Agenda

A. Brief Introduction
B. Ecosystem
C. Real World Use Cases
D. Developers Community
History of Baidu Deep Learning

2012 - DNN NLP, OCR models used in practices.

2013 - Baidu search Product: Phoenix Nest’s CTR based DNN prediction model launched.

2015 - The first ever NMT online translation engine launched.
Launched the STM-CTC based acoustic model.

2016 - PaddlePaddle went open source

2017 - Baidu News Feed recommendation system migrated to PaddlePaddle;
PaddlePaddle Fluid Released

2017 - Models in CV, Reinforcement Learning won championships in several international competitions;
Release of PaddlePaddle Suite

Release of PaddlePaddle Suite
PaddlePaddle 3.0 – Towards Maturity

PaddlePaddle 1.0

PaddlePaddle 2.0
Friendly python API, Released the core framework , model zoo and Paddle Book, Improved the ease of use and flexibility

PaddlePaddle 3.0
- July 2018
  Released PaddlePaddle 3.0
  Including functional components like EasyDL, AI Studio, AutoDL, VisualDL
- Nov. 2018
  Released PaddlePaddle Fluid 1.0
  Release PaddlePaddle Suite --Full-featured Deep Learning development kit for businesses and developers
Widely Recognized by the Government and Industry

The Only Deep Learning Technology and Application National Engineering Laboratory

PaddlePaddle has established a “Deep Learning National Team” with a number of domestic research institutions and universities.
Open Sourced Several International Competition Winning Models

The world's top technology level, leading the direction of deep learning technology

<table>
<thead>
<tr>
<th>Award-Winning Model</th>
<th>Awards</th>
<th>Awards</th>
</tr>
</thead>
<tbody>
<tr>
<td>PyramidBo Model</td>
<td>WIDER FACE (3 test subsets)</td>
<td>First place</td>
</tr>
<tr>
<td>Attention Clusters Network Model</td>
<td>ActivityNet2017/2018 kinetics</td>
<td>First place</td>
</tr>
<tr>
<td>Several Model Based on Faster R-CNN</td>
<td>Google AI Open Images-Object Detection Track</td>
<td>First place</td>
</tr>
<tr>
<td>PARL(Reinforcement Learning)</td>
<td>NIPS AI for Prosthetics Challenge</td>
<td>First place</td>
</tr>
</tbody>
</table>
Agenda

A. Brief Introduction
B. Ecosystem
C. Real World Use Cases
D. Developers Community
# PaddlePaddle Suite

Full-featured Deep Learning Suite with Comprehensive, Leading Technology

## Service Platform

<table>
<thead>
<tr>
<th></th>
<th>Modules and Components</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EasyDL</strong></td>
<td><strong>Visual DL</strong></td>
</tr>
<tr>
<td>Zero-based customized training and service platform</td>
<td>Visualization Tool for Training</td>
</tr>
<tr>
<td><strong>AI Studio</strong></td>
<td><strong>PARL</strong></td>
</tr>
<tr>
<td>One-stop development platform</td>
<td>Deep Reinforcement Learning</td>
</tr>
<tr>
<td><strong>AutoDL</strong></td>
<td><strong>EDL</strong></td>
</tr>
<tr>
<td>Network structure automation design</td>
<td>Elastic deep learning calculation</td>
</tr>
</tbody>
</table>

## Core Framework

<table>
<thead>
<tr>
<th>PaddleRec</th>
<th>PaddleCV</th>
<th>PaddleNLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intelligent Recommendation</td>
<td>Intelligent vision</td>
<td>Intelligent text processing</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Paddle Fluid</th>
<th>Paddle Serving</th>
<th>AutoDL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Features of PaddlePaddle Core framework**

<table>
<thead>
<tr>
<th><strong>Heterogeneous Computing</strong></th>
<th><strong>Parallel Training</strong></th>
<th><strong>Multiple Algorithms</strong></th>
<th><strong>Multi-End Deployment</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully supports for large-scale heterogeneous computing clusters CPU、GPU、DSP、FPGA</td>
<td>Supports multi-machine multi-thread asynchronous training and synchronous training mode</td>
<td>Personalized recommendation, image classification, semantic segmentation, face detection, machine translation, reading comprehension, lexical analysis, sentiment analysis</td>
<td>Rapid Deployment Multiple mobile end support</td>
</tr>
</tbody>
</table>
### Large-Scale Heterogeneous Computing Cluster

