infinitybatch

survey

Axel Ország-Krisz Dr.

Richárd Ádám Vécsey Dr.

What is infinitybatch?

Infinitybatch is an open source solution that helps deep learning developers to train with bigger batch size than it could be loaded into GPU RAM.

The core concept of the idea comes from the fact that GPU time is expensive and the usage of own GPU cluster or a cloud based GPU service has to be optimized to be cost efficient. Furthermore, developers and researchers regularly have limited access to GPU. However, CPU based training mostly allows higher batches than a normal GPU could provide, it is much slower. Infinitybatch helps to use GPU during training with bigger batch size thanks to the special unloading and uploading process that manages the GPU RAM to avoid memory overrun.

The name of infinitybatch is a little deceptive for the first catch. It does not mean we could really use batch size that is infinite. The highest limit of batch size will always be the length of the training dataset. Secondly, infinite batch cannot exist since the resources are always limited, such as GPU RAM, CPU RAM, GPU speed, CPU speed, storage, FSB bandwidth or the bandwidth of a network.

We only provide to use bigger batch size than GPU could be provide during a normal training process. That's why inifinitybatch does not necessarily mean big batch. For example, when a special training process on a common GPU can be trained with batch size 4, the developer could be train with batch size 32 with infinitybatch. However, if the developer want use really big batch, it can be done with our solution.

We do not want to compete with GPU cloud services, but we can help the users to use services smart and well prepared. Infinitybatch promises only to expand the limits. This is a great and yet simple idea that helps researchers and developers to test a lot of ideas with different hyperparameters on machines with limited resources and train the finalized idea on a dedicated machine or cloud service. Also it can be a good solution for students, who have mid range GPU hardware, to try different batch sizes and learn and understand the concept of deep learning model development and the behavior of the training process.

Survey

We decided to take a survey about GPU-RAM and Batch size with some added extras such as deep learning frameworks and source of knowledge. The main targets were programmers, developers and researchers. In the public call, we speak about the survey with the focus on connection between deep learning and GPU utilization, we didn't speak about infinitybatch. The cause behind this behavior is we don't want to influence the responders. There are questions that affect negatively on a survey accuracy. We made this survey via the internet with the cooperation of a very small population. That's why we tried to avoid any biases, manipulation or influence as possible to get the most accurate result.

We published the public call on the following sites under own real name or under own personal account:

[18. 08. 2020.]

- Facebook
  - ↳ Artificial Intelligence and Machine Learning [private group]
  - ↳ Udacity Alumni Network [private group]
  - ↳ PyTorch Scholarship Challenge Program from Facebook [private group]
  - ↳ Udacity Community - Europe [private group]
  - ↳ Udacity Scholarship Participants [private group]
  - ↳ Udacity AI Hungary [private group]
  - ↳ Developer Circles: Machine Learning [private group]
- Slack
  - ↳ Udacity Alumni > nd-deep-learning

[19. 08. 2020.]

- Slack
    - ↳ Udacity Alumni > random

[20. 08. 2020.]

- Reddit
    - ↳ artificial
    - ↳ MachineLearning
    - ↳ pytorch
    - ↳ tensorflow

[23. 08. 2020.]

- Reddit
    - ↳ deeplearning

This was the full version of public call:

Hi my fellow programmer,

I am programmer who work in the field of deep learning. Me and my colleagues usually make lectures and statistics about fields that we are interested in. That's why we created a survey about the connection between deep learning and GPU utilization. If you have 3-4 minutes, please help us by answering our questions.

https://forms.gle/2L6nUfxSN5KyVo786

Thank you for your time and answers,
Axel / Richard

Short description about survey

The survey is started with a short description. Its goal is to hide the survey's real focus. Maybe it seems deceptive that it is the only way to minimize the distorting effect. When people know we create a batch based research, it is a great chance that their responses adjust to their supposition about hypothetically perfect answers to the problem of batch size instead of the real life scenario.

Dear Responder,

We created a survey about the connection between deep learning and GPU utilization. We are programmers who work in the field of deep learning. Beside the work, we are learning. This small project is a part of a big journey. We usually make lectures and statistics about fields that we are interested in. Please help us by answering our questions. We would like to know what our fellow colleagues from around the world think about this topic. There are no right or wrong answers. If you help us, we will be very-very happy.

Thank you for your time and answers.

Cheers,
Axel and Richard

Questions

1. **What describes you better? You are a... [O]***

   - programmer
   - developer
   - researcher
   - gamer
   - student
   - hobbist
   - none of the above

2. **Do you have your own GPU? [O]***

   - Yes
   - No

3. **What is the memory size of your GPU (in Gigabytes)? [O]***

   - I don’t have own GPU
   - 2 or less
   - 4 (more than 2 but no more than 4)
   - 6 (more than 4 but no more than 6)
   - 8 (more than 6 but no more than 8)
   - 10 (more than 8 but no more than 10)
   - 12 (more than 10 but no more than 12)
   - more than 12

4. **How do you usually train your models at home? [O]***

    - I don’t use my personal computer to train.
    - CPU
    - single GPU
    - GPU cluster
    - cloud service
    - other

5. **How do you usually train your models at workplace / school? [O]***

    - I don’t use my workstation / school computer to train.
    - CPU
    - single GPU
    - GPU cluster
    - cloud service
    - other

6. **Do you try to different batch size during training process? [O]***

    - Always
    - Usually
    - Rarely
    - Sometimes
    - Never

7. **How often do you get memory error (such as out of CUDA memory or any other memory overrun) message during training? [O]***

- Always
- Usually
- Rarely
- Sometimes
- Never

8. **Which framework do you use? [M]***

- Caffe
- Chainer
- DL4J
- Keras
- MATLAB
- Microsoft CNTK
- MXNet
- ONNX
- OpenVINO
- PaddlePaddle
- PyTorch
- TensorFlow
- Theano
- other

9. **How familiar with these libraries? [R]***
never heard / never used it / just tested it / use it rarely / use it regularly

- Aequitas
- Seaborn
- Scikit-Learn
- SciPy
- Matplotlib
- NLTK
- NiBabel
- NumPy
- OpenCV
- Pandas
- PyBigData
- Pydicom
- PyTessy
- PySpark
- Tensorboard

10. **Do you think finetuning the hyperparameters matters for the performance of a model? [O]***

- Yes
- No

11. **Do you finetune hyperparameters? [O]***

- Yes with manual settings
- Yes with automated way
- No

12. **Where are you from? [O]***

- Europe
- North-America
- Mid- and South-America
- Africa
- Australia and New Zealand
- Middle-East
- South and West Asia
- North and Middle Asia
- I prefer not to say it.

