

Breaking the cycle of data imbalance: the role of optimized loss functions

ARSLAN, Janan^{1,2}

1. Paris Brain Institute (Institut du Cerveau - ICM), UMR 7225, 75013, Paris, France.
 2. Institut National de la Santé Et de la Recherche Médicale, U 1127, 75013, Paris, France.
-

The utilization of artificial intelligence (AI) in healthcare has several benefits attached to it. When AI is applied to diagnostic imaging, it can discern image features and colors at greater bandwidth and with greater resolution than humans due to digital image processing, often discriminating features not visible to human vision. Outsourcing repetitive and time-consuming components of diagnostics to an AI-based machine or software frees up the clinician to focus on the more holistic side of medicine, including but not limited to clinical management, integrative medicine, and emotional support. AI can also quickly, and cost effectively, process megabytes of data. However, there are circumstances in which megabytes or high-quality data may not be readily available. For example, in the event of rare diseases, we may only have a handful of cases at our fingertips. There may be other circumstances in which labels and annotations required to run AI-based methods are scarce. Whole slide images are an excellent illustration of this, as the images are often large in nature, making it a challenge for pathologists to easily annotate all regions of interest; particularly small and not easily visible structures, such as blood vessels. These extreme cases can lead to severe data imbalances that often slow down the process of achieving high-performing AI models. Several pathways to overcoming such obstacles are actively being explored, including image augmentation and synthesis. Other potential solutions lie in mathematical statistics, particularly with the optimization of loss functions. Loss functions can break or make the performance of an AI-based model. For example, in the case of semantic segmentation, one loss function may generate clean and highly accurate segmentation outputs, while another can generate blurry outputs with no discernible borders. These instances can occur even when the models are trained under the same conditions and use the same hyperparameters. In this presentation, the application of several loss functions is examined in detail. The theory behind these loss functions is explored, followed by discussions on how the application of loss functions can improve situations in which data imbalances are a reality. The presentation will act as a quick guide on loss functions and conclude with future directions on how mathematical statistics can be explored to optimize our current AI learning processes.