<table>
<thead>
<tr>
<th>Baidu AI Open Platform</th>
<th>Baidu Unified Deep Learning Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PaddlePaddle</strong></td>
<td></td>
</tr>
<tr>
<td><strong>k8s</strong></td>
<td><strong>Normandy</strong></td>
</tr>
<tr>
<td><strong>Docker</strong></td>
<td><strong>Resource Scheduling</strong></td>
</tr>
<tr>
<td><strong>Matrix Container</strong></td>
<td><strong>Resource Management</strong></td>
</tr>
<tr>
<td><strong>Hardware (CPU, GPU, FPGA, ...)</strong></td>
<td><strong>AFS Distributed file storage</strong></td>
</tr>
</tbody>
</table>

*Open source or open module*
Supports Parallel Training of Dense Parameters and Sparse Parameters

**Large-scale dense parameter**

- Data0
- Data1
- Data2
- Data3

**Ultra-Large-Scale sparse parameter**

- Data0
- Data1
- Data2
- Data3
- Data4

Computationally intensive tasks such as image classification and machine translation

Parameter synchronization mode: Synchronous Collective operation

GPU parallel training speed surpasses similar frameworks in mainstream tasks

CTR estimation, semantic matching, and other tasks with large data throughput

Parameter synchronization mode: Asynchronous large-scale sparse parameter server

CPU based ultra-large-scale asynchronous training is unique, supporting 100 billion scale parameters, hundreds of nodes parallel training
Keep Building the Most Complete Model Collection

<table>
<thead>
<tr>
<th>Scenario</th>
<th>PaddleRec – Recommendation</th>
<th>PaddleRec - CV intelligence</th>
<th>PaddleRec - NLP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Feed</td>
<td>Intelligent marketing</td>
<td>Search engine</td>
</tr>
<tr>
<td></td>
<td>Intelligent marketing</td>
<td>Video analysis</td>
<td>Machine translation</td>
</tr>
<tr>
<td></td>
<td>Video analysis</td>
<td>Medical imaging</td>
<td>Intelligent dialogue</td>
</tr>
<tr>
<td>Models set</td>
<td>Provision of many classic recall and ranking algorithms</td>
<td>Autonomous driving</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Covers all the cv application scenarios</td>
<td>Industry inspection</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fulfills mainstream NLP tasks</td>
<td>public sentiment</td>
<td></td>
</tr>
</tbody>
</table>

| Models set                    | DeepCTR                      | Image classification       | Chinese                   |
|                               | GRU4Rec                      | Object detection           | semantic matching         |
|                               | Text label                   | Face detection             | Comprehension             |
|                               | DeepCTR                      | OCR                        |                           |
|                               | Sequence semantic recall     | Semantic Segmentation      |                           |
|                               | Multi-view Simnet            | GAN                        |                           |
|                               | Metric learning              | Video classification       | Chinese semantic          |
|                               | Semantic Segmentation        |                            | segmentation              |

Application examples

- Baidu feed
- Haokan video
- Baidu Map
- Baidu OCR
- Baidu feed
- Baidu Nuomi
- Baidu
- Baidu translation
## Multi-platform Service Deployment

<table>
<thead>
<tr>
<th><strong>Paddle Serving</strong></th>
<th><strong>Paddle Mobile</strong></th>
<th><strong>Paddle Anywhere</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Flexible adaptation to multiple inference engines</td>
<td>• Multiple hardware platform support: ARM CPU, Mali GPU, Qualcomm DSP, FPGA</td>
<td><img src="image1.png" alt="Image of a car" /></td>
</tr>
<tr>
<td>• Compatible with mainstream engine TensorRT</td>
<td>• Fixed point quantization</td>
<td><img src="image2.png" alt="Image of a circuit board" /></td>
</tr>
<tr>
<td>• Inference API, lib library</td>
<td>• Low precision and efficient quantitative calculation</td>
<td></td>
</tr>
<tr>
<td>• CPU, GPU performance deep optimization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Forward pass specific optimization</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Paddle Mobile** supports multiple hardware platforms such as ARM CPU, Mali GPU, Qualcomm DSP, and FPGA.

