MSc assignments at Naspers (2018/2019)

Naspers is a global internet and entertainment group and one of the largest technology investors in the world. The Naspers AI team helps designing AI and machine learning into everything we do, to serve the needs of our customer better and to continue building businesses that empower people and enrich communities. We use AI to improve existing processes by making them more efficient. We take an ‘AI-by-design’ approach to deliver new products and services for the online platforms we work with and invest in.

Contact information

Telephone 0235692399
Website https://www.naspers.com/

Representatives

Dmitri Jarnikov (d.s.jarnikov@tue.nl)
Description

Keywords: Natural language processing, Semi-Supervised learning, Clustering, Named Entity Recognition.

OLX Group, part of Naspers, is group of trading platforms present in more than 40 countries in over 50 languages. The OLX platform makes it easy to connect people to buy, sell or exchange used goods and services. Every month, more than 350 millions of people use it to find and sell cars, real estate, home appliances, musical instruments, consumer electronic goods and more. A seller posts a ‘listing’ or an ‘ad’, typically composed of a title, textual description and a set of images. The listing database grows as fast as 60+ millions of new items per month.

Each product family shares a set of attributes that describes the product. For example, a wrist watch can be described with attributes such as brand, gender, style, strap material, colour and many others features. Typically, such attributes are defined by product manufacturers and are often based on common knowledge of what is relevant for the product. In the case of OLX platform, the sellers provide their own descriptions of a product which might include such attributes in an unorganized fashion.

Project Objective

The main objective of this project is to design a system that extracts the relevant attributes from the user’s descriptions obtained from the classifieds platform. The system should also be capable of identifying for such attributes what is the range of their ‘typical’ values. For example in the description “Golden wrist watch” the solution should identify the attribute “Color” with value “Gold”.

Dataset

Unlabeled data [title, description]. ~3000 categories, ~1M samples.

References


What will the work environment look like for the student?

Naspers is a global internet and entertainment group and one of the largest technology investors in the world. The student shall expect:

- Technology driven environment;
- Caring meritocracy;
- Multicultural organisation;
- High-performance culture.

As well we offer:

- A competitive monthly internship compensation;
• Offices in Amsterdam and Eindhoven;
• Weekly (optional) seminars where a team member or a colleague from our portfolio company presents a state-of-the-art machine learning topic;
• An opportunity to bring a machine learning solution to production by collaborating with one of our portfolio companies;
• Monthly company all-hands meetings with cool activities and a dinner.

What are your expectations of the student?

We are looking for highly motivated students that are passionate about Machine Learning technologies and have a drive to learn. The student should be experienced in Python and comfortable with frameworks such as Tensorflow and Torch.
Class Activations for Weakly Supervised Object Detection
At Naspers

Description

**Keywords:** Computer vision, Convolutional Neural Networks, Image Classification, Weakly-Supervised Learning (object localization).

OLX Group, part of Naspers, is a network of trading platforms present in more than 40 countries in over 50 languages. The OLX platform makes it easy to connect people to buy, sell or exchange used goods and services. Every month, more than 350 million users visit the platform to find and sell cars, real estate, home appliances, musical instruments, consumer electronic goods and more. A seller posts a ‘listing’ or an ‘ad’, typically composed of a title, textual description and a set of images. Every month 60+ million new listings are posted.

Naspers currently has a dataset with about ~10 million images labeled by users who post objects for selling them on classifieds platforms (~3000 classes). The labels provided by the users classify the complete image into a certain class, we would like to design from these labels a detector capable of identifying the location of an object within the image just from these ‘weak’ labels.

**Project Objective**

The goal of this project is to create a (semi-)automated solution that can ‘translate’ our image classification dataset into an object detection dataset. One possible approach consists of training several image classifiers with different architectures on the image classification dataset. Then use all of them for classifying a single image, intersect the areas that contribute positively to the classification and finally create a bounding box around the object from the intersection.

**Dataset**

labelled image data [image, category]. ~3000 categories, ~10M images.

