**Machine Learning for large-scale video flow classification: Improving classification performance**

This thesis project involves Machine Learning applied to large data sets with practical relevance. Currently, traffic in the Internet comes from a range of applications such as Web surfing, Video viewing, online gaming, Spotify and VoIP (Skype) conversations. These applications have different demands on the network with regards to throughput and delay. Inside the network it can therefore be useful to differentiate between different classes of applications. One method to do this is to look inside packet data to classify a network flow to an application type, using Deep Packet Inspection (DPI).

However, end to end encryption is becoming more common for Internet traffic, making DPI impossible for flow classification. Another possible approach to classify flows is using packet size distributions, packet direction distribution and similar features that are not dependent on looking inside packets, and thus can be used also for encrypted traffic. Using machine learning it is possible to train models that based on such characteristics are able to classify flows into different application classes.

Previous KaU thesis works have looked into several aspects of using machine learning to classify flows. Two initial thesis works created an infrastructure for processing the large data sets involved and performed initial machine learning test using the SVM and Random forest approaches. Last year’s two thesis works examined the classification and run-time performance of two variations of forest based classifiers in more detail, and implemented and evaluated a high-speed C-based forest classifier. The proposed thesis work will continue previous work by exploring avenues for improving the machine learning classification performance. Possible avenues for exploration include using clustering to improve the quality of training data, and further feature engineering to tweak the feature set used for classification. The exact topic will be settled in discussion with the prospective students and Sandvine.

This thesis work is done with Sandvine (ex Procera Networks), but much of the work will be done in Karlstad with a couple of visits to company premises in Varberg and Malmö. The work will be done using Python and associated libraries. Prior Python experience is not a requirement, but good general programming proficiency and willingness to learn Python is important. Previous Machine Learning experience is not required, but strong curiosity for the area is advantageous. This thesis work is the continuation of the very successful thesis cooperation between Sandvine and KAU, where four previous KAU CivIng students already have successfully completed their thesis work using Python and Machine Learning techniques.

The technical work part of the thesis will be jointly supervised, with a Sandvine senior technical engineer and in close cooperation with KaU researchers active in the area. The writing part of the thesis will be supervised by the KaU supervisor (i.e one of the active researchers).

Monetary compensation will be provided from Sandvine for students participating in this thesis project.

For questions or more details, please contact Johan Garcia at [johan.garcia@kau.se](mailto:johan.garcia@kau.se), or visit me at room 21F426.