**Paddle Anywhere** includes fixed point quantization, which enables low precision and efficient quantitative calculation.
Deep Learning Optimizations for today’s challenges

- Bigger the scale of data; More complicated the model structure; Larger feature size
- The model is getting more complicated
- Demanding industrial requirements
- Limited Memory & Video memory
- Requirement of Calculation Time is harsh

Deep Learning Efficient Decoding Method

- Quantification
  - Log Domain quantification
  - Product quantification
  - Binary network
  - Low precision operation
- Parameter sharing
  - Hash Net
  - Multi-Seed Random Hash
- Pruning
  - Pruning and Retraining
  - Dynamic Network Surgery
- Feature Optimization
  - Pyramid DNN

- Speed optimization
- Memory optimization
- Memory & Speed Optimization
PaddlePaddle Assistive Tools and Platform

<table>
<thead>
<tr>
<th>PARL</th>
<th>Visual DL</th>
<th>AutoDL</th>
<th>EasyDL</th>
<th>AI Studio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep</td>
<td>Visualized Deep Learning Tool</td>
<td>Automatic Network Structure Design</td>
<td>Zero skill required deep learning training and service platform</td>
<td>One-stop development platform</td>
</tr>
<tr>
<td>Reinforcement learning</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

One-stop development platform
PARL
Tools for Reinforced Learning

Won NIPS 2018 AI Prosthetics Challenge

Target Driven DDPG + Bootstrapping
One thousand of CPU + Single GPU
Visual DL
Visualize the overall Process of Training and Inferring

**Scalar**

![Scalar Visualization](image)

**Six components on visualization**

- Two SDK: C++, Python
- Supports ONNX

**Histogram**

![Histogram Visualization](image)

**ONNX network graph**

![ONNX Network Graph](image)
<table>
<thead>
<tr>
<th><strong>AutoDL Design</strong></th>
<th><strong>AutoDL Transfer</strong></th>
<th><strong>AutoDL Edge</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Network design</td>
<td>Create transfer model with small amount of data</td>
<td>Adapt to edge computing</td>
</tr>
<tr>
<td></td>
<td>Search for several neural networks with excellent performance and different structures</td>
<td>Network complexity optimization based on classic model, suits better for mobile deployment</td>
</tr>
<tr>
<td></td>
<td>Transfer pretrained models to new applications</td>
<td></td>
</tr>
</tbody>
</table>
AutoDL Design
Better than manual design

The network designed by AutoDL has the precision of 98% on CIFAR10 image classification dataset Surpassing classic network designed manually.
### Network Optimization for DL
- Based on classic DL models
- Optimization on network complexity, suitable for mobile devices

### Optimization Results
- Remain accuracy
- Highly compressed model parameters
- Run more AI tasks within the same computational capability

<table>
<thead>
<tr>
<th></th>
<th>Before suppressing</th>
<th>After suppressing</th>
<th>Suppression ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parameter amount</td>
<td>Precision</td>
<td>Parameter amount</td>
</tr>
<tr>
<td>SoundNet on ESC-50</td>
<td>13.00M</td>
<td>66.00%</td>
<td>0.07M</td>
</tr>
<tr>
<td>ResNet on CIFAR-10</td>
<td>11.17M</td>
<td>94.18%</td>
<td>0.82M</td>
</tr>
<tr>
<td></td>
<td>21.28M</td>
<td>94.72%</td>
<td>1.69M</td>
</tr>
<tr>
<td></td>
<td>23.52M</td>
<td>95.16%</td>
<td>3.97M</td>
</tr>
<tr>
<td>DenseNet on CIFAR-10</td>
<td>6.96M</td>
<td>95.13%</td>
<td>1.75M</td>
</tr>
<tr>
<td>VOC Object detection</td>
<td>26.29M</td>
<td>77.51%</td>
<td>20.94M</td>
</tr>
<tr>
<td>Goods identification for retailers</td>
<td>31.36M</td>
<td>84.55%</td>
<td>22.09M</td>
</tr>
</tbody>
</table>
AutoDL Transfer
Efficiency Modeling with small dataset

Network transfer
- Transfer pretrained networks to new applications
- Network design automation, less time consuming

Works better than classic models
- Need less samples
- Improve original model’s capability

AutoDL Transfer--Comparing with classic models

- Home decoration
- Bird classification
- Furniture classification
- Psoriasis Classification

