Paralelní trénování hlubokých neuronových sítí

Bc. Ondřej Šlampa

Fakulta informačních technologií

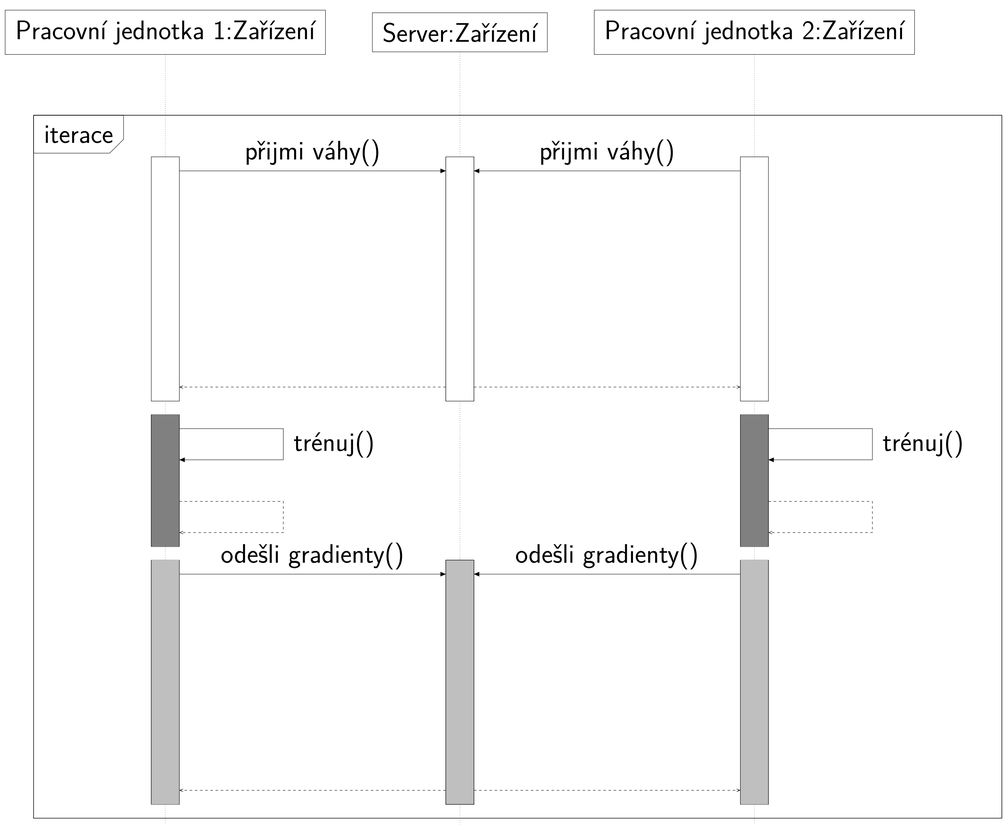
May 24, 2017

Synchronní trénování

Hlavní částí mé práce je vytvoření způsobu jak odhadovat výhodnost požití distribuovaného trénování pro danou síť. To je založeno na tom, že trénování je možné rozdělit na výpočet gradientů a komunikaci. Délku výpočtu gradientů je možné naměřit na jedné jednotce. Délka komunikace je možné vypočítat podle počtu vah. Z těchto dvou hodnot je možné vypočítat délku celého trénování.

