Master Thesis

Gabriele Angeletti

April 4, 2017

Contents

1 Acknowledgements

Acknowledgements section.

2 Abstract

Abstract Saenko et al. 2010, yeah Yoo et al. 2016, moar cite Selvaraju et al. 2016
spp

3 Introduction

3.1 Domain Adaptation

Problem of dealing with situations in which the training distribution is different from the test distribution. We assume that the source (train) data is abundant and labeled, while the target (test) data is only partially labeled. The joint distribution of data and labels is, of course, unknown. $P_s(X, Y)$ is the joint distribution of source data, while $P_t(X, Y)$ is the joint distribution of the target data. In a standard classification problem, we assume $P_s(X, Y) = P_t(X, Y) = P(X, Y)$. In Domain Adaptation instead, we have $P_s(X, Y) \neq P_t(X, Y)$. Marginal Distributions: $P_s(X)$, $P_s(Y)$, $P_t(X)$, $P_t(Y)$. Conditional Distributions: $P_s(X|Y)$, $P_s(Y|X)$, $P_t(X|Y)$, $P_t(Y|X)$. If target has labels, we call the problem Supervised Domain Adaptation. If target is unlabeled, we refer to it as Unsupervised Domain Adaptation. One Approach to domain adaptation is that of representation learning, i.e. build a mapping from source representations to target representations so that $P_s(X)$ approaches $P_t(X)$. Many algorithms work with fixed length representations, i.e. they extract (say fc7) features from S and T and then they learn a mapping from one vector to the other in order to near the two representations. Other algorithms instead performs both representation learning and domain adaptation in the same end-to-end learning architecture. Different Approaches to this matching of the feature space distributions are: reweighing the sample in the source domain, find an explicit geometric transformation, modifying the feature representation itself.

3.2 Deep Learning

3.3 Convolutional Neural Networks

4 Related Work

5 Approach

5.1 Grad-CAM

Gradient-weighted Class Activation Mapping Selvaraju et al. 2016. Technique for producing visual explanations about why a Convolutional Neural Network opted for a certain class. The technique is applicable to a wide variety of tasks, ranging from image captioning to visual question answering to reinforcement learning, and also to a wide variety of CNN architectures. In the context of this work, we'll use this technique to produce visuals for an object classification task. This technique was designed as a visualization tool to better understand how CNNs reason and to improve their interpretability, which is a fundamental property to have for AI systems that wants to be deployed in the real world. We extracted the Grad-CAM visualization from the last convolutional layer because, as also the authors or the technique note, the last conv. layer is the best compromise between high-level semantics and spatial localization. The former is a general property of deep learning models: the further the layer from the input, the more high-level and meaningful features one can expect to be extracted. The latter is a property of Convolutional Feature Maps, which are designed to retain as much spatial information as possible about the data. The high-level architecture of the method can be seen in the following figure.

Basically, the input image is forward propagated through the network in order to compute the softmax scores of the last layer, which we call $y^c = f(x)$, where $f()$ is the input-output network mapping. If we were training the network, at this point we would have computed the gradients of the last layer neurons with respect to some loss function, like cross-entropy. Instead, in Grad-CAM, we manually create these gradients. In particular, we set them to one for the neuron corresponding to the class of interest, while all the other neurons are set to zero.

$$\frac{\partial y^c}{\partial \theta_{output}} = [0, 0, \ldots 1, 0, 0, \ldots]$$

Then, we backward propagate these gradients through the network untill the last convolutional layer, where the Grad-CAM map is created. Assume

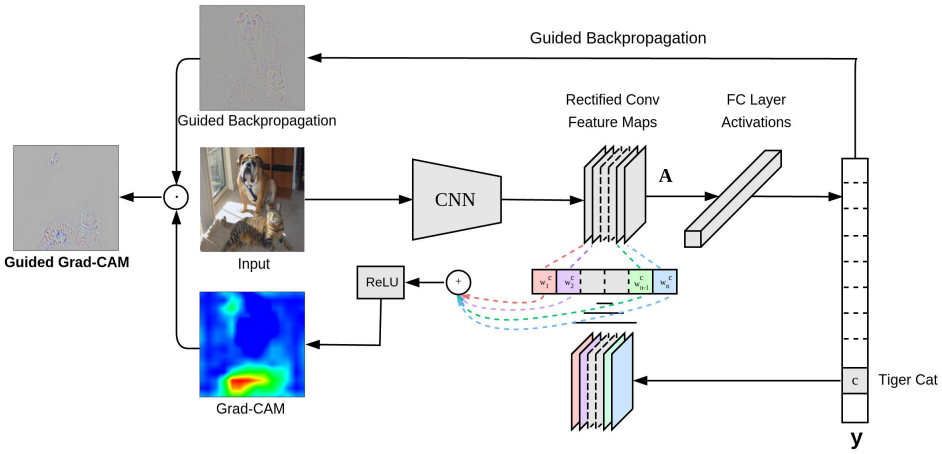

Figure 1: Grad-CAM Architecture, taken from Selvaraju et al. 2016

that the last conv. layer feature maps are in $\mathbb{R}^{K\times H\times W}$, where H and W are respectively the height and the width of each feature map, and K is the number of feature maps. The map that will be generated is thus in $\mathbb{R}^{H\times W}$. The generation procedure has multiple steps:

1. Compute the weight of each feature maps. We want a number measuring how much each filter contributed to the final score. It is therefore natural to compute this number as the derivative of the output gradients with respect to the weights of the filter. These derivatives are then global average pooled in order to get a single number:

$$w_k^c = \frac{1}{Z}\sum_{i=1}^{h}\sum_{j=1}^{w}\frac{\partial y^c}{\partial F_{i,j}^k} \tag{1}$$

   w_k^c is the weight of feature map k with respect to target class c, while F is the feature map itself. We can see that the derivative is positive if an increase of the value of the pixel $F_{i,j}^k$ yields an increase of the value of y^c.

2. The Grad-CAM map is obtained by performing a rectified weighted linear combination of the forward feature maps:

$$M^c = ReLU\left(\sum_{k=1}^{K} w_k^c F^k\right) \tag{2}$$

The $ReLU()$ function simply set to zero all the negative values. The Grad-CAM technique uses this function to retain only the values that have a *positive* influence on the class decision. Once the Grad-CAM map is generate,

it is then upsampled to the pixel space using bi-linear interpolation, then a heatmap is generated from the values and the heatmap is point-wise multiplied with the output of the Guided-BackPropagation procedure Springenberg et al. 2014, to produce the final result Guided-GradCAM.

5.2 Spatial Pyramid Pooling

5.3 Domain-Multiplicative Fusion

6 Experiments

6.1 Office 31

6.2 ICW

6.3 LookBook

7 Comparison

8 Conclusions

References

Saenko, K. et al. (2010). "Adapting visual category models to new domains." In: *Proc. ECCV.*

Selvaraju, Ramprasaath R. et al. (2016). "Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization". In: *CoRR* abs/1610.02391. URL: `http://arxiv.org/abs/1610.02391`.

Springenberg, Jost Tobias et al. (2014). "Striving for Simplicity: The All Convolutional Net". In: *CoRR* abs/1412.6806. URL: `http://arxiv.org/abs/1412.6806`.

Yoo, D. et al. (2016). "Pixel-Level Domain Transfer". In: *Proc. ECCV.*

List of Figures

List of Tables