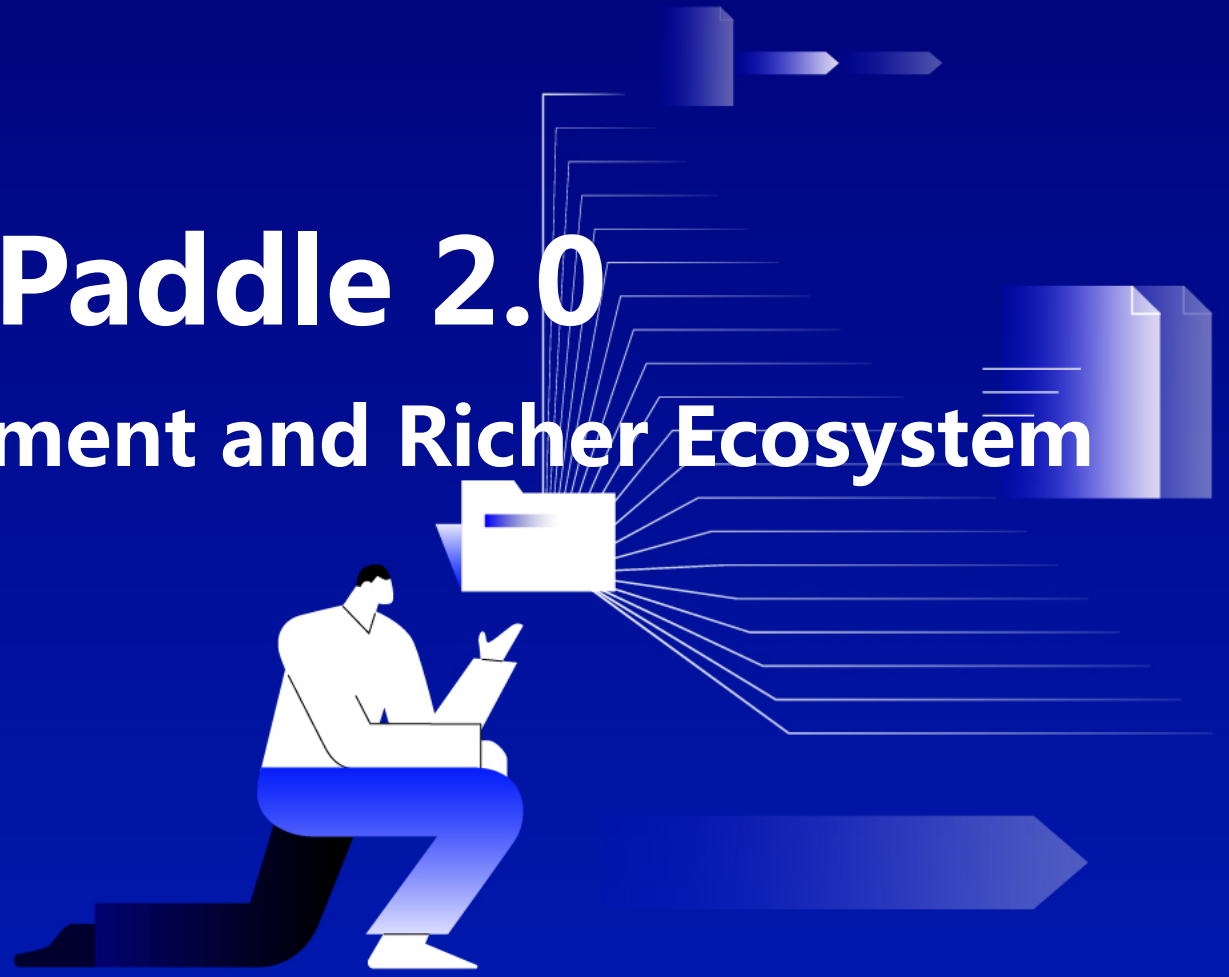
ONNX on PaddlePaddle 2.0

——Broader Deployment and Richer Ecosystem

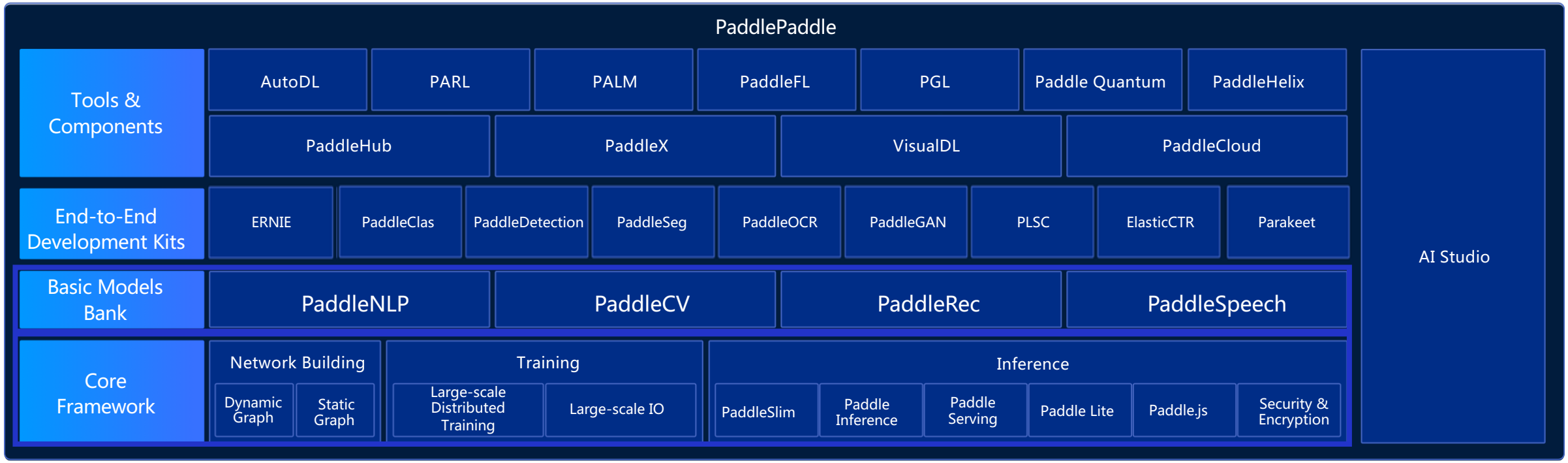
2021.03 Wranky Wang



Contents

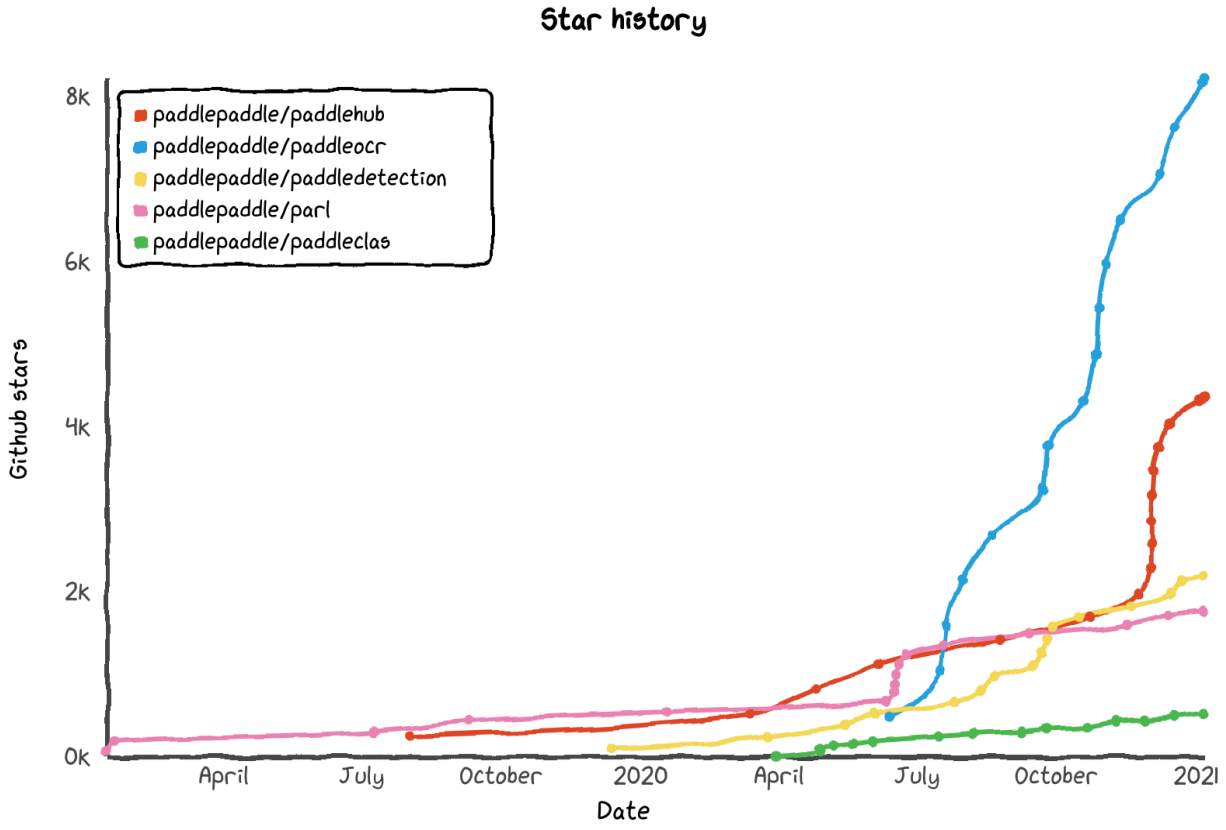
Overview of PaddlePaddle

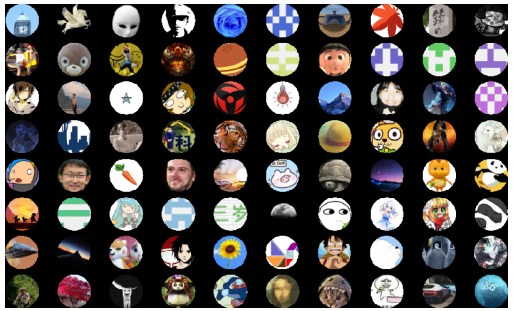
- Agile Framework for Industrial Development of Deep Neural Networks
