**CRAIG:** Hi, I'm Craig Smith and this is Eye on AI.

**CRAIG:** This week, I talk to Ma Yanjun, General Manager of the AI Technology Ecosystem at Baidu, the Chinese AI giant. Dr. Ma leads Baidu's development of AI frameworks, including PaddlePaddle. We talked about how PaddlePaddle stacks up against other frameworks more popular in the West, about Baidu's development of large language models and the direction of AI research in China more generally.

**CRAIG:** Before we begin, let's take a moment to thank our sponsor, ClearML, an open-source MLOps solution. You can give them a try at clear.ml - tell them Eye on AI sent you.

**CRAIG:** Meanwhile, I hope you enjoy the conversation with Dr. Ma as much as I did.

**CRAIG:** Yanjun, can you introduce yourself and tell us how you got to Baidu where you went to school and that sort of thing. Give us some background on who you are. And what you do at Baidu.

**MA:** Yeah. So, I worked on natural language processing during my PhD. After my graduation from Qinghua university in China. I went to Ireland for five years, so I have a PHD degree in computer science from city university in Dublin. My PhD degree is 2009 and then I come back to Beijing to work for Baidu.

**CRAIG:** And you're in charge of the PaddlePaddle product group. Is that right?

**MA:** Yeah. I'm looking after PaddlePaddle development and products and its ecosystem. Yeah.

**CRAIG:** We're going to talk a little bit about PaddlePaddle, but it would help if you explained for those listeners that aren't familiar with machine learning frameworks, what exactly they do in the machine learning pipeline and then how PaddlePaddle came to be created.

**MA:** Sure. Yeah. So actually, PaddlePaddle is derived from parallel distributed deep learning. So, we normally call it a deep learning framework. It is used to develop deep learning algorithms. If you are an algorithm research scientist, then you can use such tools to efficiently develop your own algorithms, to try your own ideas using these tools, to implement your own algorithms using these tools, and deploy your own algorithms in real applications.

**MA:** It's more of a toolkit to help research scientists and engineers to efficiently develop algorithms and deploy them in real applications. So, developing deep learning application is more of a pipeline. You have to annotate data and then feed the data into a deep learning framework for training.

**MA:** Then after the training process, you get a deep learning model. So basically, this model is a condensed version of the data you have observed. Then you can use this deep learning frameworks to deploy your model into an application. And then you can have a real AI application.

**CRAIG:** Can you tell us a little bit about its development? And I'll just say, I remember when Eric Schmidt was in China., for the DeepMind Go competition with Ke Jie, he introduced TensorFlow, which is Google's deep learning framework. And at the time Schmidt believed that TensorFlow would be adopted very widely in China, but shortly thereafter, Baidu introduce PaddlePaddle.

**CRAIG:** So, I'd like to hear a little bit about PaddlePaddle development, and then the penetration it has in the Chinese market, how it compares with TensorFlow or PyTorch or other frameworks. And then how it's being adopted outside of China.

**MA:** Yeah, that's a really good question. Actually, inside Baidu, because there are many, developers inside Baidu to deploy AI applications.

**MA:** So, we started to develop such tools as early as 2012. At that time, many of our engineers spend a lot of time debugging their own software. At that time to implement a deep learning algorithm takes a lot of time. So, Baidu started to develop such tools and then it was open sourced in 2016. And TensorFlow was open sourced, I think, it is 2015. So, this is a timeline. Even now in China, TensorFlow still is used quite a lot among industries and inside universities as well. It is a very good framework.

**MA:** And also, besides TensorFlow, Meta, Facebook also has a deep learning framework called PyTorch. It is also very popular in China. There are some differences between these different frameworks. PaddlePaddle is more oriented for industrial use compared to others. And we dedicated a lot of effort to reduce the barriers to entry for individuals and companies.

**MA:** It's not just a deep learning framework. We developed a lot of end-to-end development kits and other different components and tool kits in order to help developers better use deep learning frameworks and to reduce the gap between such tools and real applications. So, I think this. How PaddlePaddle is different from TensorFlow or PyTorch.

**MA:** And as you asked, PaddlePaddle is quite a popular in China after several years development after it's open source, because it is open-source software, everyone can have access to it. It's grown quite quickly, simply because I think it is easier to use for Chinese developers.

**CRAIG:** Are there other deep learning frameworks in China that, that are competing with PaddlePaddle?

**MA:** Yes. As I mentioned, TensorFlow and PyTorch is also used quite lot. And in China, there are a few local, deep learning frameworks as well, open sourced in 2020. But PaddlePaddle was open sourced quite early.

**MA:** So, it's in a better shape compared to other local deep learning frameworks.

**CRAIG:** And you said that PaddlePaddle is more suitable for industrial use. And are you implying then that PyTorch and TensorFlow are more useful in a research setting? What's the distinction you're making.

