Lecture IA: Improve the Interpretability of Attention: A Fast, Accurate, and Interpretable High-Resolution Attention Model by Tristan Gomez et al.

Dominique Fourer

IBISC - Team SIAM

15 novembre 2021

Plan

1. Introduction
2. Proposed BR-NP Attention Model
3. Results
4. Conclusion

Attention model [Nadaraya 1964, Watson 1964]

Motivation

- **Attention** model is a promising way for trying to explain and validating trained deep neural architectures.
- Attention can provide insights about how a model is operating (saliency map)
- Mostly related to region of interest in recognition tasks (image captioning, machine translation, speech recognition, text processing, etc.)

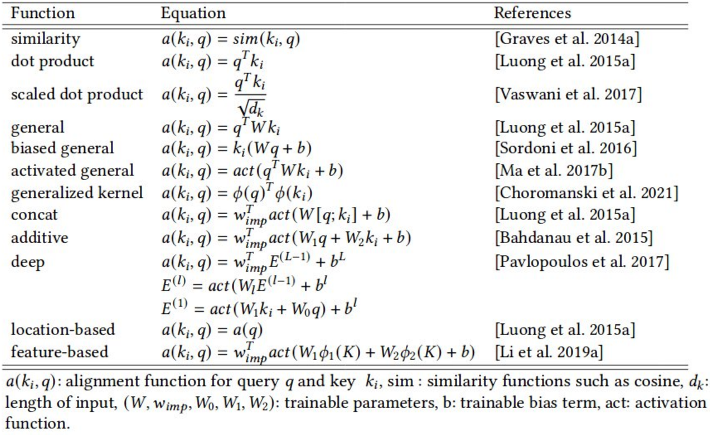
Principle

- Given a training dataset $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n))\}$, we want to predict $\hat{y}$ from x
- A naive estimator will provide : $\hat{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Naradaya-Watson proposed : $\hat{y} = \sum_{i=1}^{n} a(x, x_i)y_i$ where $a(x, x_i)$ corresponds to the relevance of x_i to predict x (alignment function).

First attention models designed for deep learning by Bengio et al in 2014 [1]

[1] * Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv :1409.0473, 2014.

Related work



Function	Equation	References
similarity	$a(k_i, q) = sim(k_i, q)$	[Graves et al. 2014a]
dot product	$a(k_i, q) = q^T k_i$	[Luong et al. 2015a]
scaled dot product	$a(k_i, q) = \frac{q^T k_i}{\sqrt{d_k}}$	[Vaswani et al. 2017]
general	$a(k_i, q) = q^T W k_i$	[Luong et al. 2015a]
biased general	$a(k_i, q) = k_i(Wq + b)$	[Sordoni et al. 2016]
activated general	$a(k_i, q) = act(q^T W k_i + b)$	[Ma et al. 2017b]
generalized kernel	$a(k_i, q) = \phi(q)^T \phi(k_i)$	[Choromanski et al. 2021]
concat	$a(k_i, q) = w_{imp}^T act(W[q; k_i] + b)$	[Luong et al. 2015a]
additive	$a(k_i, q) = w_{imp}^T act(W_1 q + W_2 k_i + b)$	[Bahdanau et al. 2015]
deep	$a(k_i, q) = w_{imp}^T E^{(L-1)} + b^L$	[Pavlopoulos et al. 2017]
	$E^{(l)} = act(W_l E^{(l-1)} + b^l$	
	$E^{(1)} = act(W_1 k_i + W_0 q) + b^l$	
location-based	$a(k_i, q) = a(q)$	[Luong et al. 2015a]
feature-based	$a(k_i, q) = w_{imp}^T act(W_1 \phi_1(K) + W_2 \phi_2(K) + b)$	[Li et al. 2019a]

$a(k_i, q)$: alignment function for query q and key k_i, sim : similarity functions such as cosine, d_k: length of input, $(W, w_{imp}, W_0, W_1, W_2)$: trainable parameters, b: trainable bias term, act: activation function.

Figure : Summary of alignment functions [2]

[2] Chaudhari et al. An Attentive Survey of Attention Models.https://arxiv.org/pdf/1904.02874.pdf.