- Support Ultra-Large-Scale Training of Deep Neural Networks
- Accelerated High-Performance Inference over Ubiquitous Deployments
- Industry-Oriented Models and Libraries with Open Source Repositories



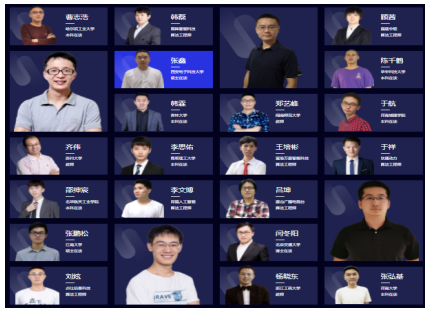
Community of PaddlePaddle

Over 61k Stars, many projects enter Github trending list

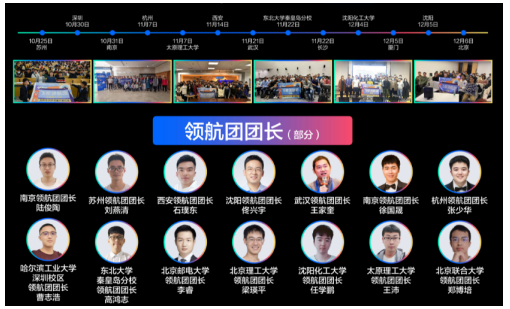




More than **160k** accumulated commits
More than **5000** contributed developers*



97 PPDE, 7 PPSIG



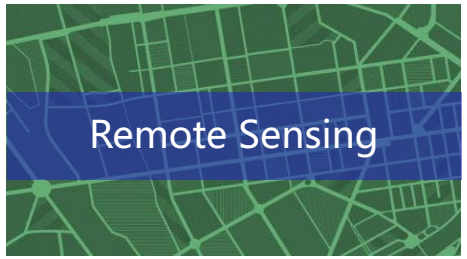
132 Urban/College self-organized community

AI contests covering **580** colleges

* Contributed by issue or PR.