**MA:** PyTorch and TensorFlow are quite popular among AI researchers. They require quite a lot of AI background. But In China, many of the developers do not really have much AI background. They are trying to use AI in their work, but they do not have that much AI background. So, in order to increase the use of AI in different industry sectors and to better serve the needs of a wider community of developers, PaddlePaddle provided as I mentioned earlier, quite a few low threshold toolkits, these toolkits easier to use. So, this is the distinction, rather than it's not on the side of research and industry, it is more on the ease of use of such tools. So, it can be used by a wider community.

**CRAIG:** Yeah. One of the things about these frameworks is when you start using a framework you then are in the ecosystem of the company that built that framework even if it's open source, in much the way that with, Google, Microsoft, and Apple have their word processing, spreadsheet presentation software. And if you use one, it's easier to move between them within the Microsoft ecosystem or the Apple ecosystem.

**CRAIG:** Is that a disadvantage for PaddlePaddle outside of China? Because certainly in the United States, Google and Microsoft dominate AI tool space.

**MA:** Yes, outside China, PyTorch and TensorFlow are used much more than PaddlePaddle, but inside China currently PaddlePaddle is quite popular already. Even though PaddlePaddle is open source, we have both English and Chinese documents. Also, we do know there are quite many developers outside China using PaddlePaddle as well. But in comparison, TensorFlow or PyTorch have more developers outside of China.

**CRAIG:** I'm familiar with the very early days of Baidu. I just had Andrew on the podcast a week or so ago.

**CRAIG:** And Baidu was early in deep learning in China. A lot of the main deep learning companies in China came out of what they call the Baidu brain trust. Can you talk a little bit about how Baidu is developing deep learning and how PaddlePaddle is aiding that? Whether you're particularly focused on any industry verticals - manufacturing or healthcare - and how you see deep learning, developing in China.

**MA:** So basically, Baidu has a lot of demands on deep learning frameworks, inside Baidu's own development process. When we develop such tools, basically it is driven by the internal use of it at the beginning.

**MA:** For example, Baidu's own AI applications, including search, recommender systems and conversational AI speakers, et cetera. These products require a deep learning framework to have high speed, low latency, easy to use for algorithm developers, these kinds of different properties that a deep learning framework must have.

**MA:** So, at the beginning, we consider the flexibility and the ease for production, those two sides of this framework. So first of all, it must be flexible for developers to use because there are many different applications.

**MA:** Only if it can be flexible, then it can be used to develop a large amount of different applications. That's the first point. Then in real production it must have high speed and low latency. So, we test this deep learning framework in Baidu extensively. Then we can have a better open-sourced project.

**MA:** If at the very beginning, if you just open source, it and nobody uses Then it's very difficult for developers to use such tools. So, this is how we develop it. A recycle of usage and technological development.

**MA:** Then it can evolve quickly and have better performance and user experience in different industry sectors outside Baidu. Currently in China, there are many different industry sectors adopting AI. To name a few, for example, transportation has been adopting AI quite a lot, to monitor traffic lights and to improve traffic efficiency etcetera. And then in China, there are many manufacturing companies. So, these companies are very interested in improving their productivity because for manufacturing companies, basically they have two points they really care about first of all, to lower their costs.

**MA:** Secondly to improve their efficiency in production. So, most of these sectors require a lot of labor and it may be hard for many of them to recruit that many workers to work. AI can help in such cases.

**MA:** And beside these two, financial services use a lot of AI already because they are quite similar to internet companies. They have more data, and their services require lots of AI.

**MA:** And besides these, smart cities, even agriculture Nowadays, we have some adoption already. For example, in China, there is a county called the Shouguang Shandong province. That's a large base for China's vegetables. So, there are lots of vegetables sheds to be maintained during the whole year, not just in summer or spring.

**MA:** In order to better manage such vegetable bases, they started to deploy AI. For example, only two people and one app can manage dozens of vegetable sheds. So, we can see that in China, actually, AI deployment is quite diverse. And you can also say it is quite fragmented as well. Different sectors have their own demands, and they use it for different purposes.

**CRAIG:** Yeah, that's actually very interesting about the vegetable sheds. Is the AI monitoring water and humidity and temperature and setting those variables? Is that how it's being used?

**MA:** Not just this kind of usage. it can also be used to monitor the growth of different vegetables.

**MA:** For example, for tomatoes, whether it's ready to be picked. Or other different points the farmers care about. There are quite a few points farmers care about, for example, which plant is ready to be picked. So, it can improve the efficiency.

**CRAIG:** Do you have a sense at all of whether deep learning has penetrated Chinese industry to a greater degree than it has in the United States, do you have any feeling for that?

**MA:** Yeah, I do. So actually, based on my observation, it really depends on the industry sector itself. It varies quite a lot. Internet already hundred percent using AI, and financial services. I think there is a very big fraction using AI. It is more than half.

