VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION - VGG

Edjalma Queiroz da Silva
Programa de pós-graduação em Ciências da Computação
Mestrado e Doutorado

Instituto de Informática - UFG

Goiânia, 11 de Novembro de 2016

Sumário I

UFG

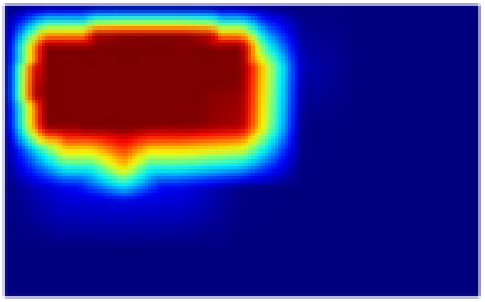
Sumário II

UFG

- No processo de classificação de Imagem, é interessante saber qual ou quais partes da imagem são relevantes para a classificação.
- Dessa forma, o método apresentado por Oquab, Bottou, Laptev e Sivic objetiva descobrir essas áreas utilizando técnica de aprendizado fracamente supervisionada.
- Assim, cada “frame” descoberto, será, classificado e pontuado. A classe de maior pontuação será a escolhida como “mapa de ativação”.

Classificação de Imagem

original image

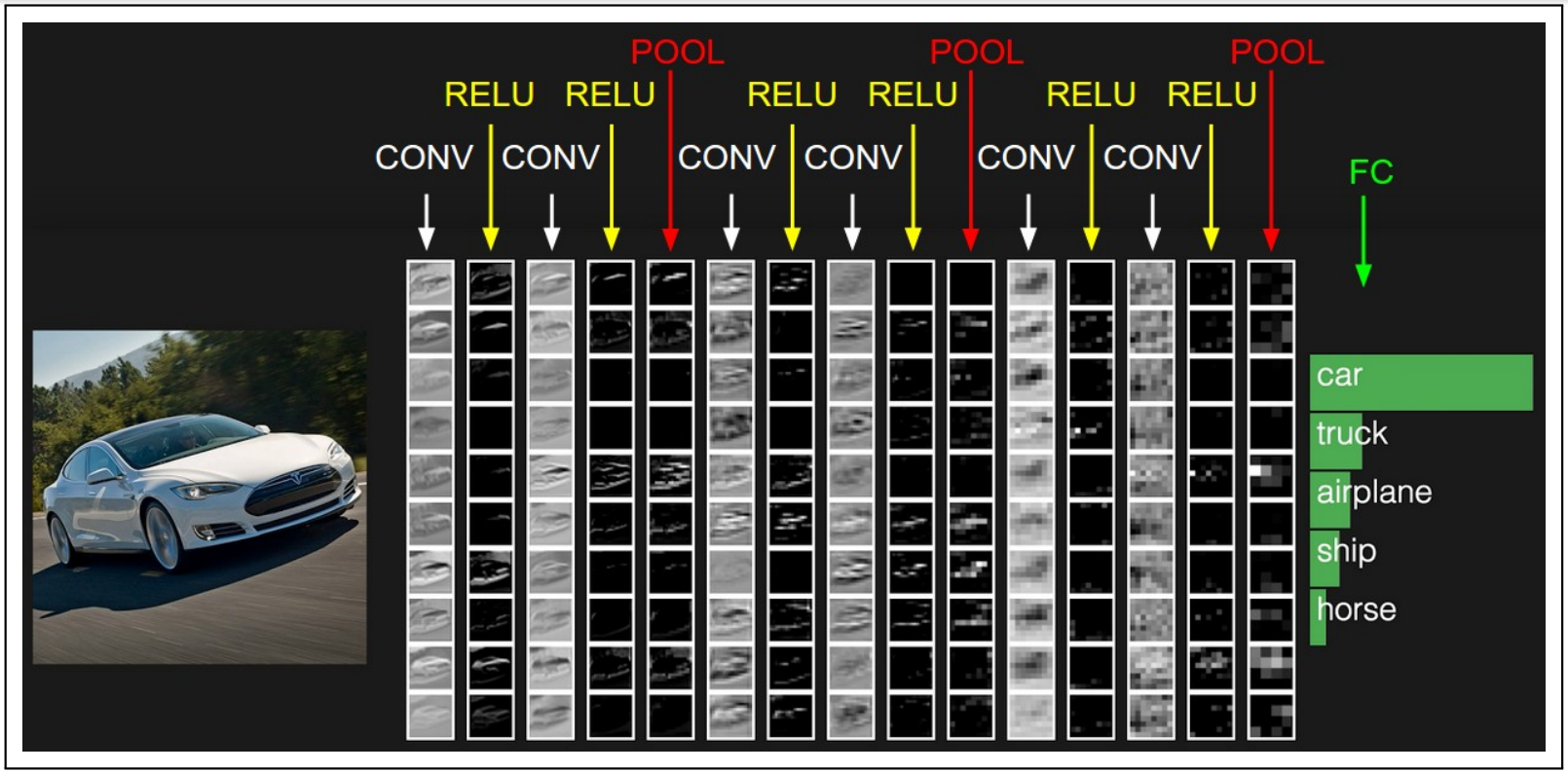


motorbike feature map

max prediction

Oquab, Bottou, Laptev, Sivic. *Is object localization for free? weakly-supervised learning with convolutional neural networks.* CVPR 2015

Classificação de Imagem



Camadas utilizadas para construção de ConvNets

- **Entrada:** Entrada para a Rede. Geralmente em RGB, assim uma imagem de tamanho 32x32, será entendida como 32 de largura, 32 de altura e 3 canais de cores (R, G e B).
- **CONV:** camada responsável por calcular a saída dos neurônios que estão conectados as regiões de entrada. Se caso adotássemos 12 filtros, isso pode resultaria em algo como 32X32X12. Ao final, como resultado da camada de Convolução, teremos uma imagem que gera um conjunto de outras imagens como saída.