Application of PaddlePaddle

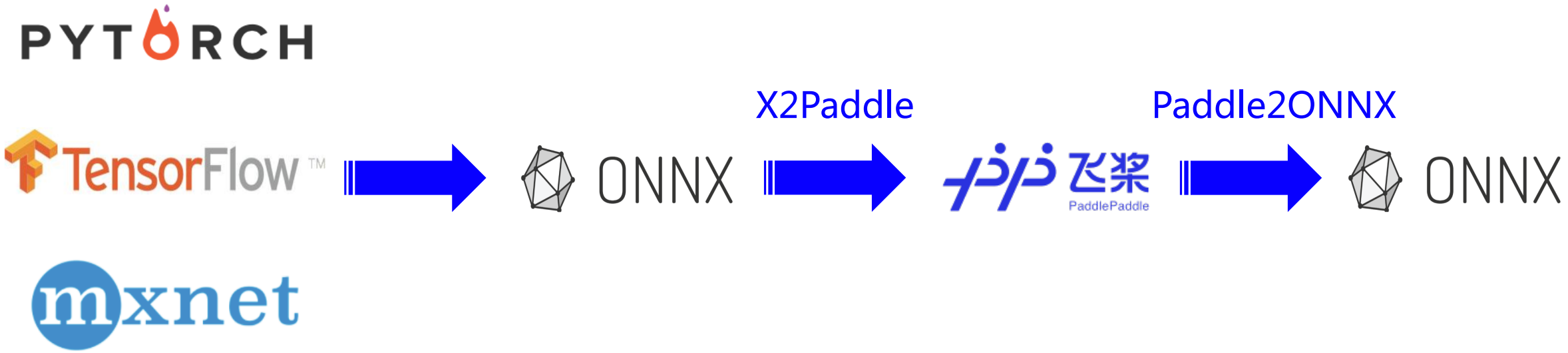
Industry



Company

intel
NVIDIA
中信银行 CHINA CITIC BANK
inspur 浪潮
HUAWEI
Cambricon 寒武纪
China unicom 中国联通
中国南方电网 CHINA SOUTHERN POWER GRID 广东电网有限责任公司电力科学研究院
比特大陆 BITMAIN
深圳证券交易所 SHENZHEN STOCK EXCHANGE
千千音乐 Qian Qian Music

PaddlePaddle and ONNX



X2Paddle

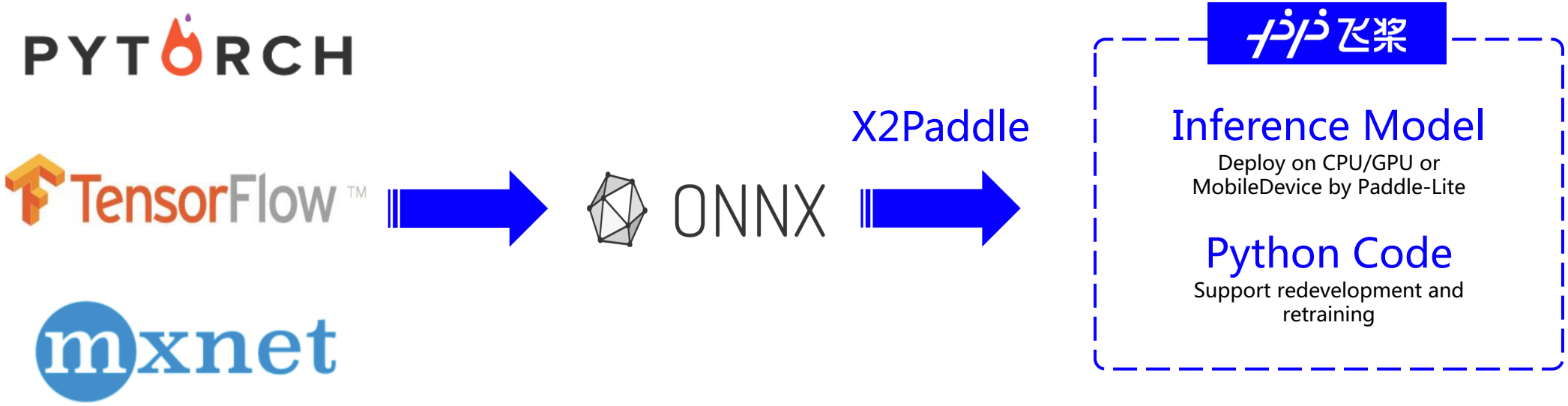
- Enrich the pretrained models for PaddlePaddle
- Support model deployment of other frameworks