**MA:** I think more than half of the financial companies are using AI already. But when it comes to other industry sectors. For example, smart cities must use AI, but not every city is getting smarter. Only big cities spend money to make it smarter currently. So big cities already using a lot of AI in their management.

**MA:** When it comes to manufacturing, it's a much more complex situation. So, for top companies, in manufacturing, they start to focus on their efficiency using AI already. In many cases, they use AI for different activities in their process, they have a very long pipeline.

**MA:** Quality inspection requires AI more than others. And it also depends on what they are producing. When it comes to others like transportation, I think there is a strong trend to adopt AI. So, from my feeling, the recent two or three years I can feel there is a momentum to a faster adoption.

**CRAIG:** You mentioned transportation. Baidu is very active in the robotaxi space. can you talk a little bit about that? Or autonomous vehicles in general?

**MA:** Yeah, currently robotaxi is already in operation in many different sites inside China. As it accumulates more data to train AI system better together with the infrastructure upgrades, for example, the roads become more and more intelligent, then overall its performance will be much better.

**CRAIG:** PaddlePaddle features a number of toolkits and libraries. Can you talk about some of the more popular ones, Paddle Detection and Paddle SEG and some of the other toolkits and libraries?

**MA:** Yeah, this is the strong side of PaddlePaddle. We have quite a number of different development tool kits based on PaddlePaddle. The main point of these toolkits, it can significantly reduce the development costs because it is much easier to use.

**MA:** For example, for deep learning framework, we have about 1000 APIs. That's a lot of interfaces for developer to use. It is very difficult to master all these different APIs but for these kinds of end-to-end development kits it'll normally have 10, 20 APIs to use, focused on a specific scenario.

**MA:** For example, Paddle SEG, it can be used for segmentation of images and for Paddle detection, it can be used for. Object detection. So different tools have their own applications scenarios which fits the developers’ requirements as well. When a developer is trying to look for a tool kits, he knows what he wants to look for. So, these tools have only 10, 20 APIs and also have the whole pipeline of AI development from data processing to training, to model compression, to the adaptation to different hardware, and then how to deploy them in different systems, for example, in windows or in Linux operating system or on an Intel chip or on a Nvidia chip, there are many different complexities in order to deploy these AI technologies. These tools are aiming to resolve such complexities. So, the developers can use these tools very easily. That's why it is quite popular in China right now to use PaddlePaddle they are using such toolkits and not just the tool itself.

**MA:** These tools also contain many different pre-trained models. These pre-trained models have their own features as well. The main feature is we have a very good balance between model precision and speed performance. So normally for AI system, many of them, if you have high accuracy or high precision then it's normally very slow. Then if it is very fast, it has poor performance in precision or accuracy. So how to balance these two is very important. We did a lot of work to do such things in order to maintain its precision performance. And in the meantime, to reduce the model size, we only keep a very small fraction of the model to be deployed on real applications. Then the developers are very happy about such models. They can deploy them on a camera or maybe on a cell phone. The model is very small and very fast, then it is easier to use. So, this is our experience in building PaddlePaddle

**CRAIG:** have you worked on. Baidu's large language model, Ernie Titan?

**MA:** Yeah. Sure. PaddlePaddle is an important part in training such models. These models are very hard to train, so it must have a strong, deep learning framework to help them. I'm looking after the distributed training of massive models.

**MA:** Basically, Paddle’s name is PArallel, Distributed, Deep LEarning. So, at the very beginning, we have a very strong focus on processing large amounts of data and to train very big models. Actually, Baidu search and the recommender system use a lot of large models already many years ago. For a recommender system, they normally require a technology inside a deep learning framework.

**MA:** They call it a parameter server technology. It can help to train sparse models that can be used in recommender systems and search. But for models like Ernie or Bert these kinds of models, they are dense models. The training of such models require much more effort. It's more complicated.

**MA:** So, for PaddlePaddle, we developed several different kinds of technologies to better serve the training of such a model.

**CRAIG:** Ernie, the largest version of that has 260 billion parameters, is that right? Can you talk about how that model is being used and are you developing even larger dense models?

**MA:** This model is already very large. It is one single model that has 260 billion parameters. It is the largest already. But nowadays there are many different models that is not one single model. They use quite a few different models and merge them. The parameters may grow by even more. So it isn't, in my opinion, it's not the bigger, the better actually. Currently we are trying to exploit many different ways to better use such big models. First of all, it can be used directly.

**MA:** The inference of such models, when I talk about inference, just use it online. It is a big challenge because it is too large. It takes many seconds or even minutes, quite a few minutes to get one result. In such cases, it can be used for scenarios like text, generation, image, generation.