**What will the work environment look like for the student?**

Naspers is a global internet and entertainment group and one of the largest technology investors in the world. The student shall expect:

- Technology driven environment;
- Caring meritocracy;
- Multicultural organization;
- High-performance culture.

As well we offer:

- A competitive monthly internship compensation;
- Offices in Amsterdam and Eindhoven;
- Weekly (optional) seminars where a team member or a colleague from our portfolio company presents a state-of-the-art machine learning topic;
• An opportunity to bring a machine learning solution to production by collaborating with one of our portfolio companies;
• Monthly company all-hands meetings with cool activities and a dinner.

What are your expectations of the student?

We are looking for highly motivated students that are passionate about Machine Learning technologies and have a drive to learn. The student should be experienced in Python and comfortable with frameworks such as Tensorflow and Torch.
Deep Listing Corrector

At Naspers

**Keywords:** Natural Language Processing, Information Extraction, Seq2Seq.

OLX Group, part of Naspers, is a network of trading platforms present in more than 40 countries in over 50 languages. The OLX platform makes it easy to connect people to buy, sell or exchange used goods and services. Every month, more than 350 million users visit the platform to find and sell cars, real estate, home appliances, musical instruments, consumer electronic goods and more. A seller posts a ‘listing’ or an ‘ad’, typically composed of a title, textual description and a set of images. Every month 60+ million new listings are posted.

**Project Motivation**

The Personalization & Relevance team at OLX is increasing transactions between buyers and sellers by showing the most relevant content to the users. The team is developing personalization technologies and optimization strategies that fuel the discovery of new items and have a direct impact on OLX’s bottom line. We are researching deep learning techniques to make it easier for users to describe the item they are selling.

**Project Objective**

The goal of the project is to propose a solution that can automatically improve a textual description of a listed item provided by the user on a classified platform. The solution should be a *domain-specific* sequence-to-sequence model capable of processing a description of a product and generate a corrected version.

**Dataset**

Listings data [text, image(s), category]. ~3000 categories, ~1M listings.

**What will the work environment look like for the student?**

Naspers is a global internet and entertainment group and one of the largest technology investors in the world. The student shall expect:

- Technology driven environment;
- Caring meritocracy;
- Multicultural organization;
- High-performance culture.

As well we offer:

- A competitive monthly internship compensation;
- Offices in Amsterdam and Eindhoven;
- Weekly (optional) seminars where a team member or a colleague from our portfolio company presents a state-of-the-art machine learning topic;
• An opportunity to bring a machine learning solution to production by collaborating with one of our portfolio companies;
• Monthly company all-hands meetings with cool activities and a dinner.

What are your expectations of the student?

We are looking for highly motivated students that are passionate about Machine Learning technologies and have a drive to learn. The student should be experienced in Python and comfortable with frameworks such as Tensorflow and Torch.
Finding Optimal Feature Sets in Supervised Classification by Building an Opensource Python Library

At Naspers

Description

Keywords: Feature selection, Machine Learning.

Taking advantage of the cloud computation power is a common practice nowadays in many supervised classification tasks such as fraud detection. However, using large number of features in models brings the cost of long training times, model overfitting problems and maintenance issues for data pipelines. Thus, models should be built on minimum number of features, and yet, keep the maximum prediction power.

Feature selection is the process of finding optimal subset of features (predictive variables) that maximizes the prediction performance of the machine learning model. Embedded feature selection, one of the most recent and popular techniques, incorporates model building attributes into feature selection process and determines feature importance with relatively low cost.

One problem with embedded feature selection techniques is that feature attribution methods are inconsistent (Lundberg 2017), especially in the existence of highly correlated features. Another problem is that embedded techniques are highly dependent on the structure of the classification models. For example, the list of top-N features from a model using bagged decision trees is usually different than the top-N list of another model using boosted decision trees.

Project Objective

The goal of the project is to build a unified feature selection framework determining feature importance in a consistent manner and finding the optimal feature subset for the given classification task automatically. The framework will be implemented as an opensource Python library and be compatible with popular ML packages such as Scikit-learn, XGBoost, LightGBM and CatBoost.

