Abstract:

At Facebook, machine learning provides a wide range of capabilities that drive many aspects of user experience including ranking posts, content understanding, object detection and tracking for augmented and virtual reality, speech and text translations. While machine learning models are currently trained on customized data-center infrastructure, Facebook is working to bring machine learning inference to the edge. By doing so, user experience is improved with reduced latency (inference time) and becomes less dependent on network connectivity. Furthermore, this also enables many more applications of deep learning with important features only made available at the edge. This paper takes a data-driven approach to present the opportunities and design challenges faced by Facebook in order to enable machine learning inference locally on smart phones and other edge platforms.

Abstract:

At Facebook, machine learning provides a wide range of capabilities that drive many aspects of user experience including ranking posts, content understanding, object detection and tracking for augmented and virtual reality, speech and text translations. While machine learning models are currently trained on customized data-center infrastructure, Facebook is working to bring machine learning inference to the edge. By doing so, user experience is improved with reduced latency (inference time) and becomes less dependent on network connectivity. Furthermore, this also enables many more applications of deep learning with important features only made available at the edge. This paper takes a data-driven approach to present the opportunities and design challenges faced by Facebook in order to enable machine learning inference locally on smart phones and other edge platforms.

Abstract:

At Facebook, machine learning provides a wide range of capabilities that drive many aspects of user experience including ranking posts, content understanding, object detection and tracking for augmented and virtual reality, speech and text translations. While machine learning models are currently trained on customized data-center infrastructure, Facebook is working to bring machine learning inference to the edge. By doing so, user experience is improved with reduced latency (inference time) and becomes less dependent on network connectivity. Furthermore, this also enables many more applications of deep learning with important features only made available at the edge. This paper takes a data-driven approach to present the opportunities and design challenges faced by Facebook in order to enable machine learning inference locally on smart phones and other edge platforms.