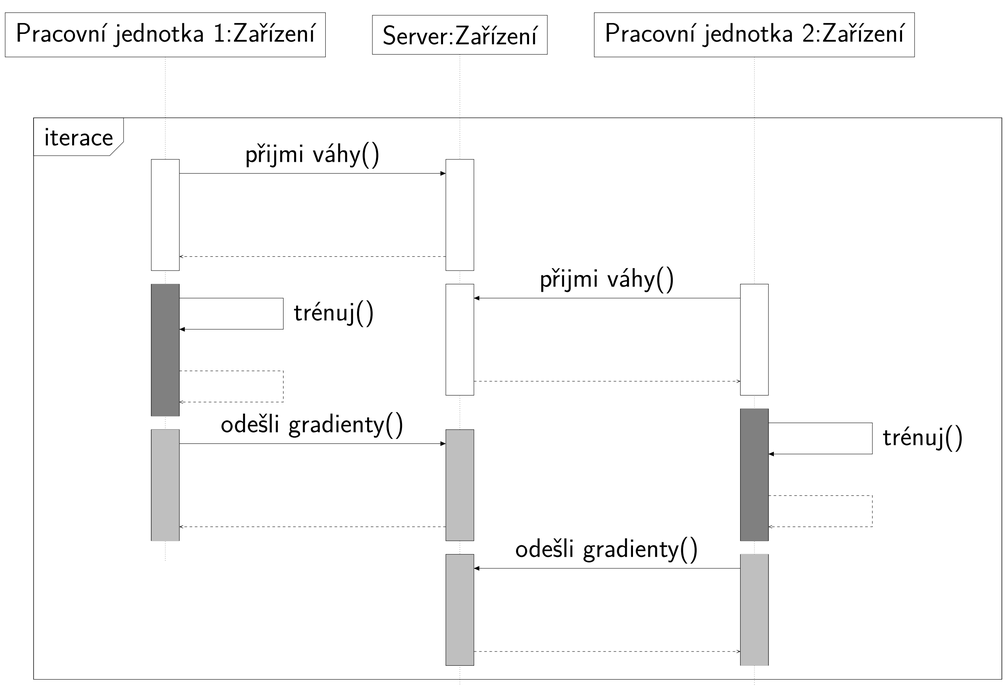
Nejhorší rozložení komunikace

Obě jednotky komunikují se serverem současně, to způsobí, že rychlost přenosu dat bude poloviční a délka přenosu dvojnásobná.



Nejlepší rozložení komunikace

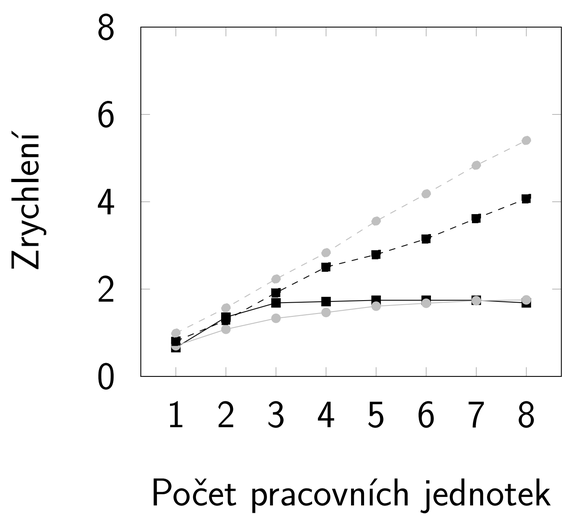
Při tomto rozložení dochází k maximalizaci prokládání výpočtů a komunikace. Pokud jedna pracovní jednotka komunikuje se serverem, druhá počítá nebo čeká.



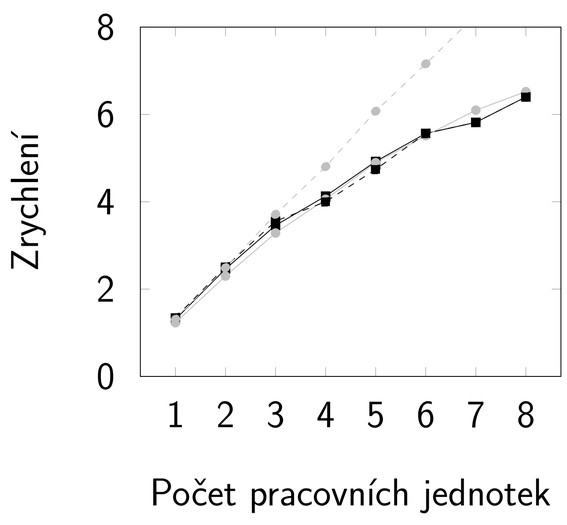
Porovnání odhadů a měření synchronního trénování

Úvahy ilustrované v předchozím bloku jsem použil k výpočtu odhadů zrychlení výpočtu na několika pracovních jednotkách. To je zobrazené na následujících grafech.

