AI: From Theory to Industry

Iuri Frosio[0000−0002−7230−4287], Giuseppe Fiameni[0000−0001−8687−6609], and
Piero Altoè

NVIDIA
{ifrosio, gfiameni, paltoe}@nvidia.com

Abstract. We live in the rising era of Artificial Intelligence (AI), which is revolutionizing the world we live in, from the advent of autonomous vehicles to the possibility of performing automatic medical diagnoses, and beyond. Nonetheless, the birth of new AI technologies and their adoption in the real world is not always a smooth process. Based on our working experience at NVIDIA, one of the leading companies in the AI world, we report our recommendations for the development, deployment and adoption of new AI technologies in industries, from a technical point of view.

Keywords: Artificial intelligence· Technology transfer· Machine learning.

1 Introduction

When picking a new research project, any researcher should answer a set of questions to justify their choice. Given the large amount of possibilities offered by the advent of deep learning, machine learning, or more generally AI in all its declinations, providing satisfying answers to this set of questions is nowadays even more important. Without claiming to be exhaustive, as in practice we only touch the technical aspects of the art of picking new research projects, without considering social, ethical or philosophical aspects, we describe our experience with the development and deployment of novel AI technologies at NVIDIA, one of the leading companies in the AI world. We provide a few recommendations to follow, based on one, main aim: what we study and develop in research, should eventually be used in the real world or spur new research activities.

In the next Section, we introduce the set of questions that should be answered before starting a new research project. In the following one, we use a real case [1] to illustrate how we applied these principles during the development of a research project at NVIDIA.

2 Questions for new AI research projects

As researchers, we are often driven by curiosity. Thus, the first question we should answer about a new research project, is the following one:

- **Question #1: is it interesting?** The research project we pick should aim at discovering new scientific knowledge, and tickle our curiosity. A project that does not unveil new insights or points of view is hardly an interesting one. Fortunately for AI researchers, the entire AI world is full of opportunities for the development of novel algorithms, hardware, methods, and technologies — in other words, it's easy to find topics that stimulate our curiosity while consequently keeping our motivation high. On the other hand, the AI space is also particularly crowded and characterized by extremely short publication times: a careful literature analysis is more than mandatory to guarantee that the research project we want to invest our time in, is really a novel one.

However, picking an *interesting* project is not sufficient to guarantee that the it will be deployed in real applications. To this aim, we believe a second question has to receive a positive answer:

- **Question #2: is it relevant?** An open problem is an important one if the market, the consumers, and/or some industries show interest in it. This kind of information can be collected through surveys or, better, from a close interaction with industry. Researchers rarely have direct access to this kind of information, although they often speculate about the future of a new technology. In this case, it is important to strike the right compromise between being visionary and understanding the need expressed by industry and its willingness to invest in a given, new technology.

A third fundamental question requires a positive answer, and this is about the feasibility of the project. Although positively answering this question is a task for researchers, the constraints are generally given by the demanding industries. More formally, the third question is:

- **Question #3: is it feasible?** Researchers have to answer this question on the basis of their knowledge of the existing solutions, their limitations, and innovations that could to be reasonably introduced to complete the project. On the other hand, the specific constraints, for example the maximum amount of computational power or energy or the maximum latency admissible to complete a task, should be collected directly from the recipient industries / final costumers. While assessing the feasibility of the project, researchers should also keep in mind that a high level of technology readiness (TRL) [2] and support for operational standards may be highly appreciated by recipients that intend to employ their results into real products.

Last but not least, researchers tend to be naturally ambitious and are often requested to look not one, but two steps ahead with respect to future technological developments. Thus, a last question is the following:

- **Question #4: is it all?** The identification of the aim of a research project is as important as the identification of the limitations of the newly proposed technology. Therefore, researchers should be careful identifying such limitations in an early stage, together with potential threats that could invalidate the output of the project, and proactively identify future development directions. The discussion about these limitations together with the recipient industry / customers should also be performed as early as possible, to verify that the predicted outcome is satisfying for the recipient.

3 Vision based cheat detection in videogames

Many research projects in NVIDIA are carried out with the interns (see the left panel in Fig. 1). The scientific knowledge and know-how acquired during the development of these projects is transferred to NVIDIA and the large (research) community through the writing of papers, by including new technologies in GPUs and libraries, and so forth. In parallel, field operation personnel can collect relevant industrial problems and constraints, while providing consultancy and support to other industries.

In the specific case that we report here as example, we tackled the problem of visual detection of cheating activities in videogames. We identified this problem as *relevant* (question #2), after collecting feedback from companies operating in the videogames space and from public reviews showing that gamers are often annoyed by cheaters and prone to leave the game in case they meet one [1].

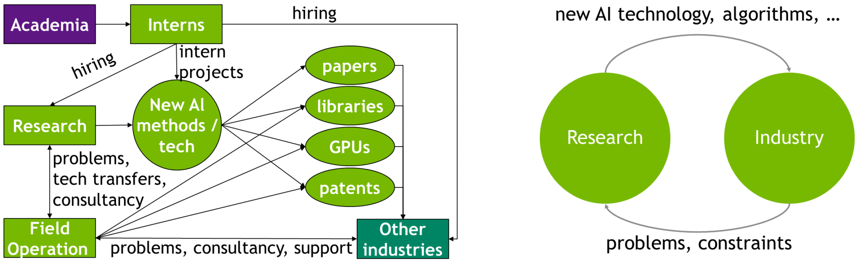

Fig. 1. The left panel represents one possible workflow of the development and deployment of novel AI technologies, from research to industry. The right panel illustrates the the effective interaction model between research and industry for the industrial transfer of AI technologies.

We also identified the problem as *interesting* (question #1), as existing anti-cheat solutions never leveraged the power of deep learning for visual cheat detection before the development of our method; furthermore, during our investigation, we were also able to explore the problem of the identification of out-of-training-distribution data at inference time, which is scientifically relevant for the deployment of robust machine learning methods in many other fields.

During the development of our method, we analyzed its *feasibility* (question #3) by taking into account constraints such as the minimal additional latency required to guarantee a high-quality gaming experience, as well as the need for privacy that does not allow the transmission of screenshot images. These led us to the development of a lightweight deep neural network for cheat detection, that can run on the local machine without adding a significant latency and without requiring data transmission.

Finally, our previous research experience in the field of adversarial attacks, suggested us that any anti-cheating deep neural network could be easily fooled by coders with knowledge of adversarial attacking technique, therefore we trained a robust network using an adversarial protection method and successfully verified its sufficient accuracy even under attack (question #4).

4 Conclusion

We have presented our point of view on the development of effective AI research projects that are aimed to be deployed in real world applications. Without touching ethical, sociological, or philosophical aspects of AI, that should anyway be discussed and taken into consideration, we suggested that a successful AI project should answer positively to the set of four questions presented here. In practice, this requires a two way interaction between research and industry as the one represented in the right panel of Fig. 1, where open problems and constraints are collected by research from inputs coming from the industry, so that research can develop and ship effective and useful AI technologies.

References

1. Jonnalagadda, A., Frosio, I., Schneider, S., McGuire, M., Kim, J.: Robust vision-based cheat detection in competitive gaming. Proc. ACM Comput. Graph. Interact. Tech. **4**(1) (apr 2021). https://doi.org/10.1145/3451259, https://doi.org/10.1145/3451259
2. Mankins, J.C.: Technology readiness levels-a white paper (1995)