13. **Where did you learn about machine learning, deep learning? [M]***

- College / University
- Course at your workplace
- Online course
- alone with DIY learning (eg: book, Youtube)

14. **Where did you meet first with machine learning, deep learning? [O]***

- College / University
- Course at your workplace
- Online course
- Online video sharing site
- Book
- none of the above

15. **What are your favourite techniques to optimize the performance of a model? Please, describe them. [F]**

Method

The questions can be different type like only choice [O], multiple choice [M], a rating matrix [R] or free-response questions [F]. Most of them are mandatory that a “*” sign represents after the question type marker.

First of all, our survey isn’t representative, since the filling is voluntary and we made a public call via internet. It seems a good assumption that people who are committed to the community is overpopulated in the responders. This is not a bad situation, but we must to remember, there are lonely wolf programmers around the world who does not like using social media or any other community channel for programmers. We did not have the resources to get demographically absolutely representative data.

However, we are developers, data science is the part of our daily routine. That’s why we cannot miss to talk about the lacks or weakness of survey. The goal of this survey to show the trends in the world. We think despite of limitation are mentioned above, this little survey is capable of filling this position.

Question 1

We are curious how people describe themselves. This question means nothing standalone, but it can refine the whole picture. If a lot of none of above occur, we can remove those answers for avoiding distortion.

Question 2

Our project’s persona has own GPU at home or at the company / school. This question measure the first target group.

Question 3

Size of GPU-RAM might be interesting. If a lot of user has limited resources, it is good for tools like infinitybatch. However, a user with huge GPU-RAM can also profit from our solution.

Question 4

Infinitybatch has very different use cases. We can add something into training. The biggest advantages pop up when user uses own GPU or GPU cloud. The only group lagging behind is the CPU-based trainers. This question is about home computer.

Question 5

Infinitybatch has very different use cases. We can add something into training. The biggest advantages pop up when user uses own GPU or GPU cloud. The only group lagging behind is the CPU-based trainers. This question is about work or school computer.

Question 6

It seems unnecessary to ask about the frequency of batch size modification, we are very curious the portion of “never” answers.

Question 7

Infinitybatch is very useful who get CUDA memory overrun messages usually at least since it helps to avoid this type of error.

Question 8

We want to know what is the ratio between PyTorch and every other popular framework. We made an other statistic about this topic that is based on github searches.

Question 9

This question focuses on the popular libraries. However, it contains two tricky answers. First is “PyBigData” that does not exist. There are pip packages with similar names, but this is a fiction that can help filter responders who want to make shiny and proudable responses. The second tricky answer is PyTessy. It is one of our projects and we put it just for fun.

Question 10

The purpose of this questions is to lead up the next question.

Question 11

Manual finetuners are the member of our target audience. Who use automatic method might find infinitybatch interesting as well.

Question 12

The location question is always sensitive. However, most of people live in freedom, there are a lot of causes while people do not want to share their location. That's why we made answers that embraces huge areas and a unique answer for people who do not want to answer this question at all.

Question 13

It reveals the correlation between the source institute of learning and the habits. This question is not related to our project but we are curious about answers, since we learned from online courses and DIY.

Question 14

The previous question is a multichoice one, so the first source should be concretized

Question 15

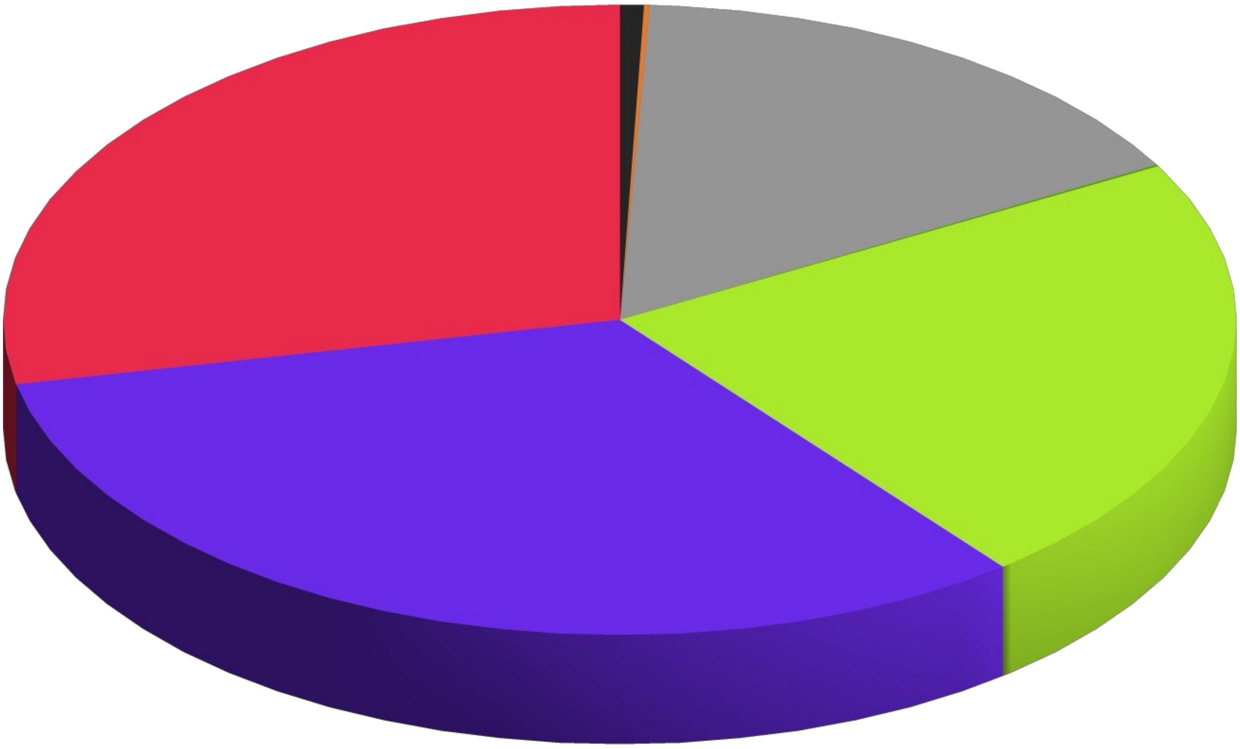
At the beginning we wanted to create a question about responder's favourite optimization techniques. The responders could answer on a scale. In the end, we dropped the idea and switched the question into a non mandatory, free response question. Collecting and ordering answers are harder, but the data is more valuable. Offering a lot of options could confuse the responder and force him into a very strange situation where they could rate higher the less used techniques. The unique answers of every technique could be correlated even if there is no real correlation. Nobody wants to see that they use mostly one or two techniques. Or maybe they are deceptived by scale and made them connect a higher frequency of usage with the given techniques than the method occurs in real life. We want to avoid any type of conflicts like these and try to get accurate answers as possible.

