**量化**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 模型名称 | 量化前大小 | 量化后大小 | 压缩前准确率 | 压缩后准确率 | 压缩前推理时间 | 压缩后推理时间/ms |
| mobilenet | 17.1M | 4.3M | 70.7% | 57.7% | 18 | 12.38 |
| Inception\_v1 | 26.1M | 6.8M | 69.8% | 66.6% | 21.2 | 26.56 |
| Vgg\_16 | 540.4M | 135.1M | 71.5% | 68.7% | 93.9 | 131.5 |
| Vgg\_19 | 561.2M | 140.3M | 71.1% | 68.0% | 110.3 | 159.9 |
| Resnet\_50 | 100.2M | 25.3M | 75.2% | 72.9% | 43 | 64.5 |
| Inception\_v2 | 43.9M | 11.3M | 73.9% | 71.9% | 31.5 | 44.4 |
| Resnet\_101 | 174.5M | 44.1M | 76.4% | 73.8% | 62.3 | 98.4 |
| Resnet\_152 | 235.9M | 58.73M | 76.8% | 73.5% | 82 | 147.6 |
| Inception\_v3 | 92.4M | 24.2M | 78% | 75.4% | 60.3 | 168.8 |
| Inception\_v4 | 167.2M | 42.5M | 80.2% | 78.4% | 107.2 | 206.5 |

**剪裁**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 模型 | 数据集 | 压缩前大小 | 压缩后大小 | 压缩前准确率 | 压缩后准确率 | 压缩前推理时间 | 压缩后推理时间 |
| Vgg16 | Imagenet | 553M | 512M | 71.5% | 68.0% | 93.9 | 60.1 |
| Resnet50 | Imagenet | 100.2M | 84.3M | 75.2% | 70.2% | 43 | 34.5 |