8.2. Networks for image classification

François Fleuret
https://fleuret.org/dlc/
Nov 24, 2020

Standard convnets
The standard model for image classification are the LeNet family (LeCun et al., 1989; LeCun et al., 1998), and its modern variants such as AlexNet (Krizhevsky et al., 2012) and VGGNet (Simonyan and Zisserman, 2014).

They share a common structure of several convolutional layers seen as a feature extractor, followed by fully connected layers seen as a classifier.

The performance of AlexNet was a wake-up call for the computer vision community, as it vastly out-performed other methods in spite of its simplicity.

Recent advances rely on moving from standard convolutional layers to more complex local architectures to reduce the model size.

torchvision.models provides a collection of reference networks for computer vision, e.g.:

```python
import torchvision
alexnet = torchvision.models.alexnet()
```

The trained models can be obtained by passing pretrained = True to the constructor(s). This may involve an heavy download given there size.

⚠️ The networks from PyTorch listed in the coming slides may differ slightly from the reference papers which introduced them historically.
LeNet5 (LeCun et al., 1989). 10 classes, input $1 \times 28 \times 28$.

(features): Sequential (
    (0): Conv2d(1, 6, kernel_size=(5, 5), stride=(1, 1))
    (1): ReLU (inplace)
    (2): MaxPool2d (size=(2, 2), stride=(2, 2), dilation=(1, 1))
    (3): Conv2d(6, 16, kernel_size=(5, 5), stride=(1, 1))
    (4): ReLU (inplace)
    (5): MaxPool2d (size=(2, 2), stride=(2, 2), dilation=(1, 1))
)

(classifier): Sequential (
    (0): Linear (256 -> 120)
    (1): ReLU (inplace)
    (2): Linear (120 -> 84)
    (3): ReLU (inplace)
    (4): Linear (84 -> 10)
)

Alexnet (Krizhevsky et al., 2012). 1,000 classes, input $3 \times 224 \times 224$.

(features): Sequential (
    (0): Conv2d(3, 64, kernel_size=(11, 11), stride=(4, 4), padding=(2, 2))
    (1): ReLU (inplace)
    (2): MaxPool2d (size=(3, 3), stride=(2, 2), dilation=(1, 1))
    (3): Conv2d(64, 192, kernel_size=(5, 5), stride=(1, 1), padding=(2, 2))
    (4): ReLU (inplace)
    (5): MaxPool2d (size=(3, 3), stride=(2, 2), dilation=(1, 1))
    (6): Conv2d(192, 384, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (7): ReLU (inplace)
    (8): Conv2d(384, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (9): ReLU (inplace)
    (10): Conv2d(256, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (11): ReLU (inplace)
    (12): MaxPool2d (size=(3, 3), stride=(2, 2), dilation=(1, 1))
)

(classifier): Sequential (
    (0): Dropout (p = 0.5)
    (1): Linear (9216 -> 4096)
    (2): ReLU (inplace)
    (3): Dropout (p = 0.5)
    (4): Linear (4096 -> 4096)
    (5): ReLU (inplace)
    (6): Linear (4096 -> 1000)
)
Krizhevsky et al. used **data augmentation** during training to reduce over-fitting.

They generated 2,048 samples from every original training example through two classes of transformations:

- crop a \(224 \times 224\) image at a random position in the original \(256 \times 256\), and randomly reflect it horizontally,
- apply a color transformation using a PCA model of the color distribution.

During test the prediction is averaged over five random crops and their horizontal reflections.

VGGNet19 (Simonyan and Zisserman, 2014). 1,000 classes, input \(3 \times 224 \times 224\). 16 convolutional layers + 3 fully connected layers.

```
(features): Sequential (  
(0): Conv2d(3, 64, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))  
(1): ReLU (inplace)  
(2): Conv2d(64, 64, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))  
(3): ReLU (inplace)  
(4): MaxPool2d (size=(2, 2), stride=(2, 2), dilation=(1, 1))  
(5): Conv2d(64, 128, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))  
(6): ReLU (inplace)  
(7): Conv2d(128, 128, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))  
(8): ReLU (inplace)  
(9): MaxPool2d (size=(2, 2), stride=(2, 2), dilation=(1, 1))  
(10): Conv2d(128, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))  
(11): ReLU (inplace)  
(12): Conv2d(256, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))  
(13): ReLU (inplace)  
(14): Conv2d(256, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))  
(15): ReLU (inplace)  
(16): Conv2d(256, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))  
(17): ReLU (inplace)  
(18): MaxPool2d (size=(2, 2), stride=(2, 2), dilation=(1, 1))  
(19): Conv2d(256, 512, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))  
(20): ReLU (inplace)  
(21): Conv2d(512, 512, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))  
(22): ReLU (inplace)  
(23): Conv2d(512, 512, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))  
(24): ReLU (inplace)  
(25): Conv2d(512, 512, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))  
(26): ReLU (inplace)  
(27): MaxPool2d (size=(2, 2), stride=(2, 2), dilation=(1, 1))  
/.../
VGGNet19 (cont.)

(classifier): Sequential (
    (0): Linear (25088 -> 4096)
    (1): ReLU (inplace)
    (2): Dropout (p = 0.5)
    (3): Linear (4096 -> 4096)
    (4): ReLU (inplace)
    (5): Dropout (p = 0.5)
    (6): Linear (4096 -> 1000)
)

We can illustrate the convenience of these pre-trained models on a simple image-classification problem.

To be sure this picture did not appear in the training data, it was not taken from the web.
import PIL, torch, torchvision

# Load and normalize the image
to_tensor = torchvision.transforms.ToTensor()
img = to_tensor(PIL.Image.open('../example_images/blacklab.jpg'))
img = img.unsqueeze(0)
img = 0.5 + 0.5 * (img - img.mean()) / img.std()

# Load and evaluate the network
alexnet = torchvision.models.alexnet(pretrained = True)
alexnet.eval()

output = alexnet(img)

# Prints the classes
scores, indexes = output.view(-1).sort(descending = True)
class_names = eval(open('imagenet1000_clsid_to_human.txt', 'r').read())
for k in range(12):
    print(f'#{k+1} {scores[k].item():.02f} {class_names[indexes[k].item()]}')
In many applications, standard convolutional networks are made **fully convolutional** by converting their fully connected layers to convolutional ones.
This “convolutionization” does not change anything if the input size is such that the output has a single spatial cell, but it fully re-uses computation to get a prediction at multiple locations when the input is larger.
We can write a routine that transforms a series of layers from a standard convnets to make it fully convolutional:

```python
def convolutionize(layers, input_size):
    result_layers = []
    x = torch.zeros((1,) + input_size)

    for m in layers:
        if isinstance(m, torch.nn.Linear):
            n = torch.nn.Conv2d(in_channels=x.size(1),
                                out_channels=m.weight.size(0),
                                kernel_size=(x.size(2), x.size(3)))
            with torch.no_grad():
                n.weight.view(-1).copy_(m.weight.view(-1))
                n.bias.view(-1).copy_(m.bias.view(-1))
            m = n

        result_layers.append(m)
        x = m(x)

    return result_layers
```

This function makes the [strong and disputable] assumption that only `nn.Linear` has to be converted.

To apply this to AlexNet

```python
model = torchvision.models.alexnet(pretrained=True)
print(model)

layers = list(model.features) + list(model.classifier)
model = nn.Sequential(*convolutionize(layers, (3, 224, 224)))
print(model)
```
AlexNet (features): Sequential (  
  (0): Conv2d(3, 64, kernel_size=(11, 11), stride=(4, 4), padding=(2, 2))  
  (1): ReLU (inplace)  
  (2): MaxPool2d (size=(3, 3), stride=(2, 2), dilation=(1, 1))  
  (3): Conv2d(64, 192, kernel_size=(5, 5), stride=(1, 1), padding=(2, 2))  
  (4): ReLU (inplace)  
  (5): MaxPool2d (size=(3, 3), stride=(2, 2), dilation=(1, 1))  
  (6): Conv2d(192, 384, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))  
  (7): ReLU (inplace)  
  (8): Conv2d(384, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))  
  (9): ReLU (inplace)  
  (10): Conv2d(256, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))  
  (11): ReLU (inplace)  
  (12): MaxPool2d (size=(3, 3), stride=(2, 2), dilation=(1, 1))  
)  
(classifier): Sequential (  
  (0): Dropout (p = 0.5)  
  (1): Linear (9216 -> 4096)  
  (2): ReLU (inplace)  
  (3): Dropout (p = 0.5)  
  (4): Linear (4096 -> 4096)  
  (5): ReLU (inplace)  
  (6): Linear (4096 -> 1000)  
)  

Sequential (  
  (0): Conv2d(3, 64, kernel_size=(11, 11), stride=(4, 4), padding=(2, 2))  
  (1): ReLU (inplace)  
  (2): MaxPool2d (size=(3, 3), stride=(2, 2), dilation=(1, 1))  
  (3): Conv2d(64, 192, kernel_size=(5, 5), stride=(1, 1), padding=(2, 2))  
  (4): ReLU (inplace)  
  (5): MaxPool2d (size=(3, 3), stride=(2, 2), dilation=(1, 1))  
  (6): Conv2d(192, 384, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))  
  (7): ReLU (inplace)  
  (8): Conv2d(384, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))  
  (9): ReLU (inplace)  
  (10): Conv2d(256, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))  
  (11): ReLU (inplace)  
  (12): MaxPool2d (size=(3, 3), stride=(2, 2), dilation=(1, 1))  
  (13): Dropout (p = 0.5)  
  (14): Conv2d(256, 4096, kernel_size=(6, 6), stride=(1, 1))  
  (15): ReLU (inplace)  
  (16): Dropout (p = 0.5)  
  (17): Conv2d(4096, 4096, kernel_size=(1, 1), stride=(1, 1))  
  (18): ReLU (inplace)  
  (19): Conv2d(4096, 1000, kernel_size=(1, 1), stride=(1, 1))  
)
In their “overfeat” approach, Sermanet et al. (2013) combined this with a stride 1 final max-pooling to get multiple predictions.

Doing so, they could afford parsing the scene at 6 scales to improve invariance.

This “convolutionization” has a practical consequence, as we can now re-use classification networks for dense prediction without re-training.

Also, and maybe more importantly, it blurs the conceptual boundary between “features” and “classifier” and leads to an intuitive understanding of convnet activations as gradually transitioning from appearance to semantic.
In the case of a large output prediction map, a final prediction can be obtained by averaging the final output map channel-wise.

If the last layer is linear, the averaging can be done first, as in the residual networks (He et al., 2015).
Lin et al. (2013) re-interpreted a convolution filter as a one-layer perceptron, and extended it with an “MLP convolution” (aka “network in network”) to improve the capacity vs. parameter ratio.

![Diagram](image)