Survey closed: **25. 08. 2020. 12:00 CEST**

Total number of filled survey: **1,439**

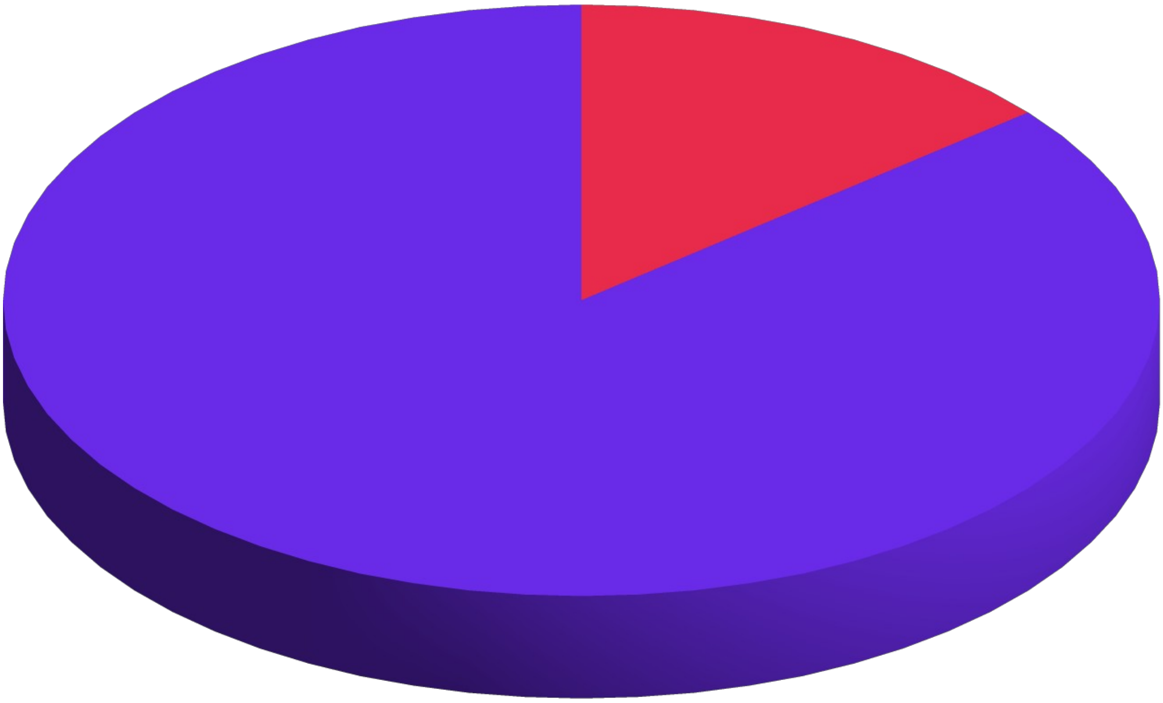
Result

What describes you better? You are a...



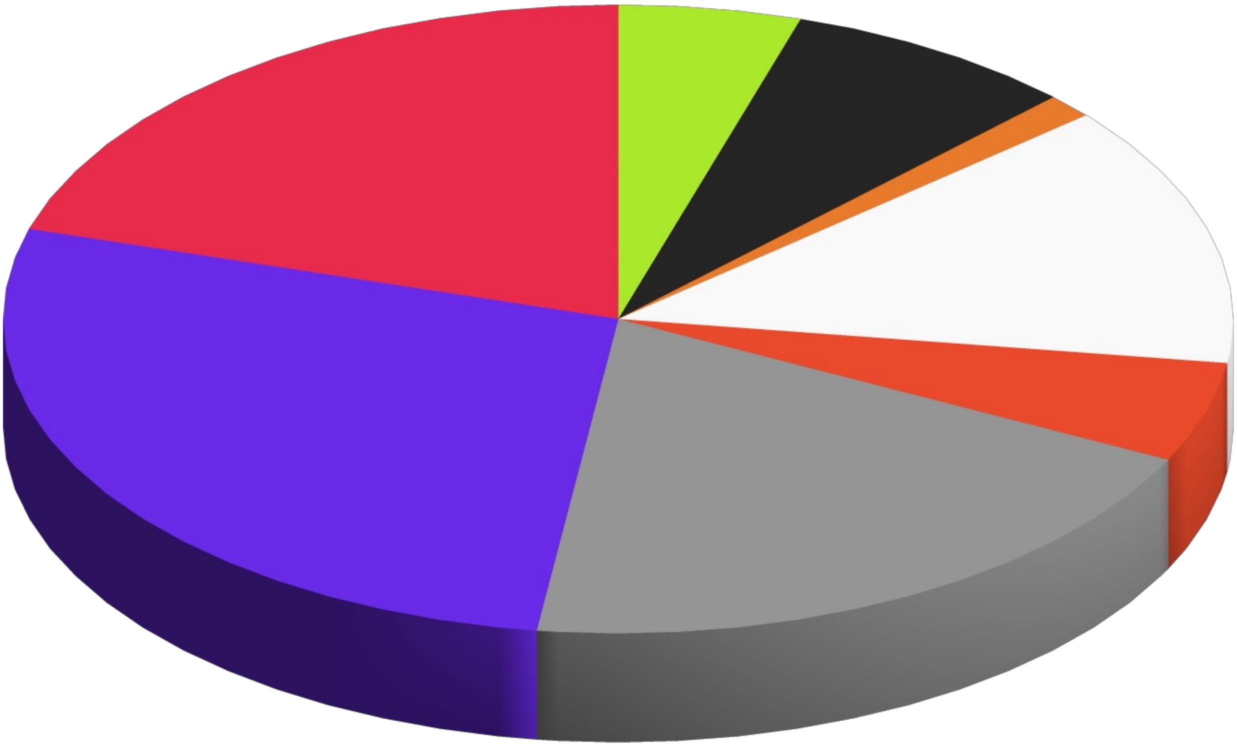
	answer	count	ratio
🟥	programmer	407	28,28%
🟪	developer	466	32,38%
🟩	researcher	323	22,45%
🟩	gamer	1	0,07%
⬜	student	231	16,05%
🟧	hobbist	2	0,14%
⬛	none of the above	9	0,63%

Do you have your own GPU?