Data

We have 500M of transaction data and 250K user data for fraud detection. We also have 200K user data for churn prediction. The library will also be validated using public data from Kaggle competitions.

References


What will the work environment look like for the student?

Naspers is a global internet and entertainment group and one of the largest technology investors in the world. The student shall expect:

- Technology driven environment;
- Caring meritocracy;
- Multicultural organization;
- High-performance culture.
As well we offer:

- A competitive monthly internship compensation;
- Offices in Amsterdam and Eindhoven;
- Weekly (optional) seminars where a team member or a colleague from our portfolio company presents a state-of-the-art machine learning topic;
- An opportunity to bring a machine learning solution to production by collaborating with one of our portfolio companies;
- Monthly company all-hands meetings with cool activities and a dinner.

**What are your expectations of the student?**

We are looking for highly motivated students that are passionate about Machine Learning technologies and have a drive to learn. The student should be experienced in Python and comfortable with frameworks such as Tensorflow and Torch.
Generalized Learning across Semantic Embedding Spaces

At Naspers

**Description**

**Keywords:** Search2Vec, Natural Language Processing, Image Text Embeddings, Query Matching.

OLX Group, part of Naspers, is a network of trading platforms present in more than 40 countries in over 50 languages. The OLX platform makes it easy to connect people to buy, sell or exchange used goods and services. Every month, more than 350 million people use it to find and sell cars, real estate, home appliances, musical instruments, consumer electronic goods and more. A seller posts a 'listing' or an 'ad', typically composed of a title, textual description and a set of images. Every month 60+ million new listings are posted.

Currently, we use search2vec (Grbovic 2016) to provide an embedding of the user’s queries and provided listings in the platform. These embeddings englobe the characteristics learned from large datasets that include the user search sessions. These search sessions consists of search queries, clicked listings and the category of the listings.

One issue that arises is that the obtained embedding spaces are typically associated with a specific geographical area. Areas differ in language as well as in the type of goods being sold or searched.

**Project Objective**

The main goal of this project is to provide a method for connecting the embedding spaces from different geographical areas that preserves the semantic concepts/products that they represent.

**References**


**What will the work environment look like for the student?**

Naspers is a global internet and entertainment group and one of the largest technology investors in the world. The student shall expect:

- Technology driven environment;
- Caring meritocracy;
- Multicultural organization;
- High-performance culture.

As well we offer:

- A competitive monthly internship compensation;
- Offices in Amsterdam and Eindhoven;
- Weekly (optional) seminars where a team member or a colleague from our portfolio company presents a state-of-the-art machine learning topic;
• An opportunity to bring a machine learning solution to production by collaborating with one of our portfolio companies;
• Monthly company all-hands meetings with cool activities and a dinner.

What are your expectations of the student?

We are looking for highly motivated students that are passionate about Machine Learning technologies and have a drive to learn. The student should be experienced in Python and comfortable with frameworks such as Tensorflow and Torch.
Description

**Keywords:** Computer Vision, Neural Image Captioning, CNN-LSTM, Natural Language Processing.

OLX Group, part of Naspers, is a network of trading platforms present in more than 40 countries in over 50 languages. The OLX platform makes it easy to connect people to buy, sell or exchange used goods and services. Every month, more than 350 million users visit the platform to find and sell cars, real estate, home appliances, musical instruments, consumer electronic goods and more. A seller posts a ‘listing’ or an ‘ad’, typically composed of a title, textual description and a set of images. Every month 60+ million new listings are posted.

**Project Motivation**

The title of a listing in an online classified marketplace affects its liquidity strongly since titles is one of the factors used for indexing/ranking in most online marketplaces. Hence, writing a suitable title is crucial for improving the product positioning. However, writing a suitable title isn't easy. In this project we are looking for a solution that can generate suitable titles from the images published by users in their product advertisements. However, unlike general image captioning certain criteria for a proper and comprehensive title must be considered in the created model.

**Project Objective**

The main objective of this project is to create an image captioning model that considers specific criteria for suitable listing titles. The obtained should also be applicable for many further usages, such as keyword extractions, hashtag generators, etc.

