Applikation för medeldyr utrustning

Adam Dahlgren

August 21, 2019

1 Description of Infrastructure

The infrastructure that we seek to build is a state of the art machine learning research system. Machine learning research is heavily dependant on high-performance specialized GPU cards. These graphical processing units are central the heavy matrix computations that machine learning relies on. The state of the art in machine learning hardware, Nvidias line of Tesla GPUs, also have Tensor cores that are designed specifically for the most common machine learning software library, Tensorflow. The infrastructure is built around the Nvidia Tesla V100, the most powerful and computationally efficient machine learning GPU available. This hardware allows for training machine learning models that would otherwise take weeks, in a few days. Not only is it possible to train larger and more advanced models due to the high capacity in terms of memory, but each cycle in the iterative development of machine learning models and theories is shortened significantly. It is important that the GPU cards are backed up by the appropriate hardware, to avoid putting a mule in front of a car. Therefore, the server architecture must ensure a access to large amounts of working memory, a high bandwidth and low latencies when accessing the large datasets usually involved in training machine learning models.

Amount applied for: 268 000 SEK TODO All, review suggested infrastructure.

Applicant: Adam Dahlgren Lindström, Department of Computing Science. Email: dali@cs.umu.se. Phone: 070-340 33 70 TODO Adam, update with Franks info

Co-applicants: TODO Lili+Son, add your own data

2 Background

The last decade has seen significant advancements in deep learning, resulting in a plethora of important applications ranging from healthcare to self-driving cars. Many of these applications will change our lives for the better, if they

have not done so already. Deep learning has given us a range of important results, e.g. in surpassing human-level accuracy in object detection[1]. This has opened up for addressing challenges that has previously been out of reach, such as answering questions about a given scene or reason about its contents and their relationships. Given the prospect of solving these difficult problems, one of the previous challenges has been handling the size and complexity of the models required.

TODO Lili+Son, review

3 Scientific Value of the Infrastructure

- Own research, becoming world leading in multimodal machine learning
- Providing Master students with state of the art hardware to compete nationally
- Have the infrastructure as an incentive for research collaborations
- Many people need this type of infrastructure, and instead of many smaller instances it is much more cost-efficient to do have a large one to share.
- Reproducability
- TODO Lili+Son, add description (a couple of sentences) of what own research would be benefited with this infrastructure
- TODO Son, list resources other than SNIC+HPC2N that are available locally and nationally.
- TODO Adam, send email to AI@UMU and get some more descriptions of research.

Complementing SNIC:

- Iterative process of developing final model can require hundreds of iterations. Using SNIC for this would 1) introduce a much too long overhead as HPC2N has anywhere from an hour to more than a day in queue times, and 2) this iterative work hogs the SNIC resources just to show that the training process crashes or that the model doesn't learn anything in its current shape.
- Difficult with larger models to try this on anything smaller than this type of infrastructure.
- Data integrity. Projects including sensitive data (such as medical data) is much easier to handle on-site, which makes such projects both easier to realize and quicker to get up and running. SNIC has no GPU resources that provides this feature.

[1] source

- For short projects, e.g. Bachelor's and Master thesis work, it is essential that the hardware is ready to use early in the projects. Applying for SNIC resources, waiting for approval, and spending time waiting in work queues can greatly affect the outcome of a project. Having access to on-site infrastructure would allow for higher quality thesis work with better results that have higher chances of being published.
- This infrastructure that complements SNIC is something most machine learning researchers need in their research process, meaning that if UmU has such a system it becomes a stronger contributor in collaborations which could make way for larger projects and more interesting collaborations.

Utrustningens vetenskapliga betydelse och institutionens/forskargruppens behov, inklusive en redogörelse för i vilken grad likartad utrustning redan finns vid fakulteten, lärostätet, och inom landet. I de senare fallen bör frågan om i vilken grad sådan eventuell utrustning söks gemensamt med medsökande skall det, under ovanstående rubruk, framgå på vilket sätt, och i vilken grad, utrustningen kommer att stödja samtliga sökandes verksamhetsområden. I det fall någon av de medsökande ej är anställda vid Umeå Universitet skall, under ovanstående rubrik, även utrustningens förväntade användning av forskare från de olika lärosätena/andra instanserna klart framgå.

4 Premises

The infrastructure will be placed in the joint computer room for the Department of Computing Science and HPC2N.

5 Responsible for the Infrastructure

Adam Dahlgren Lindström is the main responsible for the Infrastructure with the support of Tomas Forsman as the principal system administrator. The procurement process is performed by the Umeå University unit for ICT Services and System Development (ITS) together with Tomas Forsman.

6 Classification

The infrastructure is planned to operate under the definition of "User Groups", but due to the possibility for expansion it could later also be used under the definition of "Technical Platform".

7 Tillstyrkan

Ansökan skall signeras av samtliga sökande (dvs. huvud- och medsökande). Därtill skall den tillstyrkan av prefekt, innebärande att institutionen kommer

att finansiera förekommande gemensamma kostnader om ansökan bifalls.

A Lista över medsökande

TODO Adam, list from AI@UMU.

Lista över medsökande: Namn, titel, institutions- och fakultetstillhörighet samt kontaktuppgifter för samtliga medsökande.

B CV

Kort CV inkluderande en förteckning över beviljade medel samt publikationslista för de senaste 7 åren för samtliga sökande.

C Budget

Budget där sökandes och ev. medsökandes möjligheter till medfinansiering klart framgår inkluderande en plan för den långsiktiga finansieringen av utrustningen, d.v.s. ev. installationskostnader, drift samt avyttrande. Max 1 sida.

D Konsekvensanalys

TODO Lili or Son, draft what the consequences of getting this funding or not would implicate.

Konsekvensanalys omfattande både konsekvensen av att göra satsningen och att avstå från densamma. Max 1 sida.

E Offert

TODO Adam, once infrastructure is set, add this with help from Stric.

Om möjligt, offert på tilltänkt utrustning, alternativt annan allmän information om utrustningen.

F Prefekts tillstyrkande

I de fall den huvudsökande innehar en tidsbegränsad anställning, dvs en anställning som bidträdande lektor eller forskarassistent, ska ansökan tillstyrkas av prefekt i ett separat yttrande. Av detta skall utrustningens relevans i ett längre perspektiv framgå.