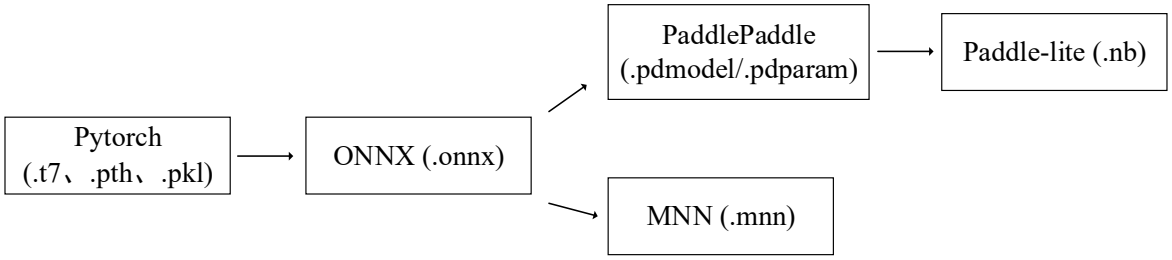
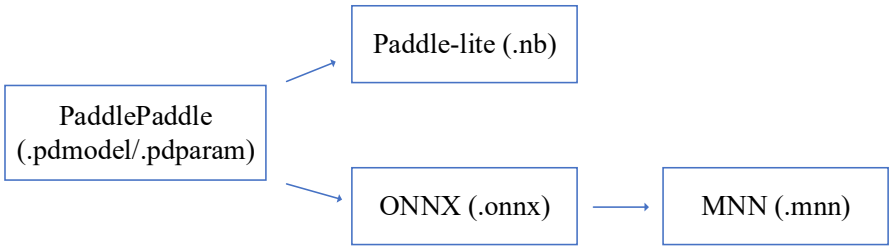
相关库版本信息

paddlehub	**2.3.0**
paddlepaddle	**2.3.2**
paddle2onnx	**1.0.1**
onnx	**1.12.0**
onnxsim	**0.4.8**
torch	**1.11.0+cpu**
torchvision	**0.12.0+cpu**
MNN	**master**
Paddle-lite	**develop**

推理框架与格式转换





预训练模型

PaddlePaddle 与 Pytorch 下载预训练模型：VGG16 不加 bn 层、Resnet50

格式转换后模型大小

训练框架	模型名	MNN (.m)	Paddle-Lite (.m)
Pytorch 模型	Torch_resnet50_sim	102.4	102.3
	torch_vgg16	553	553.5
Paddle 模型	res50_paddle2onnx_sim	102.4	102.3
	vgg16_paddle2onnx	553	553.5

时间对比

20 张测试图像

Vgg16 模型参数量过大，在 8QM 上加载时，因内存不足，导致无法加载模型进行推理。

MNN 在 8QM 上 CPU 和 GPU 时间统计：

CPU		前处理(ms)	推理(ms)	后处理(ms)	总耗时(ms)
res50_paddle2onnx_sim.mnn	Mean	1.2234	1488.54	13.4836	1503.24
	Min	0.897	1288.91	8.6465	~
	Max	3.027	1764.82	22.6033	~
Torch_resnet50_sim.mnn	Mean	1.24977	1488.49	11.9772	1501.72
	Min	0.9493	1282.31	4.310	~
	Max	2.20012	1714.07	22.499	~
GPU(Opencl as backend**)**					
res50_paddle2onnx_sim.mnn	Mean	3.37926	505.945	11.975	521.301
	Min	1.60612	490.103	9.78687	~
	Max	8.38612	523.674	17.6423	~
Torch_resnet50_sim.mnn	Mean	3.0990	503.054	3.64229	509.796
	Min	1.45512	487.663	3.528	
	Max	5.7013	517.835	3.873	

Paddle-lite 在 8QM 上 CPU 时间统计：

CPU		前处理(ms)	推理(ms)	后处理(ms)	总耗时(ms)
resnet50_sim.nb	Mean	8.6263	1761.17	0.0300	1769.83
	Min	6.6286	1665.84	0.024	~
	Max	11.2397	5388.18	0.006	~
torch_res50_arm.nb	Mean	9.1640	1802.82	0.025	1812.01
	Min	6.6468	1467.79	0.023	~
	Max	15.016	5162.91	0.028	~

以上 MNN 和 paddle-lite 在推理过程中均使用单核 cpu。

因此，不同框架训练得到模型在不同推理框架上部署时，只要模型结构一致，推理耗时的差距由部署框架导致。

用 torch_res50_arm.nb 模型在 paddle-lite 推理框架中设置 CPU 线程数量为 2 结果为：

CPU		前处理(ms)	推理(ms)	后处理(ms)	总耗时(ms)
torch_res50_arm.nb	Mean	9.3517	2112.75	0.027	2122.14
	Min	6.6361	1744.36	0.023	~
	Max	14.8877	3982.93	0.046	~

此外，在测试中对同一个模型进行多次加载推理，得到耗时结果进行比较：
MNN 结果

CPU		Test_01	Test_02
res50_paddle2onnx_sim.mnn	Mean	1503.24	1477.78
Torch_resnet50_sim.mnn	Mean	1501.72	~
GPU(Opencl as backend**)**			

res50_paddle2onnx_sim.mnn	Mean	521.301	513.029
Torch_resnet50_sim.mnn	Mean	509.796	511.737

Paddle-lite 结果

CPU		Test_01	Test_02
res50_arm.nb	Mean	1737.4	1769.83
torch_res50_arm.nb	Mean	1812.01	1676.71

MNN 框架在同一模型多次加载推理过程中，推理耗时浮动不大；paddle-lite 在模型多次加载过程中，推理耗时浮动较大。**(此结论有待进一步多次加载验证)**