**Dataset**

Listings data [text, image(s), category]. ~3000 categories, ~1M listings.

**What will the work environment look like for the student?**

Naspers is a global internet and entertainment group and one of the largest technology investors in the world. The student shall expect:

- Technology driven environment;
- Caring meritocracy;
- Multicultural organization;
- High-performance culture.

As well we offer:

- A competitive monthly internship compensation;
- Offices in Amsterdam and Eindhoven;
• Weekly (optional) seminars where a team member or a colleague from our portfolio company presents a state-of-the-art machine learning topic;
• An opportunity to bring a machine learning solution to production by collaborating with one of our portfolio companies;
• Monthly company all-hands meetings with cool activities and a dinner.

What are your expectations of the student?

We are looking for highly motivated students that are passionate about Machine Learning technologies and have a drive to learn. The student should be experienced in Python and comfortable with frameworks such as Tensorflow and Torch.
Image quality improvements using deep learning
At Naspers

Description

**Keywords:** Computer Vision, Convolutional Neural Networks, Generative Adversarial Networks.

OLX Group, part of Naspers, is a group of trading platforms present in more than 40 countries in over 50 languages. The OLX platform makes it easy to connect people to buy, sell or exchange used goods and services. Every month, more than 350 millions of people use it to find and sell cars, real estate, home appliances, musical instruments, consumer electronic goods and more. A seller posts a ‘listing’ or an ‘ad’, typically composed of a title, textual description and a set of images. The listing database grows as fast as 60+ millions of new items per month.

One of the challenges is that the listings posted by users contain photographs which might have a poor quality. Our objective is to automatically improve the quality of the provided images.

**Project Objective**

The challenge of the project is to provide a solution that takes low resolution, blurred images that may have unwanted-scene areas and produce a high-resolution images with high aesthetic quality using deep learning techniques. There are various approaches that tackle the low-resolution-to-high-resolution conversion (Kanska 2017), photo retouching (Park 2018), and automatic image cropping (Guo 2018). The assignment is to evaluate existing techniques and propose a solution that addresses multiple improvements simultaneously using user preferences as metrics for aesthetic image quality.

**References**

https://deepsense.ai/using-deep-learning-for-single-image-super-resolution/


**What will the work environment look like for the student?**

Naspers is a global internet and entertainment group and one of the largest technology investors in the world. The student shall expect:

- Technology driven environment;
- Caring meritocracy;
- Multicultural organization;
- High-performance culture.

As well we offer:

- A competitive monthly internship compensation;
- Offices in Amsterdam and Eindhoven;
- Weekly (optional) seminars where a team member or a colleague from our portfolio company presents a state-of-the-art machine learning topic;
- An opportunity to bring a machine learning solution to production by collaborating with one of our portfolio companies;
- Monthly company all-hands meetings with cool activities and a dinner.

**What are your expectations of the student?**

We are looking for highly motivated students that are passionate about Machine Learning technology and have a drive to learn. The student should be experienced in Python and comfortable with frameworks such as Tensorflow and Torch.
Description

**Keywords**: Computer Vision, Natural language processing, Deep Learning, Cross-Modal Learning, Image-text association mining.

OLX Group, part of Naspers, is a group of trading platforms present in more than 40 countries in over 50 languages. The OLX platform makes it easy to connect people to buy, sell or exchange used goods and services. Every month more than 350 million people use it to find and sell cars, real estate, home appliances, musical instruments, consumer electronic goods and more. Sellers are mostly non-professional individuals whereas a portion of the user base consists of professional sellers such as car dealers and real-estate agencies. A seller posts a ‘listing’ or an ‘ad’, typically composed of a title, textual description and a set of images. Every month 60+ million new listings are posted.

We would like to build a solution to categorize and describe listings using machine learning by utilizing both textual and visual information, so that they become easier to find. One of the challenges is that listings contain a variable number of images and the majority of the users are not professional sellers, therefore the text descriptions don’t adhere to any standards and can be truly unique. Combined with the sheer amount of data, the underlying textual and visual data has a vast variance across thousands of item categories.