Baseline | AutoDL Static model | AutoDL Dynamic model
Easy DL
Customized Platform for Training and Service

20k+ Models

Image dataset
Dialogue dataset
Voice dataset
Video dataset

Processing
Learning
Deploy
Service
AI Studio
One-stop AI Developing Platform

**Online Training Platform**

- **20k+ Programs**
- **46k+ Developers**
- **3700+ Dataset**

**Online Training Platform**

- Learning
  - Systematic tutorials
  - Coding examples
  - Classic datasets
- integration in the cloud
- Easy to use
- Efficient executions
- Free resources
- Developing
  - Python online coding
  - Predefined DL framework
  - Online training

**Large Scaled Open Dataset based on real world industry data**

- Video segments
- Recognition scenes for autonomous cars
- Machine comprehension
- Information extraction
- Knowledge extraction
- Traffic prediction
- Object labeling
Agenda

A. Brief Introduction
B. Ecosystem
C. Real World Use Cases
D. Developers Community
PaddlePaddle — Industry Application

Empower AI ability for industry with our partners 100+ cooperative corporations

Agriculture
Intelligent peach sorting
Save 90% manpower

Forestry
Worms inspection
Detection accuracy 90%

Industry
Smoking monitoring

Retail
Goods sales prediction
Decrease 30% wastage for fresh items

Human resource
Matching system by AI
5 times successful interview invitation

Manufacturing
Machine parts sorting
Double the efficiency

Petroleum
Prospection of petroleum

Telecommunication
Base station monitoring

Real estate
Building management
Save 20% electricity

Automobile
Failure prediction of charging stall
With accuracy of 90%
Partners of PaddlePaddle
Intelligent Sorting of Precision Parts

PaddlePaddle cooperates with dominant domestic enterprises performing quality assurance for rare-earth permanent magnet, to push the landing of deep learning on manufacturing sector.

PaddlePaddle assists enterprises with the landing of projects in the entire procedure.

- Analysis of Needs
- Technical Model Selecting
- Training
- Hardware Preparation
- Practical Testing

<table>
<thead>
<tr>
<th>Models of Semantic Segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>iCnet</td>
</tr>
<tr>
<td>0.1% Rate of mistaken sort (Under a mistake recognition rate at 5%)</td>
</tr>
<tr>
<td>25ms The inference speed of single part exceeds other deep learning frameworks at a rate of 20%.</td>
</tr>
</tbody>
</table>
Monitoring System for the Red Turpentine Beetles (AI insect Recognition)

PaddlePaddle Co-operates with Beijing Forestry University on "Intelligent Insect Monitoring Project"

PaddlePaddle + Baidu Map + Experts collaborated on this project

Data Collection → Model Training → Model Preparation → Capturing devices → Offline Recognition → Baidu Map → Distribution of insect population

Custom Model Optimization, Predominant Result in the field

<table>
<thead>
<tr>
<th>SSD</th>
<th>Models for Semantic Segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>The accuracy can reach 90%, which is similar to professionals</td>
</tr>
<tr>
<td>1 week VS 1 hour</td>
<td>Enhance the efficiency greatly from the manual assessment time of a week or so</td>
</tr>
</tbody>
</table>
Intelligent Candidate Matching System

CloudBrain adopts PaddlePaddle to invent an “AI HR”

- Takes full advantage of PaddlePaddle NLP capability in Chinese

- Significantly increase the rate of successful interview invitation for enterprises

<table>
<thead>
<tr>
<th>Textual Data</th>
<th>Training</th>
<th>Interview Invitation</th>
<th>Open position</th>
<th>Clicks</th>
<th>Result statistics collection</th>
</tr>
</thead>
</table>

- DSSM: Deep Structured Semantic Models
- 5倍: The increase in successful interview invitation rate
- 50%: The increase of click-through rate to the recommended posts

- 现在：企业HR发布JD后，智能招聘服务会给企业HR主动推荐合适的候选人
- 过去：求职者通过搜索JD和更新简历进行投递简历
- 传统招聘方法：时间成本和人工成本高且效率低下
- AI-EDU智能招聘服务：使成本更低，面试邀约率更高
Monitoring System for Floor Quality