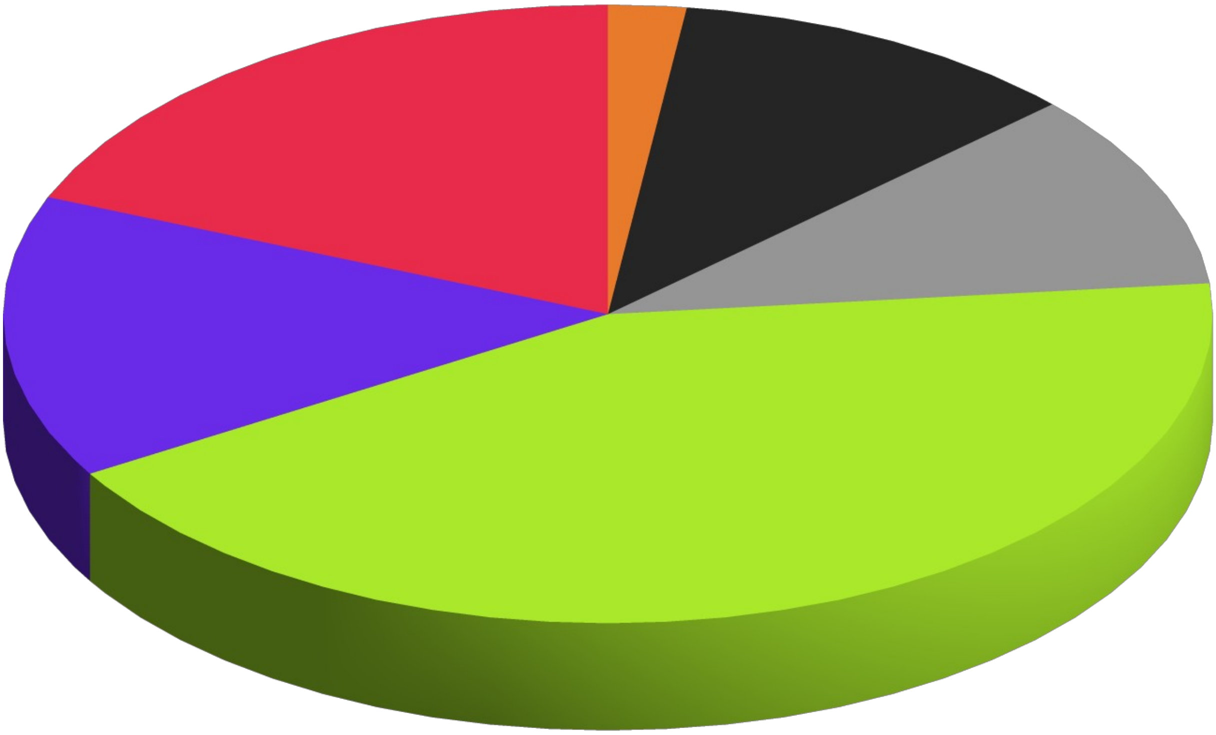
	answer	count	ratio
🟪	Yes	1237	85,96%
🟥	No	202	14,04%

What is the memory size of your GPU (in Gigabytes)?



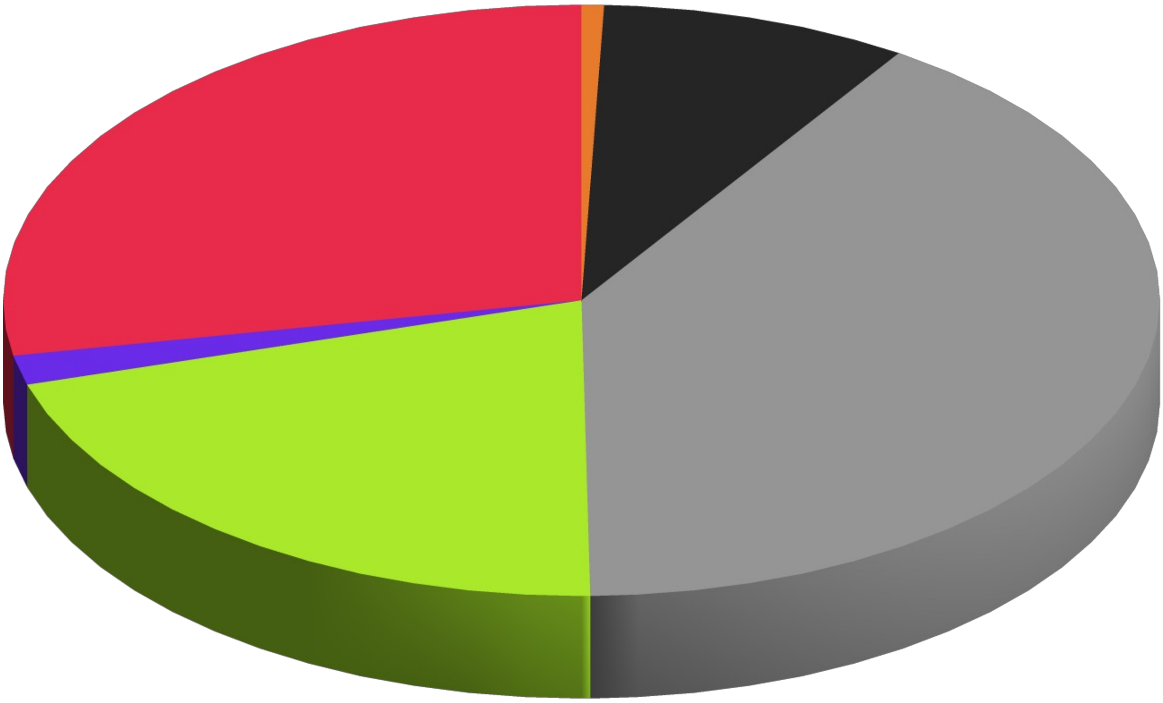
	answer	count	ratio
🟥	I don’t have own GPU	293	20,36%
🟪	2 or less	396	27,52%
▨	4 (more than 2 but no more than 4)	284	19,74%
🟧	6 (more than 4 but no more than 6)	74	5,14%
⬜	8 (more than 6 but no more than 8)	194	13,48%
🟧	10 (more than 8 but no more than 10)	18	1,25%
⬛	12 (more than 10 but no more than 12)	111	7,71%
🟩	more than 12	69	4,79%

How do you usually train your models at home?