Attention-based models in the litterature

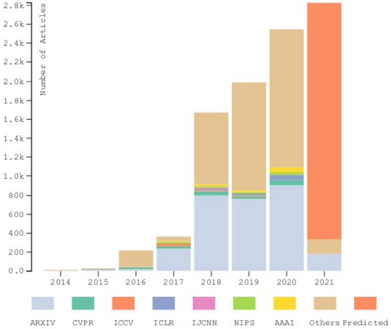


Figure 1: Works published by year between 01/01/2014 to 15/02/2021. The main sources collected are ArXiv, CVPR, ICCV, ICLR, IJCNN, NIPS, and AAAI. The other category refers mainly to the following publishing vehicles: ICML, ACL, ACM, EMNLP, ICRA, ICPR, ACCV, CORR, ECCV, ICASSP, ICLR, IEEE ACCESS, Neurocomputing, and several other magazines.

Figure : source : `https://arxiv.org/pdf/1904.02874.pdf`

Generic DL explaination approaches

- Activation map (average activation accross channels)
- Layer-wise-propagation (LRP)[3]
- Class-Activation Map (CAM, Grad-CAM, Grad-CAM++)[4]
- Guided backpropagation (eg. guided-Grad-CAM, guided-Grad-CAM++)

[3] MONTAVON, Gregoire, BINDER, Alexander, LAPUSCHKIN, Sebastian, et al. Layer-wise relevance propagation : an overview. Explainable AI : interpreting, explaining and visualizing deep learning, 2019, p. 193-209.

[4] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. CoRR, abs/1512.04150, 2015.

Proposed method (BR-NPA : Bilinear representative non-parametric attention)

Based on a CNN architecture + fully connected classification layers (eg. ResNet-50 backbone [5]).

Method Steps

1. Extract high-resolution feature maps
2. Generate Representative Feature Vectors (by grouping similar ones)
3. Concatenate the representative features vectors with classification layer

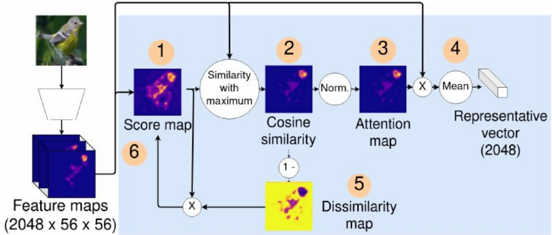


Figure 2: Illustration of the method used to group features without any dedicated module.

[5] K. He, X. Zhang, S. Ren, and J. Sun.Deep residual learning for image recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770-778, 2016.

High-Resolution Feature Map

Idea : Improving the interpretation of the reduced output feature map by increasing the resolution of the attention map without (interpolation)

Approach

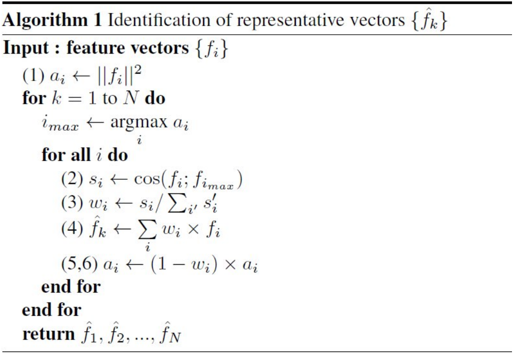
- reducing the stride of the last downsampling layers to increase the size of the feature maps from 14×14 to 56×56.
- use a distillation model proposed by Hinton [a]
- use a teacher-student model where a student network with an increased resolution imitate the lower-resolution teacher network. The training is completed using the following loss function :

$$L(\tilde{y}_s, \tilde{y}_t, y) = \frac{1}{N} \sum_i \alpha \mathrm{CE}(\tilde{y}_s, y) + (1-\alpha)\mathrm{KL}(\tilde{y}_s || \tilde{y}_t) \qquad (1)$$

where $\tilde{y}_s$ and $\tilde{y}_t$ are respectively the output of the student and of the teacher models, y being the ground truth. α is a constant parameter to balance between the cross-entropy and KullBack-Leibler divergence terms.