Paddle2ONNX

- Extend the deployment method of PaddlePaddle models
- Reduce the development cost of cooperation with PaddlePaddle

Contents

X2Paddle



Command:

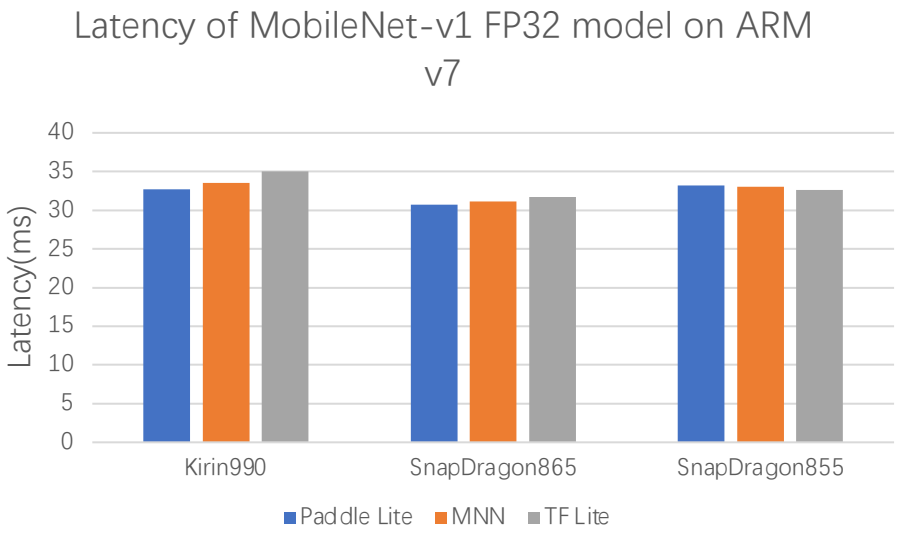
```
x2paddle --framework onnx --model resnet50.onnx  --save_dir paddle_model
```