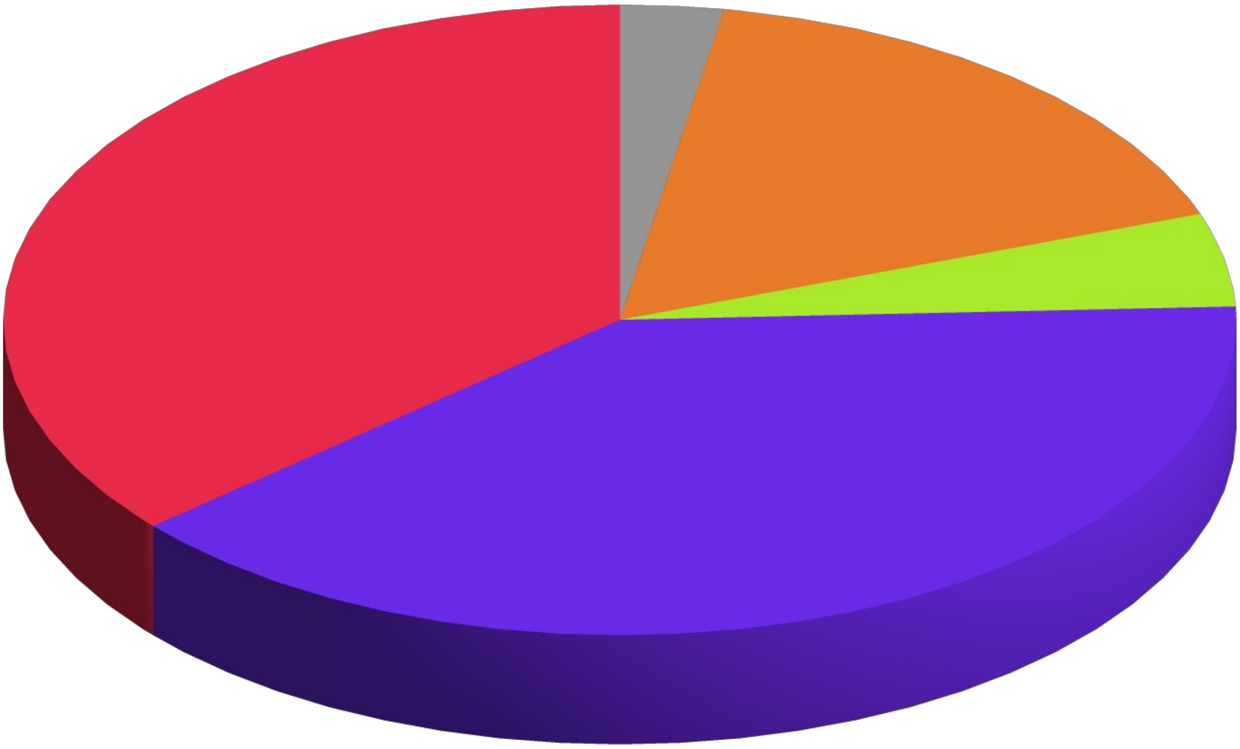
	answer	count	ratio
🟥	I don’t use my personal computer to train.	271	18,83%
🟪	CPU	213	14,80%
🟩	single GPU	618	42,95%
▩	GPU cluster	148	10,28%
⬛	cloud service	159	11,05%
🟧	other	30	2,08%

How do you usually train your models at workplace / school?



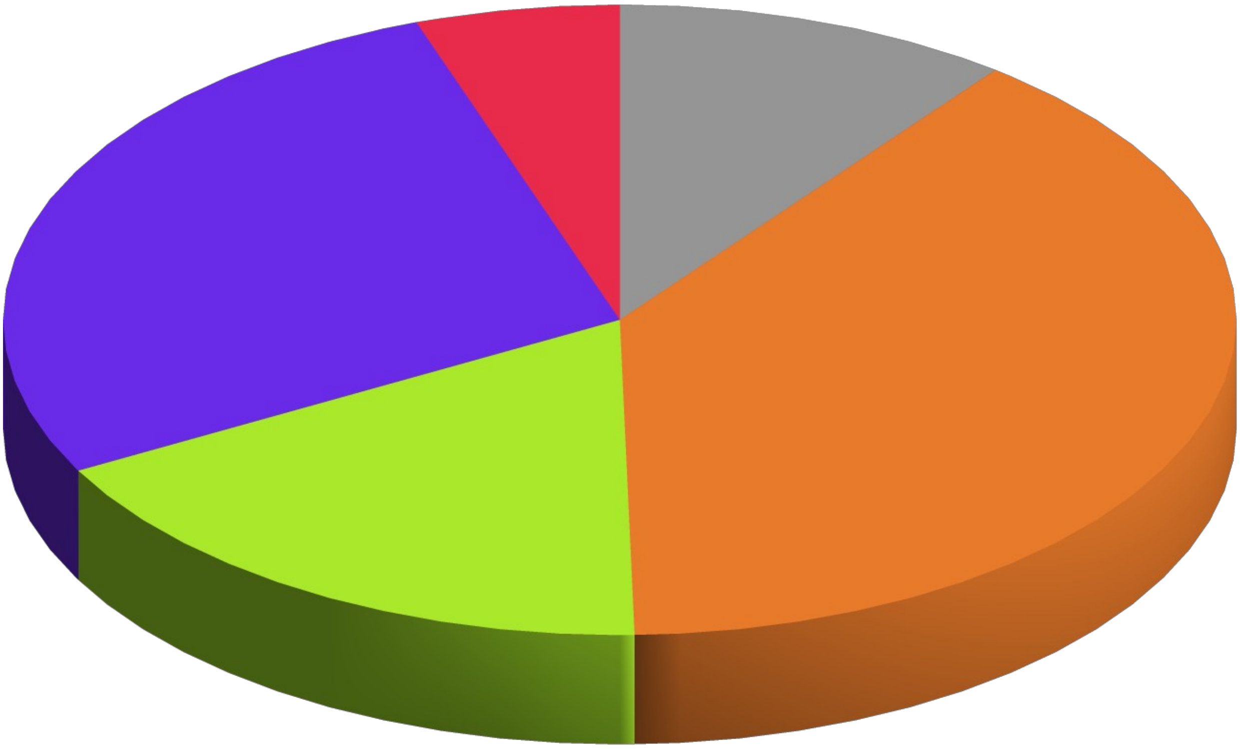
	answer	count	ratio
🟥	I don’t use my workstation / school computer to train.	403	28,01%
🟪	CPU	23	1,60%
🟩	single GPU	297	20,64%
▨	GPU cluster	583	40,51%
⬛	cloud service	124	8,62%
🟧	other	9	0,63%

Do you try to different batch size during training process?



