GPU resources for machine translation

## Intro

In this report, we provide a brief analysis into the solutions regarding deep learning infrastructures, and more specifically, what kind of computing resources are required for training a machine translation model.

Again we remind that there are two main tasks in creating a machine translation pipeline: training and inference. Training a machine translation model with deep learning architectures may require powerful computing resources, like a big Graphics Processing Unit (GPUs). GPUs accelerate the training process of neural networks, due to their ability to compute processes in parallel. On the other hand, for the inference task, which is essentially to provide a translation when a sentence is given, no powerful resources are necessarily required.

It is clear that there are three solutions: buying our own machine, using cloud resources or both (hybrid).

## In-house (ANS premises)

Pros:

* Dedicated machine
* Free to use whenever we want and easier to experiment
* Will give us the ability to train even more powerful state-of-the-art machine translation models

Cons:

* Initial cost to buy (may reach 5-6k euros)
* May still not be powerful enough to try with state-of-the-art models
* May not be easy to connect remotely outside ANS premises

## Cloud resources

Remote resources like Amazon services (AWS), Azure, Google cloud are available:

Pros:

* Unlimited resources
* Remote connection

Cons

* Cost may raise significantly
* More difficult to operate remotely

Links:

* <https://cloud.google.com/gpu> (https://cloud.google.com/products/calculator)
* <https://docs.aws.amazon.com/dlami/latest/devguide/gpu.html>

## Conclusion

In general, there are many links on the web stating that in the end it will be cheaper to buy a machine on our own. In case of not having the option for ANS premises to be have a public open tool, we may select a hybrid approach: get a decent workstation with GPU at ANS premises for training the models and then use cloud services for production purposes.

## Posts

* <https://medium.com/the-mission/why-building-your-own-deep-learning-computer-is-10x-cheaper-than-aws-b1c91b55ce8c>
* <https://medium.com/the-mission/how-to-build-the-perfect-deep-learning-computer-and-save-thousands-of-dollars-9ec3b2eb4ce2>
* https://towardsdatascience.com/data-science-in-the-cloud-239b795a5792

## Screenshot from Google’s cloud calculator