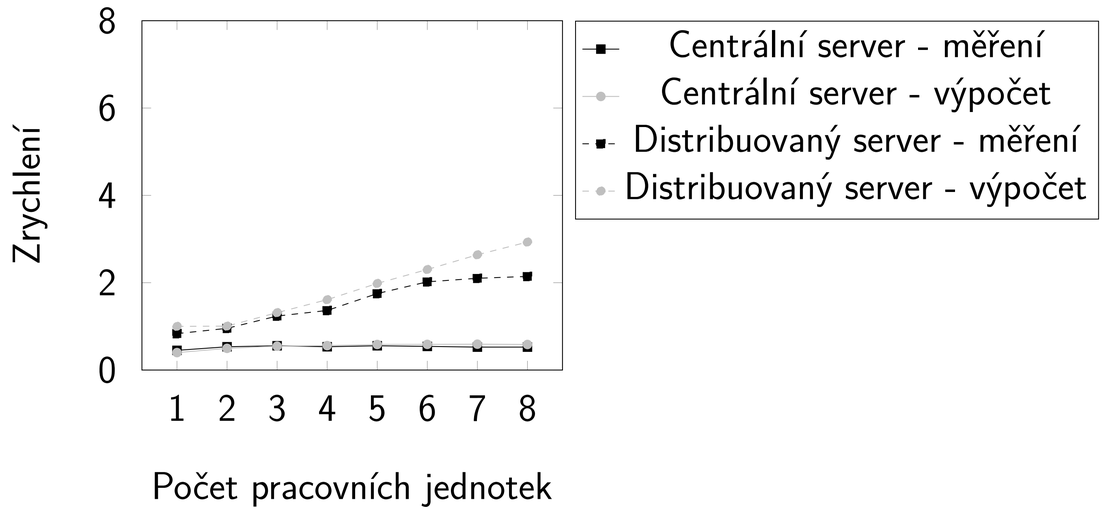
GoogLeNet



SqueezeNet



Resnet34

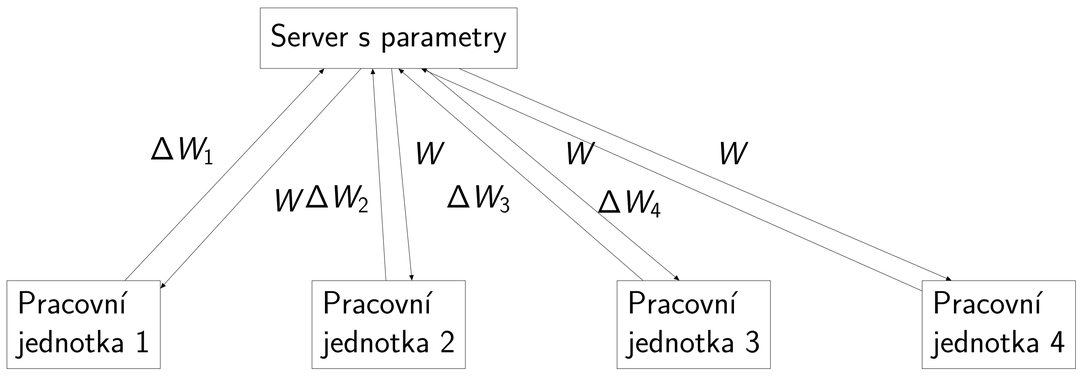


Uložení vah

Ve své práci jsem popsal dva způsoby, jak sdílet váhy neuronové sítě. Váhy musí být sdíleny mezi všemi jednotkami, aby každá jednotka měla přístup k nejnovějším váhám, které použije pro výpočet gradientů.

Centrální server

Je mechanizmus uložení vah neuronové sítě na jedné dedikované jednotce. Ostatní jednotky sdílejí váhy přes tento server. Server provádí aktualizace vah sítě. Problém je, že když více jednotek přenáší data ze serveru nebo na server, komunikace je zpomalená.



Distribuovaný server

Cílem distribuovaného serveru je odstranit hlavní slabinu centrálního serveru. Váhy sítě jsou rozděleny na N částí, každá část je uložené na jiné jednotce. Každá jednotka je tak server i výpočetní jednotka.