Sparse and discrete representations in latent spaces

Unsupervised representation learning is a very exciting direction in machine learning, particularly given the large amount of easily available unlabelled data. There are many machine learning algorithms (e.g. object detection, classification, reinforcement learning, model compression or novel sample generation) that would greatly benefit from low dimensional representation and highly expressive features. Sparse and discrete representation provide these properties.

Discrete and sparse representations are more interpretable than continuous and dense ones which are typically cryptic and understanding their behavior is an extremely difficult task. In situations where humans have to make decisions based on large amounts of data, interpretability of supporting models is fundamental to facilitate this task to people. More and more efforts are concentrated to create models that provide human-understandable justifications for their output. An example of the difference between the continuous dense representation and the discrete sparse representation is presented on the Figure 1. We can observe that the discrete and sparse representations give us much better possibilities of interpreting this model.

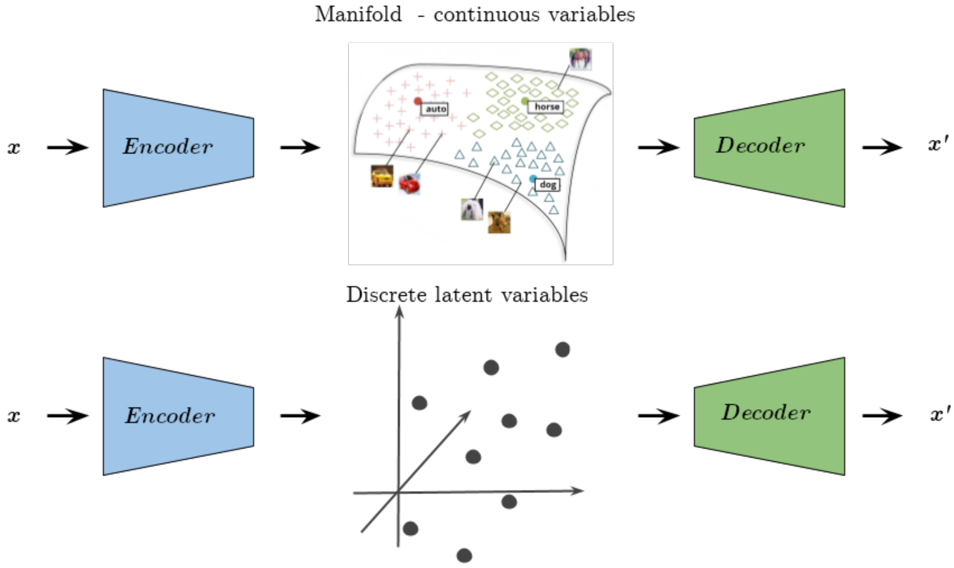

Figure 1: The figure shows two autoencoders: on the top with continuous latent variables and the bottom with discrete latent variables. Both of them take inputs as x and outputs as x'. The difference between them is that on the top model (image) each input x has its own representation in latent space, but in the second model, the representation in latent space of x can be defined by a combination of all discrete latent variables (their number is much smaller than upper model).

Discrete and sparse representations are also much more intuitive and robust than continuous and dense ones. There are many biological inspirations as well as experiments on artificial neural networks proving this statement. As many architectures are still based on continuous and dense representations, there is a necessity for exploring this topic and developing new approaches.

We consider our project as very significant because of its big potential to have positive impact on the following problems:

- *representations of data* – a large number of datasets contain inherently discrete generative factors that are difficult to detect by neural networks with continuous latent variables;
- *interpretability* – the discrete variables are easier to interpret, and sparse representation captures the data generation process itself and is biologically inspired;
- *computational efficiency* – processing sparse and discrete data requires less computer work and energy.

Creating useful sparse and discrete representations in latent spaces is a very challenging research topic. In this project, we want to address those issues. We want to propose new methods for creating discrete and sparse representations and apply them into solutions to the problems where their use will be novel and can significantly improve the results. We plan to analyze how sparse and discrete representation can help with creating self-explaining models. We will apply our methods to the various current machine learning challenges such as the big models compression (e.g, BERT model). We believe that our proposals will not only overperform accuracy of already existing models but also improve their interpretability.