(Lin et al., 2013)

As for the fully convolutional networks, such local MLPs can be implemented with 1 x 1 convolutions.

The same notion was generalized by Szegedy et al. (2015) for their GoogLeNet, through the use of module combining convolutions at multiple scales to let the optimal ones be picked during training.

![Diagram](image)

(a) Inception module, naïve version  
(b) Inception module with dimension reductions

(Szegedy et al., 2015)
Szegedy et al. (2015) also introduce the idea of **auxiliary classifiers** to help the propagation of the gradient in the early layers.

This is motivated by the reasonable performance of shallow networks that indicates early layers already encode informative and invariant features.

The resulting GoogLeNet has 12 times less parameters than AlexNet and is more accurate on ILSVRC14 (Szegedy et al., 2015).

It was later extended with techniques we are going to see in the next slides: batch-normalization (Ioffe and Szegedy, 2015) and pass-through à la resnet (Szegedy et al., 2016).
We already saw the structure of the residual networks and how well they perform on CIFAR10 (He et al., 2015).

The default residual block proposed by He et al. is of the form

\[
\begin{align*}
\text{Conv} & \quad 3 \times 3 \\
& \quad 64 \rightarrow 64 \\
\text{BN} & \quad \text{ReLU} \\
\text{Conv} & \quad 3 \times 3 \\
& \quad 64 \rightarrow 64 \\
\text{BN} & \quad + \\
& \quad \text{ReLU} \\
\end{align*}
\]

and as such requires \(2 \times (3 \times 3 \times 64 + 1) \times 64 \simeq 73k\) parameters.
To apply the same architecture to ImageNet, more channels are required, e.g.

To apply the same architecture to ImageNet, more channels are required, e.g.

$$256 \times 64 + (3 \times 3 \times 64 + 1) \times 64 + 64 \times 256 \simeq 70k$$ parameters.

They mitigated that requirement with what they call a bottleneck block:

$$256 \times 64 + (3 \times 3 \times 64 + 1) \times 64 + 64 \times 256 \simeq 70k$$ parameters.

The encoding pushed between blocks is high-dimensional, but the “contextual reasoning” in convolutional layers is done on a simpler feature representation.

<table>
<thead>
<tr>
<th>method</th>
<th>top-5 err. (test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG [41] (ILSVRC’14)</td>
<td>7.32</td>
</tr>
<tr>
<td>GoogLeNet [44] (ILSVRC’14)</td>
<td>6.66</td>
</tr>
<tr>
<td>VGG [41] (v5)</td>
<td>6.8</td>
</tr>
<tr>
<td>PReLU-net [13]</td>
<td>4.94</td>
</tr>
<tr>
<td>BN-inception [16]</td>
<td>4.82</td>
</tr>
<tr>
<td><strong>ResNet (ILSVRC’15)</strong></td>
<td><strong>3.57</strong></td>
</tr>
</tbody>
</table>

Table 5. Error rates (%) of ensembles. The top-5 error is on the test set of ImageNet and reported by the test server.

(He et al., 2015)
This was extended to the ResNeXt architecture by Xie et al. (2016), with blocks with similar number of parameters, but split into 32 “aggregated” pathways.

When equalizing the number of parameters, this architecture performs better than a standard resnet.

Tan and Le (2019) proposed to scale depth, width, and resolutions uniformly when increasing the size of a network.

Figure 1. Model Size vs. ImageNet Accuracy. All numbers are for single-crop, single-model. Our EfficientNets significantly outperform other ConvNets. In particular, EfficientNet-B7 achieves new state-of-the-art 84.4% top-1 accuracy but being 8.4x smaller and 6.1x faster than GPipe. EfficientNet-B1 is 7.6x smaller and 5.7x faster than ResNet-152. Details are in Table 2 and 4.

(Tan and Le, 2019)
To summarize roughly the evolution of convnets for image classification:

- standard ones are extensions of LeNet5,
- everybody loves ReLU,
- state-of-the-art networks have 100s of channels and 10s of layers,
- they can (should?) be fully convolutional,
- pass-through connections allow deeper “residual” nets,
- bottleneck local structures reduce the number of parameters,
- aggregated pathways reduce the number of parameters.
Image classification networks

References