Camadas utilizadas para construção de ConvNets

- **RELU:** Aplicará uma função de ativação, como $Max(0, Input)$. O Volume de dados permanecerá inalterado, ou seja, 32x32x12.
- **POOL:** Aplicará alguma operação (Max, Min, Media) de "downsample" ao longo das dimensões espaciais (largura e altura), como resultado temos uma diminuição do volume, por exemplo, saindo de 32x32x12 para 16x16x12. Outro resultado da aplicação desta camada, é a maior ativação para propagar a região de interesse do campo receptivo. Ou seja, captura o que há de mais relevante na imagem, por exemplo.
- **FC:** Fully-Connected, camada que computa a pontuação de cada classe.

VGG Net

- Trabalho publicado[1] por Karen Simonyan e Andrew Zisserman na ICLR[2] 2015.
- O uso na primeira camada de Filtros 3x3 é um diferencial em relação a de 11X11 da AlexNet's e 7x7 da ZF Net's.

[1] disponivel em: http://arxiv.org/pdf/1409.1556v6.pdf
[2] Conferencia Internacional de Representação de Aprendizado (DeepLearnig)

- Criaram 19 camadas CNN[3] utilizando estritamente filtros 3*x*3, juntamente com outro filtro 2*x*2 na segunda camada de maxpooling.
- Dessa forma, consegue simular filtros de maior dimensão, contudo, mantendo os benefícios de ser um filtro pequeno.
- Com isso, obtêm um menor número de parâmetros.