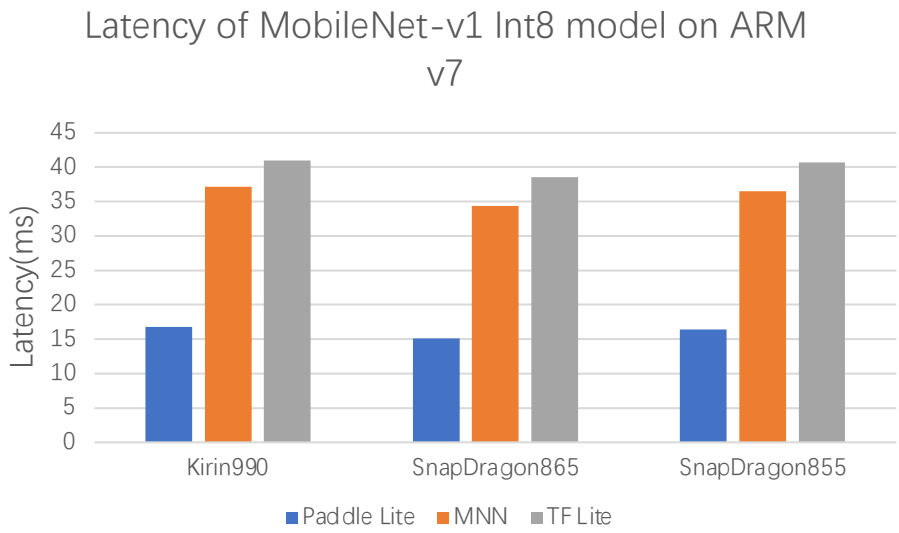
Support deployment of other framework

PYTORCH
TensorFlow
mxnet

ONNX

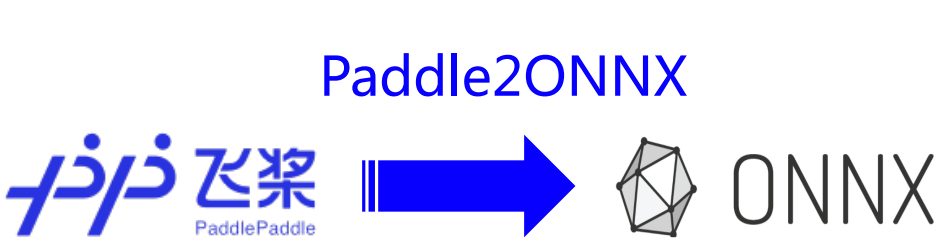
Paddle-Lite
Easy to deploy model on mobile, embeded, and IoT devices with high performance





Contents

Paddle2ONNX



Deployment

BITMAIN, cadence, CEVA, habana, HAILO, intel AI, NVIDIA, ONNX RUNTIME, MACE, Qualcomm, Rockchip, skymizer, SYNOPSYS, Tencent, tvm, TwinCAT 3, vespa, Windows

Enrich deployment methods of PaddlePaddle

Cooperation

OpenVINO™, ADLIK, NVIDIA, NVIDIA® Triton

Reduce development cost of cooperation

ONNX Export and Supported Operators

Export ONNX model by PaddlePaddle

```
import paddle
import numpy as np

model = paddle.vision.models.resnet50(pretrained=True)

toy_data = np.random.random([1, 3, 224, 224])
toy_tensor = paddle.to_tensor(toy_data)

paddle.onnx.export(model, 'resnet', input_spec=[toy_tensor])
```

Supported PaddlePaddle Operators

abs	arg_max
conv2d	flatten
matmul	multiclass_nms
lstm	nearest_interp
equal	gather
reciprocal	reduce_sum
roi_align	strided_slice
pool2d	prelu

... more than 80 operators

Supported Model Zoo

Image Classification	Object Detection	Image Segmentation	OCR	NLP
ResNet	SSD_MobileNet	UNet	**PPOCR**	TextCNN
DenseNet	YoloV3_DarkNet53	HRNet	DB	BiLSTM
ShuffleNet	YoloV3_ResNet34	Deeplab	CRNN	SimNet
MobileNet	YoloV3_MobileNet	BiSeNet		
VGG16	**PPYOLO**	GCNet		
SE_ResNext50	FasterRCNN	OCRNet		
InceptionV4				
SE_ResNet50_vd				
SqueezeNet1_0				
Xception41				
...				

More details: https://github.com/PaddlePaddle/Paddle2ONNX/blob/develop/docs/en/model_zoo.md

Join the ONNX family

Frameworks & Converters

Use the frameworks you already know and love.

To build an Open, Win-win, prosperous Ecosystem

More Information

Discover our opensource project

- https://github.com/PaddlePaddle/Paddle
- https://github.com/PaddlePaddle/X2Paddle
- https://github.com/PaddlePaddle/Paddle2ONNX

Official Website of PaddlePaddle

- https://www.paddlepaddle.org.cn/