**Project Objective**

The aim of this project is to research algorithms and methods to build robust classification models utilizing both text and images, inherently combining computer vision (CV) with natural language processing (NLP).

Consequently, we would like to deploy the trained models on distributed computing platforms to make an impact at large scale.

**Dataset**

Labelled listings data [text, image(s), category]. ~3000 categories, ~1M listings.

**What will the work environment look like for the student?**

Naspers is a global internet and entertainment group and one of the largest technology investors in the world. The student shall expect:

- Technology driven environment;
- Caring meritocracy;
- Multicultural organization;
- High-performance culture.
As well we offer:

- A competitive monthly internship compensation;
- Offices in Amsterdam and Eindhoven;
- Weekly (optional) seminars where a team member or a colleague from our portfolio company presents a state-of-the-art machine learning topic;
- An opportunity to bring a machine learning solution to production by collaborating with one of our portfolio companies;
- Monthly company all-hands meetings with cool activities and a dinner.

What are your expectations of the student?

We are looking for highly motivated students that are passionate about Machine Learning technologies and have a drive to learn. The student should be experienced in Python and comfortable with frameworks such as Tensorflow and Torch.
Multimodal Listing (Ad) Summarization

At Naspers

Keywords: Multimodal Data, Text Summarization, Computer Vision, Natural Language Processing.

OLX Group, part of Naspers, is a network of trading platforms present in more than 40 countries in over 50 languages. The OLX platform makes it easy to connect people to buy, sell or exchange used goods and services. Every month, more than 350 million users visit the platform to find and sell cars, real estate, home appliances, musical instruments, consumer electronic goods and more. A seller posts a ‘listing’ or an ‘ad’, typically composed of a title, textual description and a set of images. Every month 60+ million new listings are posted.

We are researching deep learning techniques to make it easier for sellers to describe the item they are selling. We would like to develop a tool capable of taking a listing composed of images and text given by the user of the classifieds platform and produce a summary that describes this listing.

Project Objective

The main goal of this project is to provide a solution capable of taking a listing composed of images and text and provide a summary of the most important features, characteristics, attributes that describe the object. It is important that the solution is capable of understanding which pictorial elements convey essential information and combine it with the textual descriptions given by the user to enhance current text summarization methods.

What will the work environment look like for the student?

Naspers is a global internet and entertainment group and one of the largest technology investors in the world. The student shall expect:

- Technology driven environment;
- Caring meritocracy;
- Multicultural organization;
- High-performance culture.

As well we offer:

- A competitive monthly internship compensation;
- Offices in Amsterdam and Eindhoven;
- Weekly (optional) seminars where a team member or a colleague from our portfolio company presents a state-of-the-art machine learning topic;
- An opportunity to bring a machine learning solution to production by collaborating with one of our portfolio companies;
- Monthly company all-hands meetings with cool activities and a dinner.

What are your expectations of the student?
We are looking for highly motivated students that are passionate about Machine Learning technologies and have a drive to learn. The student should be experienced in Python and comfortable with frameworks such as Tensorflow and Torch.
Real-time Personalized Recommendation System using Images, Text and User Behavior

At Naspers

Description

**Keywords**: Natural Language Processing, Computer Vision, Convolutional Neural Networks, Recurrent Neural Networks, Recommendation System, User Preferences.

OLX Group, part of Naspers, is a network of trading platforms present in more than 40 countries in over 50 languages. The OLX platform makes it easy to connect people to buy, sell or exchange used goods and services. Every month, more than 350 million users visit the platform to find and sell cars, real estate, home appliances, musical instruments, consumer electronic goods and more. A seller posts a ‘listing’ or an ‘ad’, typically composed of a title, textual description and a set of images. Every month 60+ million new listings are posted.

Majority of the users on OLX platforms have a clear intent while browsing items, e.g. they want to buy second-hand winter tires for their car. Thus, the search sessions mostly exhibit highly correlated actions, such as submitting similar search queries or clicking on similar listings. We would like to design and build a real-time recommender system that considers the specific browsing history of individual users and recommends similar listings to the user in real-time.

