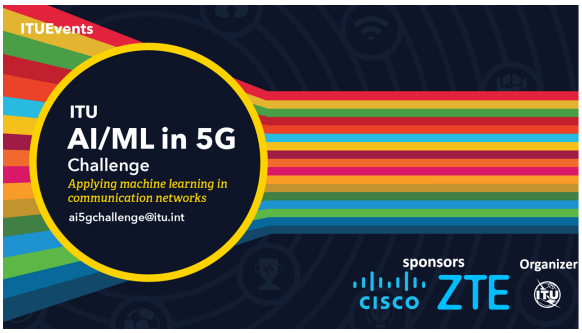
2020 DNN Inference Optimization Challenge



Adlik invites you to participate in the **ITU Artificial Intelligence /Machine Learning in 5G Challenge**, a competition which is scheduled to run from now until the end of the year. Participation in the Challenge is free of charge and open to all interested parties from countries that are members of ITU.

Detailed information about it can be found on the Challenge website, which includes the document "ITU AI/ML 5G Challenge: Participation Guidelines".

DNN Inference Optimization Challenge is organized as a part of ITU Artificial Intelligence/Machine Learning in 5G Challenge.

Background:

While Deep learning has achieved great success in many areas like audio recognition, computer vision and natural language processing, it still remains a great challenge to use DL models in environment with restrict constraints in computing cost, memory footprints or inference latency, e.g. edge computing scenarios in 5G network.

Many effective optimization technologies have been introduced to address these challenges, including model-targeted optimization, system communication optimization and hardware specific optimization. Model-targeted optimization mainly focus on compression of deep learning models, examples methods are model pruning, kernel sparseness, quantization, low rank decomposition, knowledge distillation, etc. System communication optimization methods like DNN partition tries to accelerate model inference by optimizing communication between different computing nodes or layers. Hardware specific optimization like TensorRT optimizes inference operations based on the hardware characteristics.

Task:

This problem statement focuses on the construction of general model optimization technology. Participants are required to design a general model optimization algorithm to achieve model acceleration. The target models are as follows:

- BERT
- MobileNet-V3
- ResNet 50

Participants can choose any model version and dataset as they need, and then design their own model compression or other optimization solutions, which can either be a single algorithm or a system with multiple algorithms intergrated.

Submitting:

Participants must create a private Github repository to contain their work and submit by adding AdlikChallenge as a collaborator. The repository will be made open to the judges after submission deadline and should be accessible till the end of the final event of the ITU challenge.

Things need to be submitted are as follows:

- Source code, including the whole optimzation solution and performance test demo.
- A description document. It should describe how to verify the optimization performance with the source code. Other contents of the document include but are not limited to: insight, opinion and analysis of model optimization; selected target model and reason; solution, algorithm used; description and comparison of optimization results, etc.

Evaluation criteria:

- Effect of model optimization (50%): Reasonable trade-off between accuracy and efficiency. The selected model type, loss of accuracy, compression rate of model parameters and computation will be taken into account.
- Solution advantage (30%): Whether the solution is reasonable and whether the solution has enough practicability, innovation and universality.
- Problem analysis (10%): Whether there is a deep and original insight into the problem, and whether the analysis of the key elements of the problem is accurate and reasonable.
- Completeness (10%): Whether the requirements of the competition are fulfilled according to the proposed scheme and design.

Resource:

BERThttps://github.com/google-research/bert

MobileNet-V3https://arxiv.org/abs/1905.02244https://github.com/topics/mobilenet-v3

ResNet 50: https://github.com/KaimingHe/deep-residual-networks

How to participate?

1. If you don't have an ITU account, please follow the guidance to create one for challenge registration.
2. Register on ITU AI/ML in 5G challenge website with your ITU account.
3. Fill out the ITU AI/ML in 5G Challenge Participants Survey to select problem statement **ITU-ML5G-PS-018**. You can enroll as a team with 1-4 members.
4. Begin to work on this problem and submit your results. We will begin to accept submissions from **July 1st, 2020** and the submission deadline is ~~**October 10th, 2020**~~ extended to **October 15th, 2020**.

Contact:

ITU contact: ai5gchallenge@itu.int

DNN Inference Optimization Challenge contact: yuan.liya@zte.com.cn

You can also visit our Slack Channel to find more guidance.