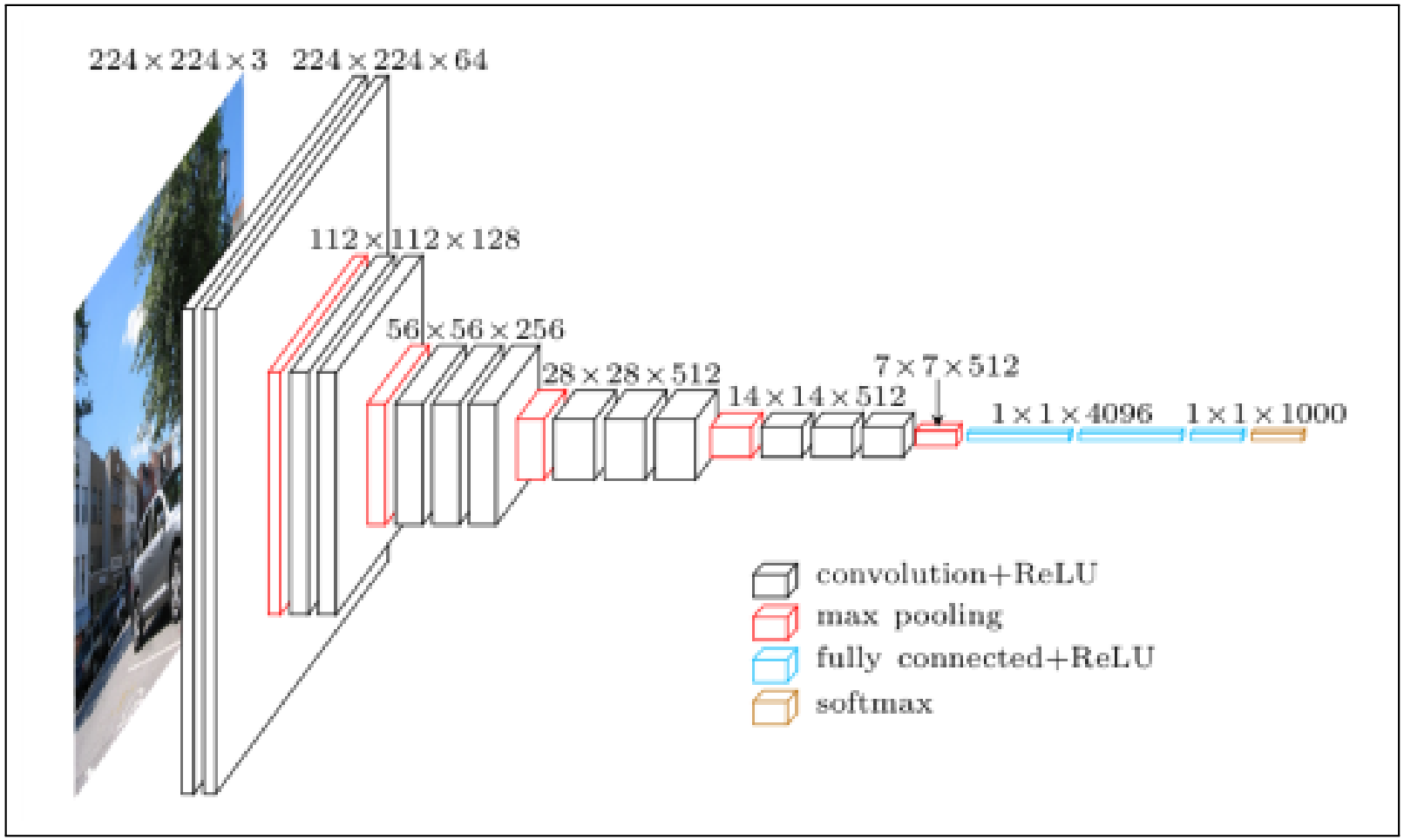
[3] Redes Neurais Convolucionais

- Outro beneficio: com duas camadas de Conv, pode-se habilitar duas camadas ReLU.
- Como o tamanho do volume de entrada diminui (por conta da Convolução e de Pooling), a profundidade da Rede aumenta devido ao aumento do número de filtros a serem aplicados
- O número de Filtros dobram após cada camada de MaxPooling. Assim, ocorre a diminuição da quantidade de dimensões espaciais, contudo, aumenta a profundidade da rede.

