# NXP Konzultace - Computer Vision 3

## Souhrnná výzkumná zpráva

Tomáš Milet, Roman Juránek

## Konzultace k akceleraci algoritmů

**Odhad výkonnosti NN.**
Ze strany NXP padl požadavek k odhadu výkonnosti typů neuronových sítí. Jednalo se především o konvoluční sítě obsahující vrstvy jako 2D konvoluce, plně propojená vrstva, pooling vrstvy apod. Odhad měl počítat s různými platformami (procesory, grafické karty pomocí OpenCL). Ze strany fakulty došlo k doporučení algoritmů pro odhad nejlepší možné výkonnosti a nejhorší možné výkonnosti. Výkonnost neuronových sítí závisí na implementaci inferenčního enginu a hardware, na kterém běží. Fakulta doporučila postup jakým spočítat odhad s přihlédnutím k implementaci daného inferenčního enginu a hardware na kterém běží.

**Akcelerace pomocí OpenCL.**
Další požadavek ze strany NXP padl ohledně doporučení k akceleraci algoritmů (ne jen neuronových, sítí, ale i obecných algoritmů) v prostředí OpenCL na grafických kartách. Fakulta doporučila postupy práce s jednotlivými částmi GPGPU. Jednalo se o správné navržení velikosti pracovních skupin, sdílené paměti a množství registrů pro optimalizaci obsazení multiprocesorů na grafické kartě. Dále fakulta doporučila postupy práce s lokální pamětí, aby se minimalizovaly bankové konflikty a aby nedošlo k přetěžování lokální paměti na úkor jiných druhů pamětí. Dále fakulta doporučila postupy, jak a kdy synchronizovat vlákna na grafické kartě a navrhnout algoritmy pro zamezeni divergence vláken. Další doporučení se týkalo atomických operací zvláště v kontextu lokálních a globálních atomických instrukcí a jejich možnou náhradou pomocí změny algoritmů (například prefixovou sumou). Fakulta dále doporučila práci s warpovými operacemi (jako je ballot instrukce a shuffle instrukce), přístupy ke globální paměti a práci s dalšími částmi grafické karty (jako jsou texturovací jednotky).

**Měření výkonnosti OpenCL**
Další požadavek od NXP se týkal ohledně měření výkonnosti OpenCL algoritmů a jak zabránit nepřesným a kolísavým měřením. Fakulta doporučila postupy vestavěné v OpenCL a postupy které použít, pokud OpenCL měření selže. Jednalo se především o doporučení ohledně úsporného režimu grafických karet a podtaktování pamětí a jader GPU. Fakulta doporučila frameworky vhodné pro profilování kernelů na grafické kartě jak v podobě knihoven, tak v podobě aplikací pro různé platformy a vestavěných instrukcí v kernelu.