[a] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

Representative Feature Vector



Algorithm 1 Identification of representative vectors $\{\hat{f_k}\}$

```
Input : feature vectors $\{f_i\}$
  (1) $a_i \leftarrow ||f_i||^2$
  for $k = 1$ to $N$ do
    $i_{max} \leftarrow \underset{i}{\text{argmax}}\ a_i$
    for all $i$ do
      (2) $s_i \leftarrow \cos(f_i; f_{i_{max}})$
      (3) $w_i \leftarrow s_i / \sum_{i'} s'_i$
      (4) $\hat{f_k} \leftarrow \sum_i w_i \times f_i$
      (5,6) $a_i \leftarrow (1 - w_i) \times a_i$
    end for
  end for
  return $\hat{f_1}, \hat{f_2}, ..., \hat{f_N}$
```

- the activation a_i of each feature vector corresponds to its squared l_2-norm $||f_i||^2$
- the feature vector with the maximal activation is denoted f_{imax}
- a feature vector is considered as singular if the other ones are not similar to it (in terms of the cosine distance $\frac{x\dot{y}}{||x||||y||}$))

with $i \in \{0, 1, ...H \times W - 1\}$, $H \times W$ being the dimension of the feature map.

Comparative results

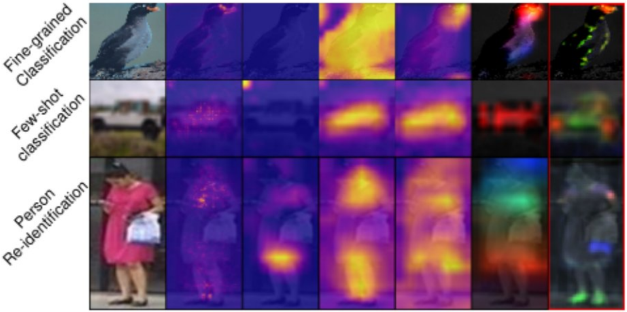
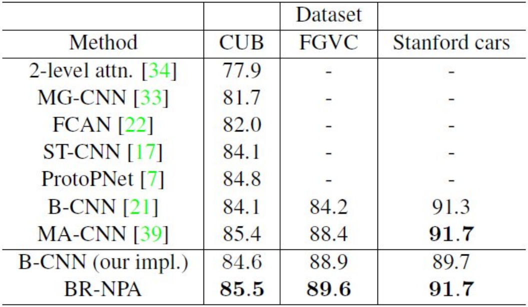


Figure : From left to right : original image, guided Grad-CAM++, Grad-CAM[5], Grad-CAM++, activation map, B-CNN, BR-NPA (proposed by the authors)

[5] Grad-CAM : Chattopadhyay et al. `https://arxiv.org/abs/1710.11063`

Fine-grained image classification

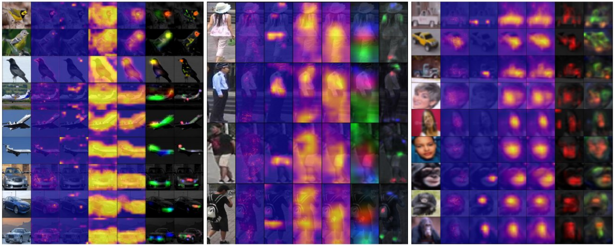


		Dataset	
Method	CUB	FGVC	Stanford cars
2-level attn. [34]	77.9	-	-
MG-CNN [33]	81.7	-	-
FCAN [22]	82.0	-	-
ST-CNN [17]	84.1	-	-
ProtoPNet [7]	84.8	-	-
B-CNN [21]	84.1	84.2	91.3
MA-CNN [39]	85.4	88.4	**91.7**
B-CNN (our impl.)	84.6	88.9	89.7
BR-NPA	**85.5**	**89.6**	**91.7**

Table 1: Performance for task of fine-grained classification.