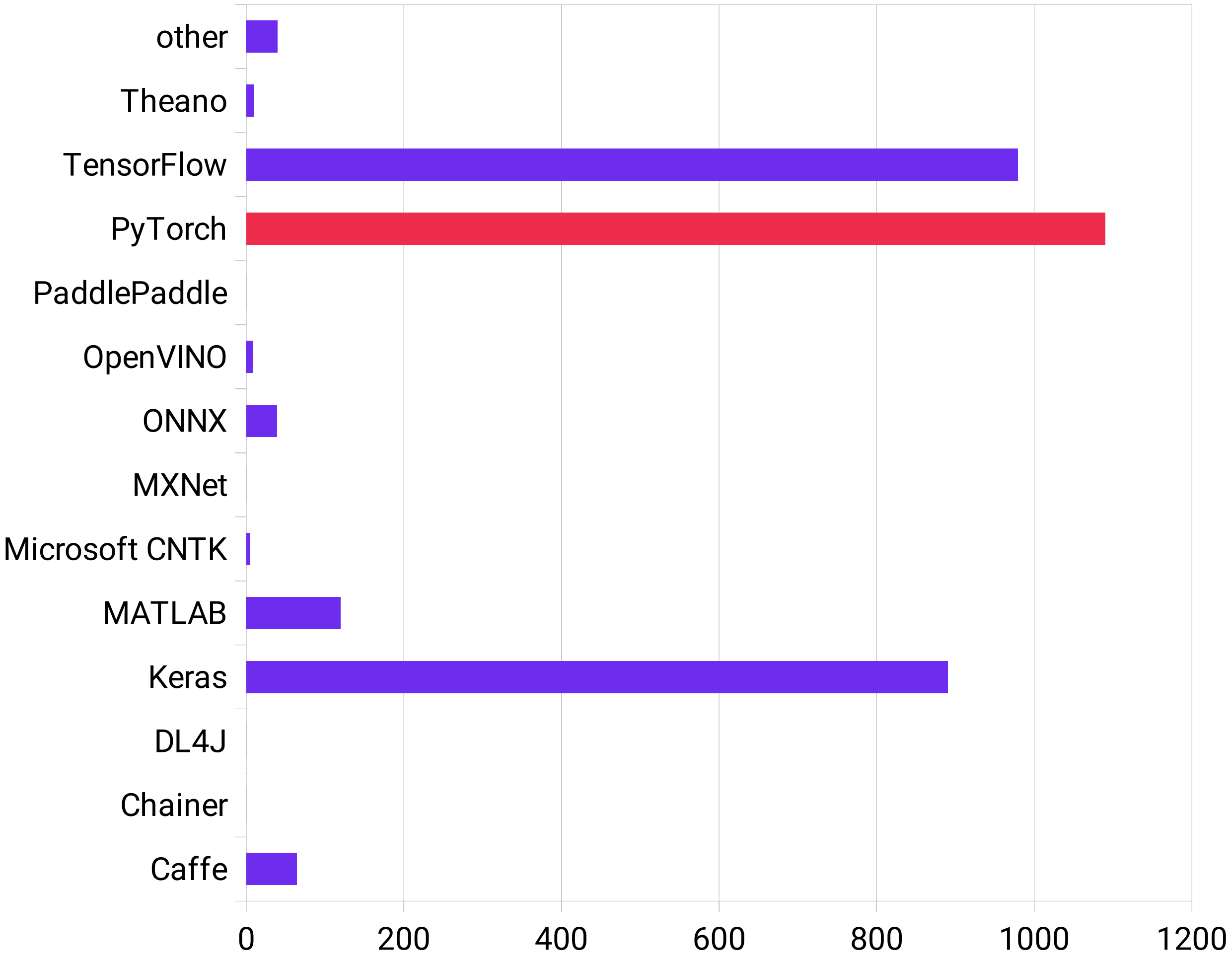
	answer	count	ratio
🟥	Always	523	36,34%
🟪	Usually	566	39,33%
🟩	Rarely	69	4,79%
🟧	Sometimes	242	16,82%
◼	Never	39	2,71%

How often do you get memory error (such as out of CUDA memory or any other memory overrun) message during training?



	answer	count	ratio
🟥	Always	77	5,35%
🟪	Usually	397	27,59%
🟩	Rarely	251	17,44%
🟧	Sometimes	564	39,19%
◼	Never	150	10,42%

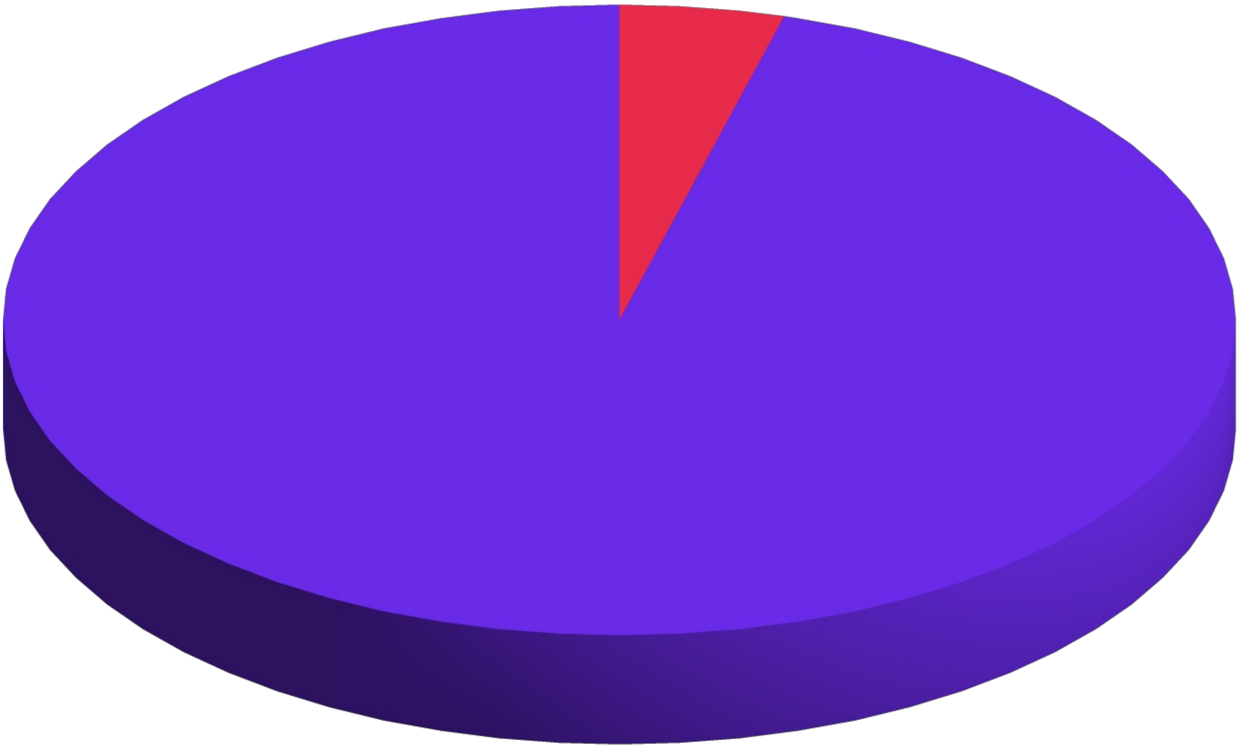
Which framework do you use?



	answer	count	ratio
🟪	Caffe	64	4,45%
🟪	Chainer	0	0,00%
🟪	DL4J	0	0,00%
🟪	Keras	890	61,85%
🟪	MATLAB	120	8,34%
🟪	Microsoft CNTK	5	0,35%
🟪	MXNet	0	0,00%