**Project Objective**

This project aims to design and build a personalized recommender system which harnesses the user behavior in combination with the information extracted from listing content, i.e. images and text. This project requires to integrate visual and textual representations (i.e. embeddings) with sequence modeling algorithms, such as LSTMs or GRUs.

**Dataset**

3.2M unlabeled listings [images, text]. 100M+ user search session data [queries, clicks, messaging signals]

**What will the work environment look like for the student?**

Naspers is a global internet and entertainment group and one of the largest technology investors in the world. The student shall expect:

- Technology driven environment;
- Caring meritocracy;
- Multicultural organization;
- High-performance culture.

As well we offer:

- A competitive monthly internship compensation;
- Offices in Amsterdam and Eindhoven;
• Weekly (optional) seminars where a team member or a colleague from our portfolio company presents a state-of-the-art machine learning topic;
• An opportunity to bring a machine learning solution to production by collaborating with one of our portfolio companies;
• Monthly company all-hands meetings with cool activities and a dinner.

What are your expectations of the student?

We are looking for highly motivated students that are passionate about Machine Learning technologies and have a drive to learn. The student should be experienced in Python and comfortable with frameworks such as Tensorflow and Torch.
Keywords: Computer vision, Natural language processing, Deep Learning, Convolutional Neural Networks, Image classification, Object detection, Transfer learning, Hashtag Engineering.

OLX Group, part of Naspers, is a network of trading platforms present in more than 40 countries in over 50 languages. The OLX platform makes it easy to connect people to buy, sell or exchange used goods and services. Every month, more than 350 million of users visit the platform to find and sell cars, real estate, home appliances, musical instruments, consumer electronic goods and more. A seller posts a ‘listing’ or an ‘ad’, typically composed of a title, textual description and a set of images. Every month 60+ million new listings are posted.

Millions of active users generate massive amount of visual and textual data on OLX platforms. Think of datasets orders of magnitude larger than ImageNet. The generated data doesn’t naturally come with desired labels which makes it difficult to utilize for supervised learning. However, there is significant amount of user behavior data such as textual search queries, clicks on listings found relevant by the user and conversations between users via messages to perform the transaction. All these signals can be considered as weak labels for the underlying data. For instance, users may submit a search query ‘PlayStation 4’ and some listings may be clicked after this search query. This already indicates that the images and text belonging to the clicked listings most likely belong to ‘PlayStation 4’ category. We would like to utilize massive amounts of user behavior data to curate large-scale labelled datasets and employ them to train and deploy large neural networks.

Project Objective

The aim of this project is to research and design algorithms and methods to utilize large amounts of user behavior data to take our Computer Vision and Natural Language Processing models to the next level. Recently, Facebook utilized 3.5 Billion Instagram images and user hashtags to pre-train CNNs to significantly improve the classification performance on ImageNet over the existing the state-of-the-art (Mahajan D. et al. (2018)).

Dataset

Labelled image data [image, category] ~3000 categories, ~10 Images. Unlabeled listing data [images, text]: 3M+ listings. Anonymized user search and click data [search queries, clicks, messaging signals]: 100M+ samples

References

What will the work environment look like for the student?

Naspers is a global internet and entertainment group and one of the largest technology investors in the world. The student shall expect:

- Technology driven environment;
- Caring meritocracy;
- Multicultural organization;
- High-performance culture.

As well we offer:

- A competitive monthly internship compensation;
- Offices in Amsterdam and Eindhoven;
- Weekly (optional) seminars where a team member or a colleague from our portfolio company presents a state-of-the-art machine learning topic;
- An opportunity to bring a machine learning solution to production by collaborating with one of our portfolio companies;
- Monthly company all-hands meetings with cool activities and a dinner.

What are your expectations of the student?

We are looking for highly motivated students that are passionate about Machine Learning technologies and have a drive to learn. The student should be experienced in Python and comfortable with frameworks such as Tensorflow and Torch.