- A priori criado para a tarefa de Localização, também apresentou bons resultados para classificação em Imagens.
- Na etapa de regressão, fazem o uso de uma “especie” de localização.
- Como toolbox, criaram o framework Caffe.
- Fazem uso da camada de ReLu após a camada de Convolução (Conv + ReLu)

- A macroarquitetura da VGG16 pode ser observada na Figura 16 (próximo slide).
- Observe que incluem uma camada de pré-processamento que pega a imagem em RGB e padronizam os pixels em valores entre 0 e 255. Após isso, subtrai-se os valores médios com base no conjunto de treino (adotado o conjunto da ImageNet).
- Observe também que há uma grande quantidade de camada de convolução seguidas por Max-pooling, reduzindo a dimensionalidade.

Macro-arquitetura



- O modelo proposto alcança a precisão de 92.7%, ficando entre os 5 melhores na ImageNet, que é um conjunto de dados de mais de 14 milhões de imagens classificadas em 1.000 classes

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB imag)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

- Durante o treinamento, a imagem para a ConvNets é fixada em 224x224
- Existe apenas um pré-processamento: subtração dos valores Medios do RGB, calculados sobre o conjunto de treinamento, de cada pixel da imagem.
- A imagem então é passada para as próximas camadas, que são na verdade uma pilha de Convoluções+ReLU.

- Na pilha de convoluções são aplicados filtros, bem pequenos, 3X3 (que é o menor tamanho para capturar a noção de esquerda/direita, centro, e acima/abaixo).
- A convolução é fixada em 1 pixel, dessa forma, o preenchimento espacial da camada de entrada de convolução, é de tal modo, que a resolução espacial é preservada após a convolução.
- A cada final da execução de uma pilha de convolução, é realizado a chamada a camada de max-pooling.

- Note que a camada de max-pooling não é executada a todo final de UMA max-pooling, e sim, de um conjunto de convolução.
- Assim, a camada de max-pooling é realizada por cinco vezes.
- Max-pooling é executada em uma janela de 2X2.

- Ao final das execuções anteriores, são acionadas 3 camadas de Fully-Connected (FC).
- Os dois primeiros têm exatamente 4096 canais cada uma
- A terceira possui 1000 canais (uma para cada classe).
- Todas as camadas de FC são equipadas comadas ocultas ReLU.

- A camada final é uma camada de soft-max.
- A configuração das camadas FC são as mesma em toda a rede.
- Autores indicam que o uso de normalização, não apontou melhora no desempenho/acurácia do algoritmo, contudo, aumenta o consumo de memória e o tempo de execução.

- Os resultados do uso da VGG demonstra que são competitos em relação a GooLeNet (Campeã no ILSVRC-2012)
- GooLeNet 6.7% de erro, contra 6.8% da VGGNet16
- São extremamente superior em relação a Clarifi (11.2%), campeã no ILSVRC-2013.
- Os resultados demonstram, ainda, a importância do aprendizado profundo na representação visual.

