Research Internship - Representation Learning

NAVEE

1 Internship offer

Navee

Navee's ambition is to become the world leader in the fight against frauds and counterfeits on the Internet. We have developed a unique technology based on reverse image search allowing us to detect these frauds in all major marketplaces as well as any standalone sites and remove them.

Navee's technology has already won numerous awards: it was chosen as a favorite of Future40 (the most promising young French startups) by Forbes and Station F, and as one of the 30 innovators of 2020 for the LVMH group. Navee has attracted a dozen customers, including three of the world's largest luxury groups and the most visited rental platform in France.

Mission and Environment

You will be incorporated into our AI team, composed of the CTO and a PhD in Computer Vision. You will also benefit from the partnership we have with CentraleSupélec researchers. Your mission will consist of:

- Research project leading to the publication of a research paper/workshop submission (see below).
- 2/3 smaller applied projects to improve Navee's fraud detection and Reverse Image Search Engine.

Skills and profile

We expect the applicants to have a strong background in applied mathematics (linear algebra and statistics) and experience in the areas of machine learning/computer vision, as well as excellent programming skills. Ideally we expect the applicants to be comfortable with GNU/Linux and have a knowledge of the Python ecosystem.

Conditions

- Location: Station F, 5 Parvis Alan Turing, 75013 Paris, France.
- Salary: 1400 € brut/month.
- Possibility to continue for a CIFRE PhD thesis after the internship.

Apply

- Book a slot for an introduction call here: https://calendly.com/mathieu-daviet/30min
- Please direct any inquiries to job@navee.co

2 Research project

Navee is working with luxury brands to detect counterfeits across all the internet. All the internet is huge, it needs to be filtered to select the most relevant content.

Every day, we are receiving hundreds of thousands posts. One post is containing a list of images, a title and a description. There is a need to select posts from very specific products (examples: Rolex Daytona, Gucci Horsebit Loafer ...) so luxury brands can focus on the counterfeits mimicking the most iconic models of their collection. The idea is to build a model to help us filter the content.

There are some constraints:

- Usually, there are only between 4 to 100 images of each particular item in the dataset. It is hard to find a lot of images when we target a specific model.
- Datasets can be quite different from one brand to another, hence a necessity of *domain adaptation*.
- There are new products to detect every month as fashion trends evolve quickly.

This problem fits well into the paradigm of *semi-supervised learning* as we have a large amount of unlabeled data and *few-shot learning* as we need to adapt quickly to the new brands and products.

Related work

Few-shot learning is a type of machine learning problem where learning is done from a few examples. This means that we have to exploit the prior knowledge we have about the task and data distribution. There are multiple ways of doing this [5]:

- We can separate the approaches that *augment the dataset*: introducing data augmentations, exploit weakly-labeled or unlabeled available data points, exploit similar datasets. To this category also belong the approaches that utilize semi-supervised learning techniques for improving few-shot learning [1, 2], it will be useful in our case.
- Approaches that *constrain hypothesis space*: metric learning, generative modeling. We are interested in this category because we want to obtain good explainable representations.
- Approaches that *alter search strategy in hypothesis space*: these are various meta-learning approaches. This is the last priority for us.

Semi-supervised learning is an approach that aims to exploit unlabeled data during training. Recent state of the art often features pseudo-labeling step for the unannotated images, with pseudo-labels employed to train a representation. Multiple modern methods exploit the data-augmentation techniques to enforce the pseudo-label consistency for the predictions of augmented samples [3, 4, 6].

Internship plan

The first step is to implement a basic few-shot learning approach that learns and exploits a robust representation for the product category. It is important to identify the limits of the basic approach and the data requirements for learning this representation.

Next, we will use semi-supervised techniques to exploit the unlabeled data and improve the representation. A challenging approach is to exploit the category-level similarity between the products (i.e., condition bag representations to be similar across the brands).

The student is also expected to study the relevant bibliography, identify relevant existing approaches and propose solutions to the advisors.

