

Peer Review

Review of: "PaRCE: Probabilistic and Reconstruction-based Competency Estimation for CNN-based Image Classification"

Jörg Martin¹

1. Physikalisch-Technische Bundesanstalt, Braunschweig, Germany

The article describes a new method for judging the reliability of an image classification produced by CNNs. The main idea is to combine several aspects into a single metric which is called PaRCE.

The overall idea seems valid and important to me.

Unfortunately, I find the theoretical description of the method really hard to grasp, which makes it impossible for me to judge the validity of the score and its evaluation in the results. Let me explain this in detail for Section 3.1 of the article. The formulas here are based on probabilistic notation. However, the formulas seem ambiguous, if not flawed, to me:

1. $f(X)$ is introduced as the “true underlying model”. This seems to refer to the conditional data distribution $p(Y | X)$ of targets given an image X which is in general assumed to allow for multiple labels for an image (cf. the works on uncertainty quantification cited by the article). However, this contradicts the statement in the next sentence where $f(X)$ is called “the true class of the image”. This might only be an imprecise formulation or simply an assumption on the data distribution (no multiple labels possible) but should be clarified.
2. This point is my main criticism: Formula (1) does not make sense. Conditional on X the objects $f(X)$ and $\hat{f}(X)$ are both either a specific class or a specific vector of probabilities adding to 1.0 (which is unclear, see point 1 above). However, in either case, they are *not random*. In particular, it is entirely unclear what a probability of an event such as $\{f(X) = \hat{f}(X)\}$ should indicate. The same is also true for equations (2), (3), and (4).

It could be that the confusions in points 1 and 2 arise from a mixture of imprecise notation and a lack of explanations. In the current form, the derivations are, however, at least from my perspective, impossible to follow, which is why I cannot judge the meaningfulness of the presented score. In case of a thorough and complete revision of the method part, I would be ready to review the article once more.

Declarations

Potential competing interests: No potential competing interests to declare.