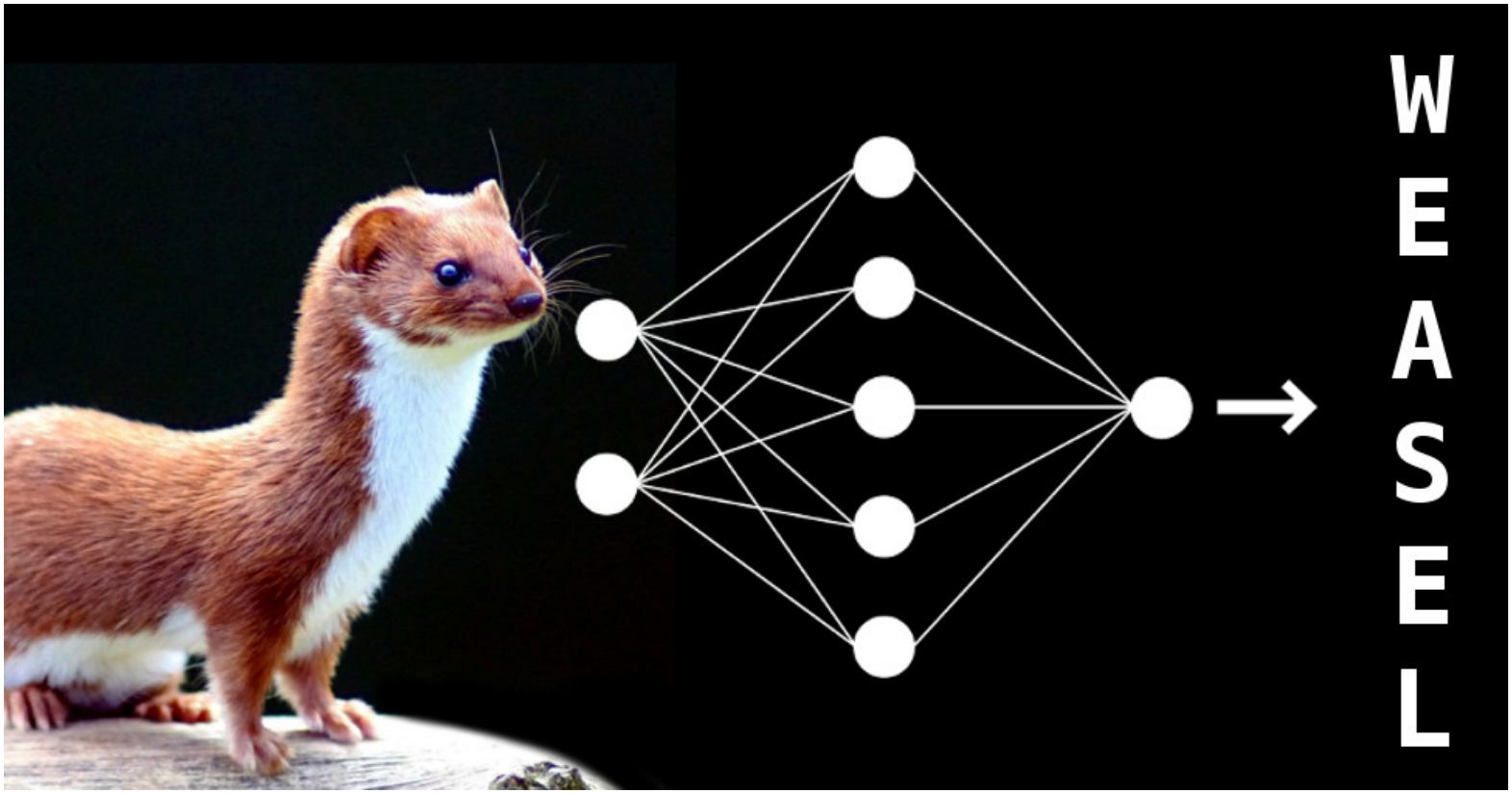
Arquivos necessário:

- arquivo com modelo de pesos
- Arquivo principal, no nosso caso, utilizando framework TensorFlow
- Arquivo com as diferentes classes existentes e classificadas
- Arquivo como entrada, para ser utilizado como exemplo de entrada para a Classificação.

Nota sobre os Pesos

- Os autores disponibilizaram publicamente o arquivo contendo os Pesos.
- Os autores fizeram o uso do framework Caffe[4].
- Foram feitas conversões para que este mesmo arquivo pudesse ser utilizado no framework Tensor Flow (vgg16_weights.npz).

[4] https://gist.github.com/ksimonyan/211839e770f7b538e2d8file-readme-md


WEASEL

Agradecimentos

- Ao INF pela oportunidade.
- Ao professor Anderson[5], que em pouquissimo tempo, na disciplina de TARP, alavancou nossos conhecimentos.

[5] http://www.inf.ufg.br/ anderson/