Appendix: Internal dataset

We have a small internal dataset (1.7 G) with brands annotated within a tree structure. In this brand names are at the shallowest level and model instances with corresponding images are at the deepest level.

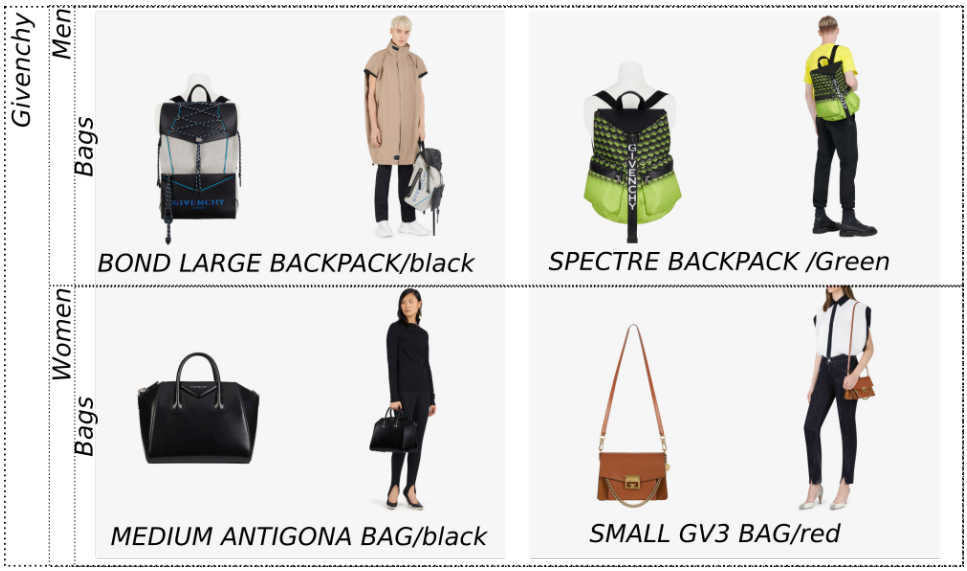

Figure 1: Examples of the fashion brand images.

```
Root
├── Givenchy
│   ├── MEN
│   │   └── BAGS
│   │       └── BACKPACKS
│   │           ├── BOND LARGE BACKPACK IN LEATHER AND CANVAS
│   │           │   └── Black
│   │           │       ├── BK506FK0SW449-02-01.jpg
│   │           │       ├── BK506FK0SW449-02-02.jpg
│   │           │       └── ...
│   │           └── SPECTRE BACKPACK IN GEOMETRIC PRINTED NYLON .7 Green
│   │               ├── BK507EK0YR734-02-01.jpg
│   │               ├── BK507EK0YR734-02-02.jpg
│   │               └── ...
│   └── WOMEN
│       └── BAGS
│           └── Antigona
│               └── MEDIUM ANTIGONA BAG
└── YSL
    └── CEINTURES
        └── Sac ceinture vintage en python laquee
            └── CARAMEL
                └── image.jpg
```

Figure 2: Example of the fashion brand tree.

References

[1] Da Chen, Yuefeng Chen, Yuhong Li, Feng Mao, Yuan He, and Hui Xue. Self-Supervised Learning For Few-Shot Image Classification. *arXiv:1911.06045 [cs]*, February 2020.

[2] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Boosting Few-Shot Visual Learning with Self-Supervision. *arXiv:1906.05186 [cs]*, June 2019.

[3] Zhongzheng Ren, Raymond A. Yeh, and Alexander G. Schwing. Not All Unlabeled Data are Equal: Learning to Weight Data in Semi-supervised Learning. *NeurIPS*, 2020.

[4] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *NeurIPS*, 2020.

[5] Yaqing Wang, Quanming Yao, James Kwok, and Lionel M. Ni. Generalizing from a Few Examples: A Survey on Few-Shot Learning. *arXiv:1904.05046*, 2020.

[6] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised Data Augmentation for Consistency Training. *NeurIPS*, 2020.