Tutorial: (Some) Best Practices of ConvNet Application

Sheng Jia
Feb. 15, 2022
(Adapted from Jenny Bao’s slides in winter 2021)
Math-heavy tutorial
-> High-level guidance
Overview

- Transfer Learning
- Label Imbalance
- Normalization
Transfer learning: idea

Instead of training a deep network from scratch for your task:

- Take a network trained on a different domain for a different source task
- Adapt it for your domain and your target task

Variations:

- Same domain, different task
- Different domain, same task

Slides from: https://towardsdatascience.com/a-comprehensive-hands-on-guide-to-transfer-learning-with-real-world-applications-in-deep-learning-212bf3b2f27a
 Freeze or fine-tune?

Bottom $n$ layers can be frozen or fine tuned.

- **Frozen**: not updated during backprop
- **Fine-tuned**: updated during backprop

Which to do depends on target task:

- **Freeze**: target task labels are scarce, and we want to avoid overfitting
- **Fine-tune**: target task labels are more plentiful

In general, we can set learning rates to be different for each layer to find a tradeoff between freezing and fine tuning.

Slides from:
https://towardsdatascience.com/a-comprehensive-hands-on-guide-to-transfer-learning-with-real-world-applications-in-deep-learning-212bf3b2f27a
## Transfer Learning: Rule of thumb

<table>
<thead>
<tr>
<th>Target Dataset</th>
<th>is small</th>
<th>Target Dataset</th>
<th>is large</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similar to Source dataset</td>
<td>Freeze</td>
<td>Fine-tune all</td>
<td></td>
</tr>
<tr>
<td>Dissimilar to Source dataset</td>
<td>Try SVM from low-level features first</td>
<td>Train from scratch</td>
<td></td>
</tr>
</tbody>
</table>

[http://cs231n.github.io/transfer-learning/#tf](http://cs231n.github.io/transfer-learning/#tf)
Transfer Learning

- Additional advice:
  - Smaller learning rate when fine-tuning

http://cs231n.github.io/transfer-learning/#tf
Task Transfer Learning

- Same domain, different tasks
- Computer Vision Taskonomy: [http://taskonomy.stanford.edu](http://taskonomy.stanford.edu)
- What is the relation between 3d keypoint detection and depth estimation?

*Taskonomy: Disentangling Task Transfer Learning*, Amir et al, 2018
Task Transfer Learning

- Same domain, different tasks
- Computer Vision Taskonomy: [http://taskonomy.stanford.edu](http://taskonomy.stanford.edu)
- What is the relation between 3d keypoint detection and depth estimation?
- Is it able to structurally represent them?

*Taskonomy: Disentangling Task Transfer Learning*, Amir et al, 2018
Task Transfer Learning

Task Similarity Tree Based on Transfering-Out

Taskonomy: Disentangling Task Transfer Learning, Amir et al, 2018
Task Transfer Learning: Result

- How significant is the discovered structure of task space?

Green: look up the taxonomy connectivities.

Gray: use random connectivities

Taskonomy: Disentangling Task Transfer Learning, Amir et al, 2018
Transfer Learning from ImageNet?

- Always better?
- ImageNet: 130M
- COCO: 8.6M

Rethinking ImageNet Pre-training, Kaiming et al, 2019
Transfer Learning from ImageNet?

Rethinking ImageNet Pre-training, Kaiming et al, 2019
Transfer Learning from ImageNet?

- With only 1k training image:
  - w/ pretrain: 9.9 AP
  - Random init: 3.5 AP
  (on validation set)

Overfitting without transfer learning
Train loss similar at convergence but validation error different this time!

Rethinking ImageNet Pre-training, Kaiming et al, 2019
Transfer Learning from ImageNet?

One conclusion:

Training from scratch can be no worse than its ImageNet pre-training counterparts under many circumstances, down to 10k COCO images.

Rethinking ImageNet Pre-training, Kaiming et al, 2019
## Transfer Learning: Rule of thumb

<table>
<thead>
<tr>
<th>Target Dataset is small</th>
<th>Target Dataset is large</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similar to Source dataset</td>
<td>Freeze</td>
</tr>
<tr>
<td>Dissimilar to Source dataset</td>
<td>Try SVM from low-level features first</td>
</tr>
</tbody>
</table>
Label Imbalance

- Semantic Segmentation
- Contour Detection
- Long tail recognition
Label Imbalance

- Reweight the loss by class ratio
- Data Resampling by class ratio
Structure of ConvNet

- Conv -> **Normalization** -> ReLU -> Pooling
Normalization layers

\[ y = \frac{x - E[x]}{\sqrt{\text{Var}[x]} + \epsilon} \gamma + \beta \]

Learnable parameters, to make sure the normalization layer can represent identity transformation

- Batch normalization
- Layer normalization
- Instance normalization
- Group normalization
BatchNorm