**Raw Material**

**Mechanical Control, Channel Transmission**

200m²/day, 400m²/day

2x single-worker processing amount

**Tagging Platform**

**EasyDL Platform**

to perform model training

**Business processing & summaries and analyses**

**Intelligent Camera**

Flaw judgment and transmit data

Model Exporting

SDK integration
Advantages of PaddlePaddle in Enterprise Empowerment

**Leading technology**
- Officially supported ICNET model
- The precision can reach 99.5%
- The inference speed is 20% higher than similar products

**Reliability**
- The only Chinese Deep Learning Framework
- Performances with stability and reliability, thanks to the internal business lines of Baidu

**More understandings for domestic enterprises**
- Official Technical Support responses within 24h
- Official Chinese Community and documentation
- Follows AI project all the way through
- Published『HuangPu Plan』 Chinese AI talent training program

Across the globe, there have been many enterprises adopting PaddlePaddle and EZDL. 30% of Chinese Enterprises have already remarked PaddlePaddle as one of Top3 deep learning frameworks.


**Agenda**

A. Brief Introduction
B. Ecosystem
C. Real World Use Cases
D. Developer Community
PaddlePaddle has a relatively high vitality at GitHub open-source community, even higher than other frameworks in the same period.
PaddlePaddle Education

10k+ of active AI studio PaddlePaddle users

Published “Certification Standard for DL Engineering” with China software association

3 training courses for 300 university teachers from 100s of schools

Publications of books and training videos
PaddlePaddle Education Ecosystem

Applications in industry

Needs for professional personnel

Resource stocking for enterprises

Technology popularization

<table>
<thead>
<tr>
<th>Research of teaching</th>
<th>Teaching</th>
<th>Teaching resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher training</td>
<td>Online Course</td>
<td>Publication</td>
</tr>
<tr>
<td>Discussion and</td>
<td>College Course</td>
<td>Technical Articles</td>
</tr>
<tr>
<td>research Certification</td>
<td>Vocational Training</td>
<td>Chinese FAQ</td>
</tr>
<tr>
<td></td>
<td>Open Course</td>
<td></td>
</tr>
</tbody>
</table>

Development of Practical Personnel

<table>
<thead>
<tr>
<th>Contests</th>
<th>Interaction</th>
<th>Certificate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Practice Contests</td>
<td>Offline Interaction</td>
<td>Deep Learning Certification</td>
</tr>
<tr>
<td>AI Algorithm Contests</td>
<td>Online Answering</td>
<td></td>
</tr>
<tr>
<td>Campus Creativity</td>
<td>Directed Social Group</td>
<td></td>
</tr>
<tr>
<td>Contests</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Support of Practical Platforms

<table>
<thead>
<tr>
<th>Compute power</th>
<th>Algorithm</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster of 100 GPU</td>
<td>100 shared example projects</td>
<td>13 directions Around 30 classic datasets</td>
</tr>
</tbody>
</table>
Whampoa College - Training the First Batch of Chief AI Architects for Chinese Industry

Baidu Established the “Whampoa College” with National Engineering and Applications Laboratory of Deep Learning

- Face-to-face communication with Baidu Deep Learning T10 Architects
- Unlock the key point of implementing DL in Baidu’s core business know how
- Analysis of the typical case of the combination of business and deep learning in Baidu AI Cooperators in Ecosystem
- Help companies use AI thinking, AI tools, and methodologies to solve real business problems

[Launch] Way Of Deep
1.19

[Second] CV Fierce
3.9-3.10

[Third] NLP Leap
4.20-4.21

[Fourth] Enhance Together
6.1-6.2

Hard Core Technology
Experimental Course throughout the entire process

20+ DL Experts Waiting to Sail Together
Cooperations with Cloud Platform

Significant AI brand influence power, and a sharing-oriented attitude towards market resource

- PaddlePaddle is about to bring profit boosts to cloud services through market share expansion
- PaddlePaddle is devoted to developing a framework in line with needs of cloud providers
- PaddlePaddle is obliged to share cloud-end solutions with partners
- PaddlePaddle is willing to share part of promotion resources with all our partners

Share Customers

Partners will be listed as important cloud SP forging a great guidance for potential users

Co-Branding

PaddlePaddle willing to share promotion resources with all our partners

Technical Support

Expert assistance for deploying PaddlePaddle to cloud
The Deep Learning Framework that Truly Stems From Industry Practice

http://paddlepaddle.org
https://github.com/PaddlePaddle