THE AI REVOLUTION: DEEP LEARNING AND WHAT´S NEXT?

by Patrick Michl

30. April, 2019

When talking about the "AI Revolution" it's difficult to narrow down a common denominator. This is not only because science fiction didn't prepare us for our first real encounters with AI, but also due to the many and varied accessions, ranging from hopes to fears.

The AI Revolution is nothing more and nothing less then a rite of passage. But to know, where this journey takes us, requires to know where it started. After about one decade of deep learning it's time to take stock of progress and review some of the most important milestones and remaining challenges.

Big-Bang!

The advent of deep learning can be traced back to Geoffrey Hinton's daredevil science article "Reducing the dimensionality of data with neural networks" (Hinton et al. 2006). It's contents may kindly be summarized in two essential observations:

1. Certain undirected graphical models, termed Restricted Boltzmann Machines (RBM), can efficiently[1] be trained to represent data by maximizing their likelihood.

2. These RBMs can be stacked together to "pre-train" deep Artificial Neural Networks (ANN), which in a subsequent "fine-tuning" step generally attain much better solutions then without pre-training.

These points are not as bloodless as they appear: Essentially, they mean that almost anything[2] can be predicted with appropriate data! Data scientists usually do not have the reputation of being exuberant, but some tears of joy must have been shed with this discovery!

The Bayesian: DBM

Although Hinton's article definitively paved the way for deep ANNs, it did not yield an explanation for the use of pre-training, nor did it provide a mathematical framework to describe it. Undeterred of these shortcomings, a group about Guillaume Desjardins greatly improved Hinton's approach by welding the stack of RBMs into a single Deep Boltzmann Machine (DBM).

Their article "On Training Deep Boltzmann Machines" (Desjardins, Courville, Bengio 2012) provides a gradient based update rule for the simultaneous effective[3] training of stacked RBMs and therefore avoids losses, caused by their stacking. Thereby the stacked RBMs - the DBM - is trained to generate a latent representation of the training data, by preserving it's dependency structure. This stategy endows it with high generalizability.

Apart of these improvements, however, DBMs provide an important hint about the very nature of pre-training: DBMs generate the sample distribution of the training data by maximizing the likelihood. This can be imagined as the inflation of a manifold that clings to the data in terms of a total least squares regression. Without pre-training, however, the ANNs only perform an ordinary least squares regression, which heavily impairs their generalizability.

The Frequentist: GAN

> *"Adversarial training is the coolest thing since sliced bread"*
>
> — Yann LeCun

A further obscurity in Hinton's article was the succession of an undirected graphical model, followed by a directed - and indeed that's the daredevil part! Superficially the parameter spaces of both models may somehow seem comparably, but they are not at all! In particular with respect to the different probability distributions, they generate. But how to solve this problem? Imagine two kids sharing toys: No wonder they always quarrel! A group about Ian Goodfellow provided a fairly straight solution: Every model get's it's own parameter space!

The article "Generative Adversarial Nets" (Goodfellow et al. 2014) proposes a model, where one ANN is trained to generate the sample distribution, while another is trained to discriminate the artificially generated samples from truly observed data. Thereby the generative network tries to fool the discriminative network by increasing it's proportion of misclassifications, while the latter tries to decrease it, which is a zero-sum game.

Due to this approach GANs by the way solved a further problem of DBMs: Since the likelihood gradient of DBMs usually is not tractable it has to be estimated, either by a Markov chain or variational inference. GANs, however, do not require such estimations. The results are impressive! In particular, the photorelistic images and videos received much attention with artificial faces and deepfakes.

What will come next?

> *"Now what belongs together will grow together"*
>
> — Willy Brandt

The zoo of deep models grows exponentially! Currently we find ourselves surrounded by many promising approaches, but there is a reason why the two approaches mentioned above - DBMs and GANs - are of paramount importance: They have a fundamental and pure character of feuding schools in statistics: The Bayesians and the Frequentists.

At this point one could draw parallels to Romeo & Juliet, which raises the idea to put them together and see what happens. Lo and behold, some people already did this! First steps in this direction, e.g. "Boltzmann Encoded Adversarial Machines" (Fisher et al. 2018) impressively demonstrate, that there is a lot of potential in this fusion! This is not by chance, as both approaches show distinctive strengths, in structure and representation. So I'll take the bet: The next big thing in deep learning is the fusion of GANs and DBMs.

But let's extend the projection further into the future. There is one thing that only received very little attention in deep learning so far: Undirected graphical models like DBMs have the capability to capture dependency structures, and not only the boring linear ones, but indeed any sufficiently smooth! This property, however, has not yet been exploited at all! Why? Simply spoken, there is a large gap in the literature, as it affects statistics as well as differential geometry and topology! Nevertheless, I am convinced that the odds of deep structural inference satisfy to take the efforts to develop a completely new branch of statistics[4].

Usually I try not to get mawkish, but the prospects about the AI Revolution somehow can be overwhelming. And no matter, how important the above aspects will turn out, after all they will still only represent a tiny chapter within the long succession of incredible advances, that await us.

1. Due to the bipartite graph structure of RBMs, repeated Gibbs sampling is rapidly mixing, which allows an efficient approximation of the log-likelihood gradient. ↩

2. The observables are required to trace out sufficiently smooth and Lipschitz-continous trajectories. ↩

3. If you stack bipartite graphs together, you still get a bipartite graph. Of course, it's a little more complicated, but under the hood that's the reason, why DBMs can efficiently be trained. ↩

4. I started the journey, to merge statistics with differential geometry and topology and would be glad, if I could inspire you with my ideas: [1, 2, 3, 4], but be warned: You could get mad (or bored)! ↩