- Internal Covariate Shift
- Compute batch statistic during training
  - Dependent on mini-batch

\[
\mu_B \leftarrow \frac{1}{m} \sum_{i=1}^{m} x_i \quad \text{// mini-batch mean}
\]

\[
\sigma_B^2 \leftarrow \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu_B)^2 \quad \text{// mini-batch variance}
\]

\[
\hat{x}_i \leftarrow \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \quad \text{// normalize}
\]

\[
y_i \leftarrow \gamma \hat{x}_i + \beta \equiv \text{BN}_{\gamma, \beta}(x_i) \quad \text{// scale and shift}
\]
BatchNorm for CNN

- Jointly normalize all the activations in a minibatch, over all locations.
- “Effective minibatch size of $N' = N \times H \times W$.
- Learn a pair of parameters per feature map, rather than per activation.
BatchNorm

- Usually, during training, BN keeps a running estimate of the mean and variance, which are used at testing time.

Recall:

\[ y = \frac{x - \mathbb{E}[x]}{\sqrt{\text{Var}[x] + \epsilon}} \ast \gamma + \beta \]

Running estimates of \( \mathbb{E}[x], \text{Var}[x] \) besides learnable parameters.
BatchNorm Example

Pytorch documentation

```python
class torch.nn.BatchNorm2d(num_features, eps=1e-05, momentum=0.1,
                           affine=True, track_running_stats=True)

- num_features: C from an expected input of size (N, C, H, W)
```

Example: convolution block in Inception Net V3

```python
class BasicConv2d(nn.Module):

    def __init__(
        self,
        in_channels: int,
        out_channels: int,
        **kwargs: Any
    ) -> None:
        super(BasicConv2d, self).init ()
        self.conv = nn.Conv2d(in_channels, out_channels, bias=False, **kwargs)
        self.bn = nn.BatchNorm2d(out_channels, eps=0.001)

    def forward(self, x: Tensor) -> Tensor:
        x = self.conv(x)
        x = self.bn(x)
        return F.relu(x, inplace=True)
```
BatchNorm -- limitations

- Performance depends on the batch size
- Difficult to apply to recurrent connections

<table>
<thead>
<tr>
<th>batch dimension</th>
<th>sequence dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
LayerNorm

- Normalize across the entire layer for each training example.
LayerNorm

Figure 2: Validation curves for the attentive reader model. BN results are taken from [Cooijmans et al., 2016].
LayerNorm Example

Pytorch documentation

```python
input = torch.randn(20, 5, 10, 10)
# With Learnable Parameters
m = nn.LayerNorm(input.size()[1:])
# Without Learnable Parameters
m = nn.LayerNorm(input.size()[1:], elementwise_affine=False)
# Normalize over last two dimensions
m = nn.LayerNorm([10, 10])
# Normalize over last dimension of size 10
m = nn.LayerNorm(10)
# Activating the module
output = m(input)
```
InstanceNorm

- Special Case: Feed-Forward Stylization
InstanceNorm

- Special Case: Feed-Forward Stylization
- Invariant to the contrast (style) of the content image
InstanceNorm

- Special Case: Feed-Forward Stylization
- Invariant to the *contrast* (style) of the content image
- **Channel-wise** normalization
InstanceNorm

- Special Case: Feed-Forward Stylization
- Invariant to the **contrast** of the content image
- Normalize over channel for each image
**InstanceNorm Example**

*Pytorch documentation*

```
CLASS torch.nn.InstanceNorm2d(num_features: int, eps: float = 1e-05, momentum: float = 0.1,
                          affine: bool = False, track_running_stats: bool = False)
```

- `num_features`: C from an expected input of size (N, C, H, W)
- By default, there are no learnable parameters, and does not track running statistics (unlike BN or LN)

```python
# Without Learnable Parameters
m = nn.InstanceNorm2d(100)
# With Learnable Parameters
m = nn.InstanceNorm2d(100, affine=True)
input = torch.randn(20, 100, 35, 45)
output = m(input)
```
GroupNorm

- Large Feed-Forward network
  - Sometimes batch size is small due to computational constraints
- How to adjust?
  - GroupNorm
GroupNorm

- Group a set of features and normalize them
  - like normalizing HOG and SIFT separately
GroupNorm
GroupNorm Example

Pytorch documentation

- num_groups (int) – number of groups to separate the channels into
- num_channels (int) – number of channels expected in input

```python
input = torch.randn(20, 6, 10, 10)
# Separate 6 channels into 3 groups
m = nn.GroupNorm(3, 6)
# Separate 6 channels into 6 groups (equivalent with InstanceNorm)
m = nn.GroupNorm(6, 6)
# Put all 6 channels into a single group (equivalent with LayerNorm)
m = nn.GroupNorm(1, 6)
# Activating the module
output = m(input)
```
**SyncBatchNorm**

- Split large batch into several and distribute them many GPUs
  - Collect the batch statistics from all devices
Any Question?