WHAT IF YOU HAD TO

INTERVIEW A MACHINE

Loria
Laboratoire lorrain de recherche en informatique et ses applications

UNIVERSITÉ DE LORRAINE

SUPERVISED LEARNING

An intelligent machine relies on datasets to learn a desired behaviour. That behaviour is encoded in the labels of the data, thanks to the interpretation of human annotators. One problem that emerges is that human bias transpires in those labels, and therefore transfers to the machine when it learns.

SUPERVISED EVALUATION

We like to quantify how a machine behaves using numerical indicators. Those indicators (accuracy, recall, f-score...) are estimated on part of the annotated data, sometimes called Ground Truth. But talking about estimation implies uncertainty in the estimates.

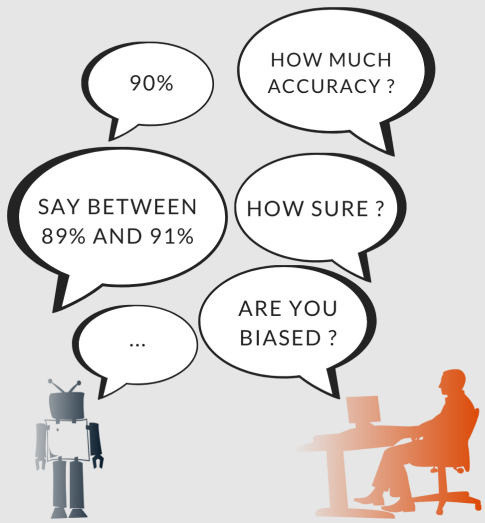
THE PROBLEM OF CONFIDENCE

On the other hand, we are witnessing our society putting a growing number of intelligent machine in different areas. Some of them have critical decision-taking power. Therefore, the notion of confidence in the behaviour of a machine becomes primordial. To quantify that in supervised evaluation, we compute confidence intervals.

ANNOTATION BIAS

But the data we use for that evaluation is also annotated, and therefore potentially biased. Knowing that, can we remain as confident about what it tells us on the behaviour of the machine ?

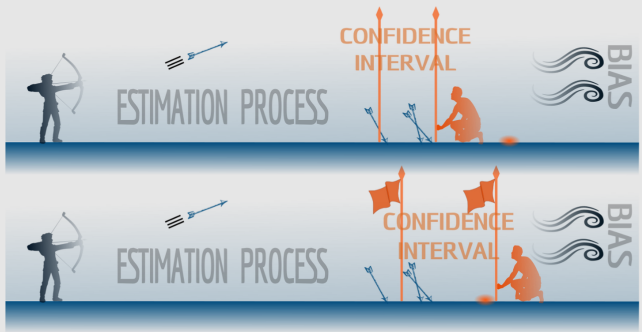
What bias means can be unclear. In the general case, let us consider that it refers to how the data labels differ from the interpretation of one entity that takes responsibility over the machine's behaviour (typically the builder).



BIAS-WISE CONFIDENCE INTERVALS

The way confidence intervals are computed only account for the uncertainty caused by testing with a limited number of data. But it is unfortunately blind to the presence of annotation bias.

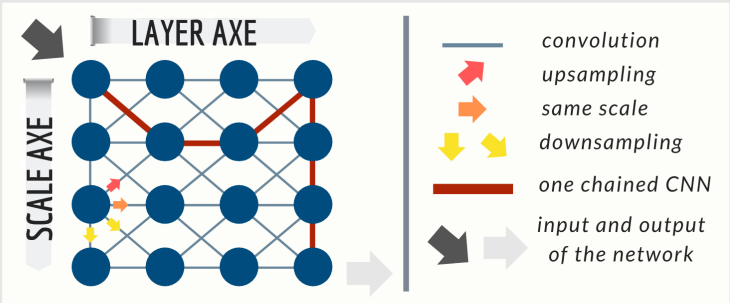
However, assuming we are able to quantify simply that bias, we can compute corrected confidence intervals.



UNSUPERVISED PERFORMANCE INDICATORS ?

Bias is hard to identify, quantify or even define. There is a need to investigate more how to detect it in annotated datasets, or to assess and mitigate its impact in the estimation phase. Another promising direction is the field of AI explainability. More precisely, what can we understand of a machine's behaviour without relying on annotated data ? Can the machine speak for itself ?

APPLICATION TO A NETWORK COMPRESSION CASE



In that spirit, we are studying the benefits of using Sensitivity Analysis tools to build such indicators and to evaluate their benefits when data is biased. In particular, we want to compare two different network pruning strategies, one that depends on the annotated data, the other based on Sensitivity Analysis tools, The purpose is to show that the second method gives better results. We apply that to the model proposed by Vebreek, the Convolutionnal Neural Fabric, which encompasses an exponentially large number of chained CNN by massively sharing weights among the convolutions, arranged in a grid.

Ilias Benjelloun - Synalp

ilias.benjelloun@loria.fr