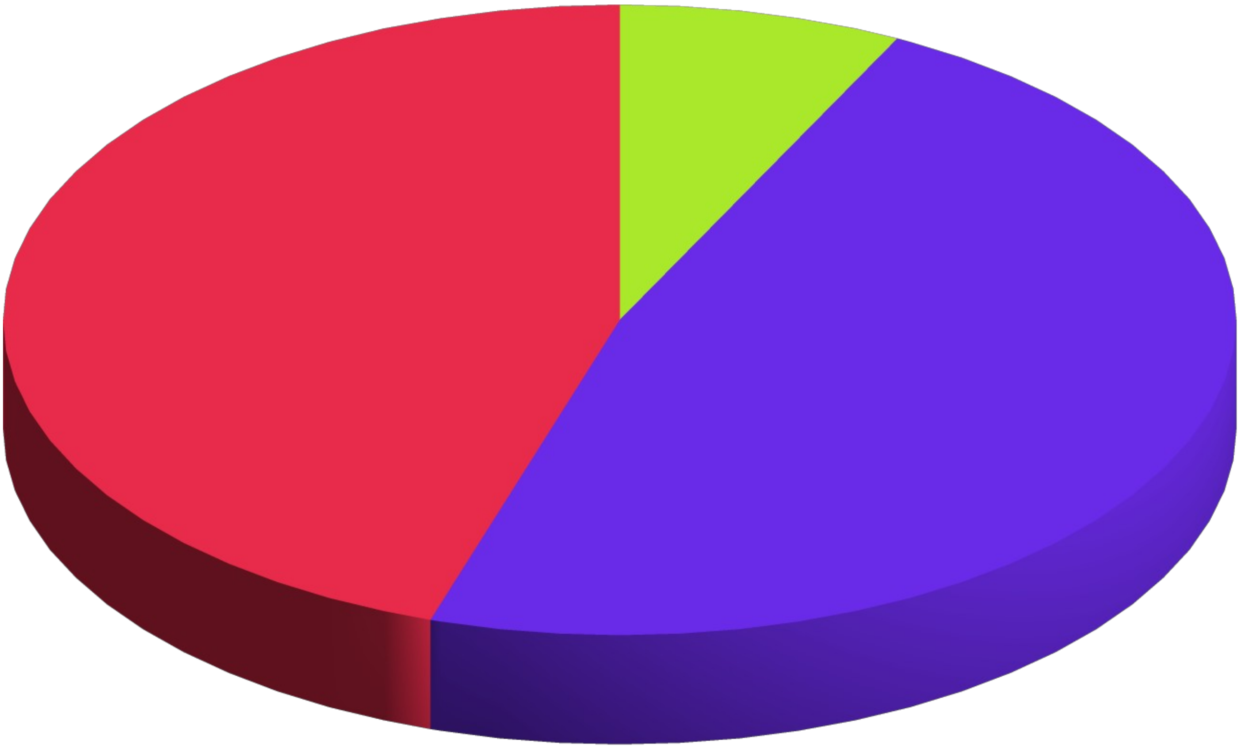
	answer	count	ratio
🟪	ONNX	39	2,71%
🟪	OpenVINO	9	0,63%
🟪	PaddlePaddle	0	0,00%
🟥	PyTorch	1090	75,75%
🟪	TensorFlow	979	68,03%
🟪	Theano	10	0,69%
🟪	other	40	2,78%

Do you think finetuning the hyperparameters matters for the performance of a model?



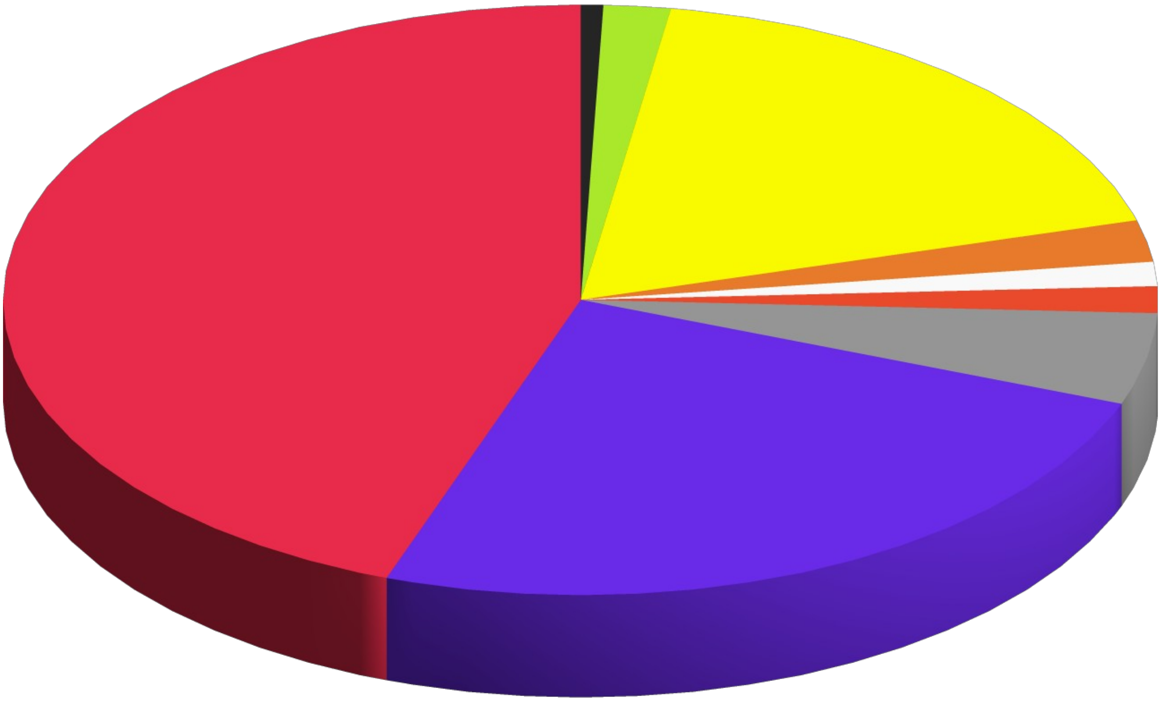
	answer	count	ratio
🟪	Yes	1377	95,69%
🟥	No	62	4,31%

Do you finetune hyperparameters?



