i = \{1, \dots, n\} \tag{2}$$

- If ℓ and g are not of the same size, then you can either "project" ℓ to the space of g, or vice versa.

How the attention works – II

Step 2: Calculate the **attention weights** a_i^s.

- Perform a *softmax* operation:

$$\boxed{a_i^s = \frac{\exp(c_i^s)}{\sum_{j=1}^n \exp(c_j^s)}, \quad i = 1, \dots, n. \qquad (3)}$$

Step 3: Calculate the final output of the attention estimator.

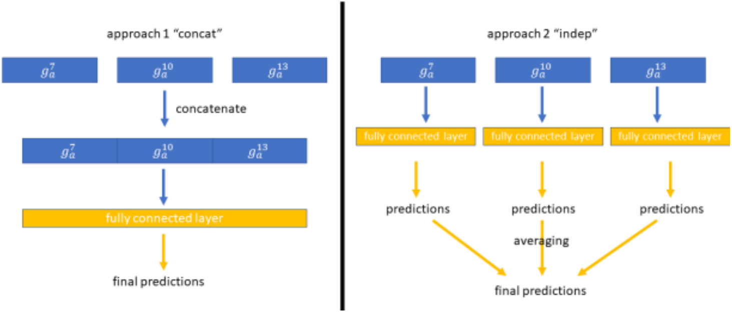
- Compose attention weights a_i^s and local features ℓ_i^s:

$$\boxed{g_a^s = \sum_{i=1}^n a_i^s \cdot \ell_i^s, \quad i = 1, \dots, n. \qquad (4)}$$

How the attention works – III

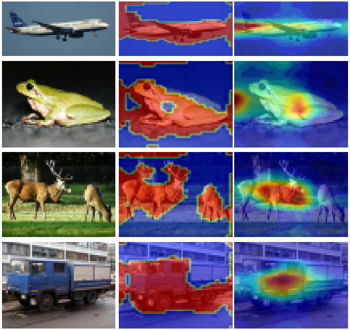
Step 4: Make a classification prediction.

- There are two approaches: *concat* and *indep*.



Results

- The attention mechanism improved performance on image data sets CIFAR-10, CIFAR-100, and SVHN (**S**treet **H**ouse **V**iew **N**umbers).



Input image Layer-10 att. (proposed) Layer-13 att.

Results – II

- *Parameterized compatibility* + *concat* performed best.

- We might be able to neglect g in *parameterized compatibility* as the attention is learned as part of the weight vector u, i. e., we could also use

$$\boxed{c_i^s = \langle u, \ell_i^s \rangle, \quad i = 1, \ldots, n. \qquad (5)}$$

Time for the tutorial!

The **exposition** (also images) followed an **online article** from *towardsdatascience.com* (`link`).

The **article** is based on
S. Jetley et al. *Learn to pay attention*. ICLR 2018.
(`https://arxiv.org/abs/1804.02391`).

The **code for the tutorial** is adapted from
`https://github.com/SaoYan/LearnToPayAttention`.