Sound Recognition with Limited Supervision

Sound recognition aims to allow a machine recognize the various sounds happening around it. There are many possible applications such as environmental monitoring, autonomous driving and automatic captioning to name a few. Due to the large number of possible sounds and environments, however, one particular challenge for state-of-the-art data-driven machine learning solutions is to gather a sufficient amount of annotated training data matching the target application. Therefore, to keep annotation effort low, training approaches are required, which can take advantage of data which is not (fully) annotated and/or not matches the target application. In this presentation, we present our contribution in the field. We present a training approach allowing a system not only recognize sounds in an audio clip but also provide sounds' on- and offset times, although on- and offsets are not annotated in the training data. Further, we present how performance can be improved by 1) pre-training the system on mismatched data and 2) integrating completely unlabelled data, i.e., only raw audio, into the training.