**MA:** It doesn't require that you have to real time get the results. So that's okay. If you use big models for such tasks. But for others, for example, if you use it for search or use it for recommender systems, it is not acceptable. If you can only get results back in a few minutes. Then we did a lot of work to streamline big models and to make it smaller, for example, to make it only have 10 billion parameters, not 260 billion. So, we can do a lot of work to make the model smaller. In the meantime, the accuracy is acceptable.

**MA:** But the speeds has been improved quite a lot. So, in search or other different applications, a model size like 10 billion could be a good size. For models like 260. It is very difficult to use it in real time, but it can be used for generation, et cetera.

**CRAIG:** Actually, that's one of my questions.

**CRAIG:** These ever-larger models are fascinating, but more from a research point of view, I don't yet see, other than GitHub and Microsoft's co-pilot I don't see any real practical applications. Are there any practical applications emerging from the very big models in China?

**MA:** I do see quite a few trends of using big models.

**MA:** Baidu's Ernie is already deployed in Baidu search online. So, this is already deployed. It must help internet companies, AI companies for their own business. That's already for sure. Then for financial services, they have similar demands for big models as internet companies.

**MA:** So, it can be used already. In China, we have seen from Baidu's ERNIE models and others we see that it is very interesting for artists, or writers. They start to use such models. They use it to generate different contents, AI generated contents.

**MA:** It has its own uniqueness. Recently we had a competition inside China among students to have their own ideas using big models. Many of them have very interesting ideas about how to use big models to serve their own purposes.

**MA:** Many of them are about arts or about maybe advertising to have better slogans for example, in their products. So, in many cases, the AI generated contents have a lot of interesting points that we have never thought about. It gives some surprises sometimes, but it does help in this space. So currently this is how it has used.

**MA:** I think it is emerging already. And I think if we can reduce the model to a smaller size and to fit a single machine, if we can do that, I think it will have much more uses in the future.

**CRAIG:** From your vantage point, you can see not only where AI is being deployed, but what kind of research, not only researchers in Baidu, but across China. If they're using PaddlePaddle, a framework then where the research is headed. Can you give us an idea what the trends are in deep learning research?

**MA:** AI had quite a few different branches over the past few years.

**MA:** For example, some people are working on speech and languages. Some people are working on computer vision in different areas, but in recent years, we can see these areas are converging. Different areas start to work on similar topics. For example, they work on cross model learning. Even though these contents have their own modality, currently deep learning is heading toward cross model learning.

**MA:** So, this is one of them. Another one we see knowledge graphs more and more used in deep learning process. So basically, previously a deep learning system basically deal with raw texts or raw images. You don't really input any knowledge into the system. It just does self-supervised learning, to gather rules outside the data.

**MA:** But in recent years, we also see that knowledge graphs is helpful to such systems. So, knowledge graph plus deep learning has been more and more popular in research. As AI is more and more used in different scenarios, different scenarios are posing different challenges to deep learning frameworks.

**MA:** For example, if you want to make some applications more real time, then you have to work on how to reduce the model size and to make it run very fast on specific hardware, for example. There is a lot of opportunities of innovation in this space as well because AI engineers normally don't know that much about industry sectors.

**MA:** Manufacturing. They have their own knowledge, quite lots of knowledge, but AI engineers, AI scientists don't really know. If these two groups of people meet together, there is a lot of possibilities to move AI forward. For example, Paddle helix, this is AI plus life science. So, these areas, they do need a lot of AI technology, but they have their own problem how to adapt to AI to better solve such problems. So that's a very specific scenario. This area has a lot of space opportunities as well. So that's the research directions.

**MA:** In the meantime, at least in China, we still see lots of industry sectors. A lot of engineers don't know how to use AI. They know AI can help a lot, but they don't really know how to use it. So, this is still a very big problem for AI penetration into different industry sectors.

**MA:** In my opinion, it's very important as well for such platforms to further reduce the barrier to entry for individuals and companies, including big models. If we have big models, actually these big models are a way of reducing the barrier. Because if you have this foundation model, you can just adapt this model to a specific scenario. Then it will work. If you don't have this one, you need to develop everything from scratch. So big model is hard for everybody to use, but if it can be foundation model it become an infrastructure, then many different companies can use it.

**MA:** This is also a way of reducing the barrier. And also, for many industry sectors, many people don't really write a lot of code. They just want to reuse AI. So, if you have a GUI, for example, have a graphical user interface that people can just click the mouse and don't really write codes or very little codes, for example, then it'd be easier for them to use.

**MA:** So how to make the AI process more automated rather than you write a lot of code to deal with the whole process, these areas are also very important for AI to better deployed.

**CRAIG:** That's it for this episode. I want to thank Dr. Ma for his time. I also want to thank our sponsor, ClearML. Please take a moment to visit them at clear.ml

**CRAIG:** And remember, the singularity may not be near, but AI is about to change your world, so pay attention.