- Evaluation on 3 datasets : CUB-200-2011, FGVC-Aircraft and Standford cars
- Results expressed in terms of Accuracy (guessed)

Person Id-reidentification



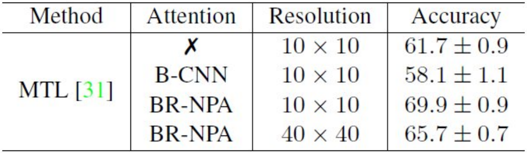
(a) Fine-grained classification (b) Person re-identification (c) Few-shot classification

Figure 3: From left to right : original image, Guided Grad-CAM++, Grad-CAM, Grad-CAM++, AM, B-CNN, BR-NPA. The first four columns indicate saliency with a yellow color, whereas the last two rows indicate saliency using brightness and use colors to indicate by which map the pixel was attended. Red, green and blue represent the 1_{st}, 2_{nd} and 3_{rd} maps respectively.

Method	Attention	Resolution	Accuracy
DG-Net [41]	✗	16×8	94.8
	B-CNN	16×8	78.4
	BR-NPA	16×8	93.6
	BR-NPA	64×32	88.2

Table 2: Performance for task of person re-identification.

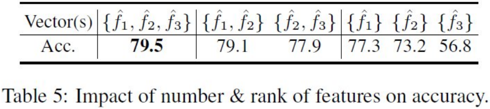
Few-shot classification (CIFAR-FS dataset)



Method	Attention	Resolution	Accuracy
MTL [31]	✗	10×10	61.7 ± 0.9
	B-CNN	10×10	58.1 ± 1.1
	BR-NPA	10×10	69.9 ± 0.9
	BR-NPA	40×40	65.7 ± 0.7

Table 3: Performance for task of few-shot classification.

Ablation study - features relevance

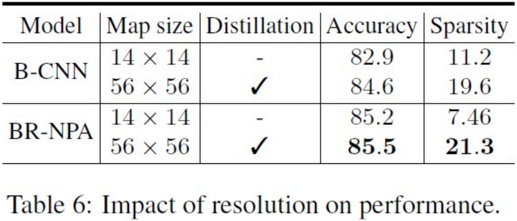


Vector(s)	$\{\hat{f}_1, \hat{f}_2, \hat{f}_3\}$	$\{\hat{f}_1, \hat{f}_2\}$	$\{\hat{f}_2, \hat{f}_3\}$	$\{\hat{f}_1\}$	$\{\hat{f}_2\}$	$\{\hat{f}_3\}$
Acc.	**79.5**	79.1	77.9	77.3	73.2	56.8

Table 5: Impact of number & rank of features on accuracy.

cf. paper, p. 9

Ablation study - feature resolution effect



Model	Map size	Distillation	Accuracy	Sparsity
B-CNN	14×14	-	82.9	11.2
	56×56	✓	84.6	19.6
BR-NPA	14×14	-	85.2	7.46
	56×56	✓	**85.5**	**21.3**

Table 6: Impact of resolution on performance.

cf. paper, p. 9

Attention map illustration 1/2

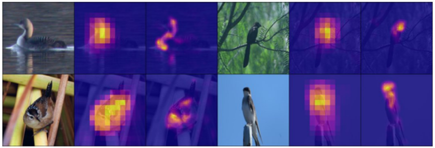


Figure 4: From left to right: original image, attention maps of CNN with lower and then higher-resolution feature maps.

Attention map illustration 2/2

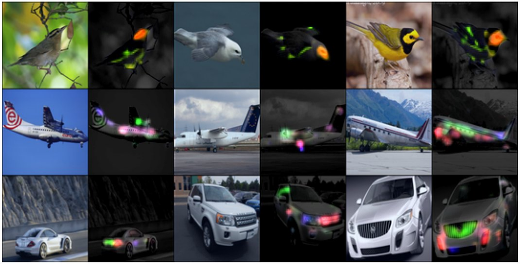


Figure 5: Attention maps produced by BR-NPA.

Conclusion

- A non-parametric attention model to integrate into existing architecture
- improve interpretability without a significant accuracy loss thanks to a higher-resolution refined feature map

Paper of the author recently submitted for publication (example code will be available after acceptance).