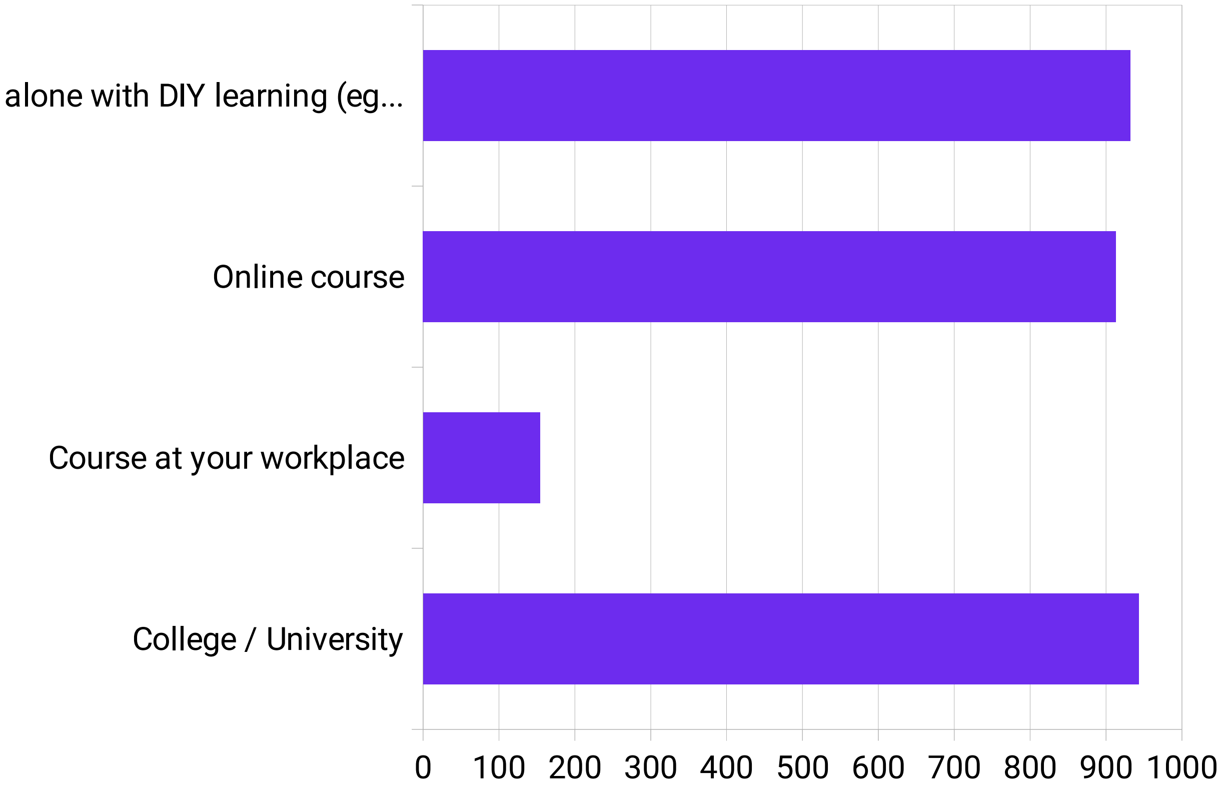
	answer	count	ratio
🟥	Yes with manual settings	648	45,03%
🟪	Yes with automated way	684	47,53%
🟩	No	107	7,44%

Where are you from?



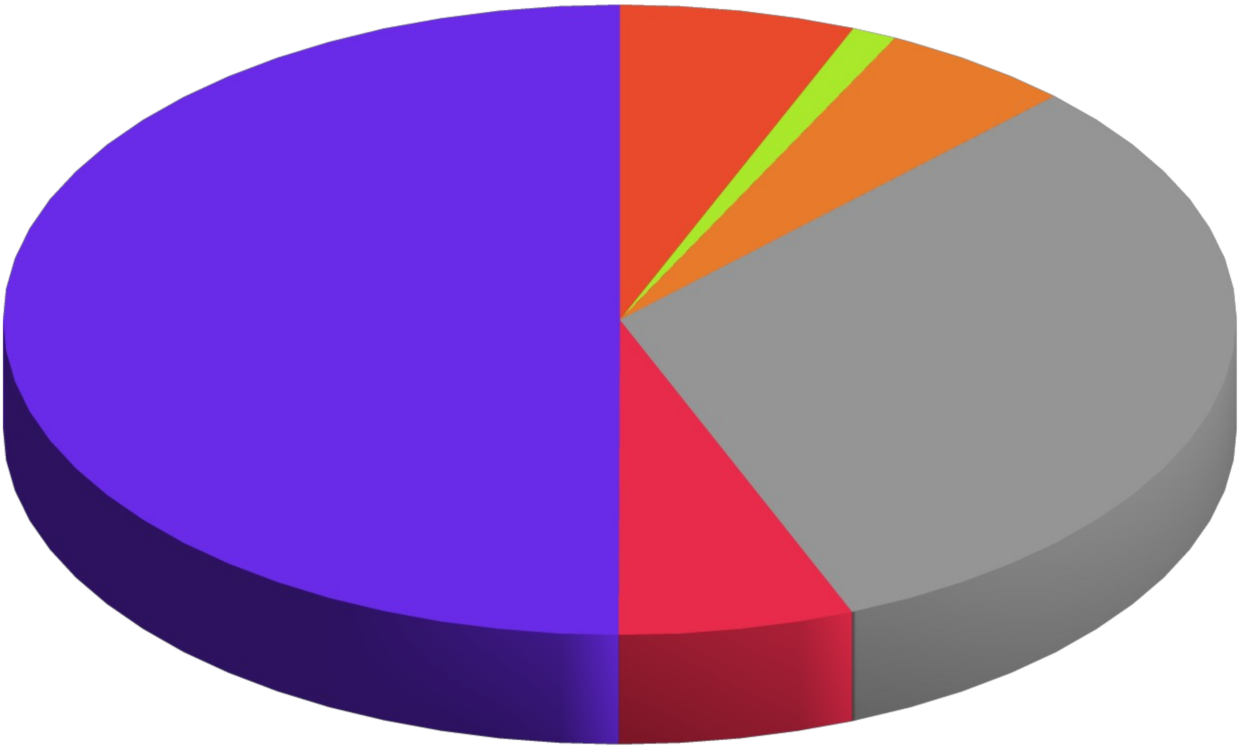
	answer	count	ratio
🟥	Europe	641	44,54%
🟪	North-America	356	24,74%
▨	Mid- and South-America	72	5,00%
🟥	Africa	21	1,46%
⬜	Australia and New Zealand	19	1,32%
🟧	Middle-East	33	2,29%
🟨	South and West Asia	261	18,14%
🟩	North and Middle Asia	27	1,88%
⬛	I prefer not to say it.	9	0,63%

Where did you learn about machine learning, deep learning?



	answer	count	ratio
🟪	College / University	943	65,53%
🟪	Course at your workplace	154	10,70%
🟪	Online course	913	63,45%
🟪	alone with DIY learning (eg: book, Youtube)	932	64,77%

Where did you meet first with machine learning, deep learning?



	answer	count	ratio
🟪	College / University	719	49,97%
🟥	Course at your workplace	89	6,18%
⬜	Online course	452	31,41%
🟧	Online video sharing site	73	5,07%
🟩	Book	17	1,18%
🟥	none of the above	89	6,18%

About us

Axel ORSZÁG-KRISZ Dr.

https://hyperrixel.com/en/axel

Richárd Ádám VÉCSEY Dr.

https://hyperrixel.com/en/richard

We are two deep learning developers and data scientists. We work on a lot of different fields of AI developing, but our main objective is a computer vision in healthcare. Nowadays, most companies like to hire developers who are outsiders, who have other skills and education. We are lawyers who have jumped from law to programming and deep learning. We had a lot of opportunities from Facebook and Udacity. These companies maintain scholarship programs to get knowledge from people around the world. We were two of them. Computer Vision Nanodegree or Deep Learning Nanodegree were supported by Facebook. Besides the knowledge, we got a lot of new viewpoints and the possibility to help others. We are thankful and during the COVID-19 epidemic, we are trying to get back something to the community and to the people around the Globe. Under these sign, we created the following projects.