Klasifikace scén a tagování obrázků.

FIT - VG20102015006 - 2014 – 02

Ing. Michal Hradiš, Ing. Aleš Láník

Fakulta informačních technologií, Vysoké učení technické v Brně

6. ledna 2015

Abstrakt

Pro zpracování obrazových dat z různých zdrojů a jejich další organizaci je důležitá informace o typu zachycené scény, případně další informace popisující daný obraz nebo úsek videa. Tyto další informace mohou být ve formě tagů, tedy jednoznačných kategorií (objekty, činnosti, události) popisujících daný mediální objekt. V rámci projektu VideoTerror jsme zkoumali hluboké konvoluční sítě, které v současnosti dosahují nejlepších výsledků v klasifikaci fotografií. Tato zpráva popisuje vlastnosti použitých sítí, jejich chování na různých úlohách a připravené technické řešení.

Obsah

1. Konvoluční sítě v počítačovém vidění

Konvoluční sítě jsou dopředné sítě, které ve svých vrstvách používají některé specifické operace a jejichž vstupem je zpravidla přímo obraz a ne předzpracované příznaky, jak je zvykem u ostatních sítí používaných v počítačovém vidění. Nejcharakterističtější operací konvolučních sítí je očekávatelně konvoluce, která aplikuje matici vah lokálně na určitou číst obrazu a tuto operaci provádí v nezměněné podobě v celém prostoru obrazu (konvoluce mohou být i jiných řádů). Výstup jedné takové konvoluce se nazývá kanál a sdružení několika konvolucí a jim odpovídajícím kanálům pak tvoří vrstvu konvoluční sítě. Základní myšlenkou konvolučních sítí je předpoklad stacionality obrazu, tedy že obraz má stejné vlastnosti v celé své ploše a že je možné ho zpracovávat stejným způsobem nehledě na pozici. Tento předpoklad je jednoznačně správný ve vrstvách blízkých obrazu. Ve vyšších vrstvách už pak může být vhodné zpracovávat různé části obrazu různým způsobem, ne což se využívají plně propojené vrstvy nebo lokálně propojené vrstvy (které ale už nejsou konvoluční).

Konvoluce je lineární operace, a vícevrstvé sítě proto musí mezi konvoluce vkládat nelinearity stejně jako v ostatních vícevrstvých dopředných sítí, které nejsou konvoluční. Kromě klasických nelinearit, jako jsou sigmoidy a hyperbolický tangens, se v konvolučních sítích často používají po částech lineární funkce (rektifikované lineární jednotky (Krizhevsky et al., 2012) a max-out (Goodfellow et al., 2013)). Tyto po částech lineární jednotky umožňují rychlejší konvergenci a pomáhají minimalizovat problém mizejících gradientů.

Další nedílnou součástí konvolučních sítí jsou pooling vrstvy. Tyto vrstvy explicitně vnášejí do sítě invarianci proti malým posunutím a jsou realizovány aplikací statistických funkcí na malá okolí a případným následovným zmenšením prostorového rozlišení. Často se používá funkce maxima (max-pooling) a v některých případech funkce průměru (avg-pooling).

Další vrstvy využívané konvolučními sítěmi jsou lokální normalizace odezev (local response normalization) a dropout (Krizhevsky et al., 2012). Lokální normalizace odezev může například normalizovat lokální odezvy, aby měly konstantní energii. Tato normalizace může být aplikována pro každý konvoluční kanál zvlášť nebo i přes kanály. Efektem takovéto normalizace je inherentně vyšší odolnosti sítě proti změnám osvětlení, případně vzájemné potlačování lokálně nemaximálních odezev neuronů.

Konvoluční neuronové sítě mají dlouhou historii v oblasti počítačového vidění. Již na konci sedmdesátých let Kunihiko Fukushima (1980) navrhl první funkční sítě, které mají mnoho společného s konvolučními sítěmi používanými v současné době. V druhé polovině devadesátých let dosáhl významných úspěchů s konvolučními sítěmi Yan LeCun a další (1998) na problému rozpoznávání ručně psaných číslic. Toto řešení bylo úspěšně nasazeno například pro čtení směrovacích čísel při přepravě a třídění zásilek.

Problémem spojeným s konvolučními sítěmi, ale i s ostatními dopřednými sítěmi trénovanými pomocí optimalizace objektivní funkce na základě gradientů parametrů získanými pomocí zpětného šíření chyb, byla po dlouhou dobu nemožnost trénovat hluboké sítě, protože gradienty mají tendenci během propagace sítí mizet nebo naopak explodovat. V takových případech je velmi obtížné hluboké sítě efektivně trénovat. Tento problém v roce 2006 z části vyřešili Salakhutdinov a Hinton (2006) pomocí předtrénování jednotlivých vrstev

sítě jako Restricted Boltzmann Machine a pospojování takových vrstev do jedné sítě, která může být dále optimalizována v klasickém supervised režimu. Tato unservised inicializace umožnila trénovat i velmi hluboké sítě s více než desítkou vrstev, ale našla si jen menší využití ve spojení s konvolučními sítěmi.

Výrazný průlom ve využití konvolučních sítí znamenalo vítězství Krizhevkého a dalších z University of Toronto v Large Scale Visual Recognition Challenge 2012 s využitím velkých konvolučních sítí trénovaných na GPU (Krizhevsky et al., 2012). Tato síť, která měla 60 milionů parametrů 650 tisíc neuronů, výrazně přesahovala svou velikostí dosud ostatní dosud použité sítě, dosáhla v ILSVRC 2012 úspěšnosti 15.3 %, což znamenalo výrazný posun oproti do té doby používaným přístupům založeným na vizuálních slovech (úspěšnost 26.2 %). Tento úspěch nastartoval zájem o konvoluční sítě a od té doby byly obdobné sítě aplikovány ve velké škále problémů.

Přístupy založené na konvolučních sítí v současné době dosahují nejlepších výsledků v úlohách klasifikace obrazu (Szegedy et al., 2014), lokalizace obličejových bodů (Sun et al., 2014), odhad pózy lidského těla z fotografie (Toshev et al., 2014), rozeznávání osob podle obličeje (Taigman et al., 2014), detekci objektů (Girshick et al., 2013) a mnoha dalších.

2. Nástroje a výpočetní možnosti

Pro experimenty s konvolučními sítě, naše skupina používá framework Caffe[1]. Caffe je napsané v C++ s využitím CUDA a cuDNN [2] a je možné je take používat z jazyků Python a Matlab. Caffe obsahuje základní stavební prvky konvolučních sítí (konvoluční vrstvy, plně propojené vrstvy, lokální normalizaci odezev, dropout, max-pooling, …), základní optimalizační metody (stochastický gradientní sestup, AdaGrad (Duchi et al., 2011) a Nesterov's accelerated gradient (Nesterov, 1983)) a připravené zdroje dat. Caffe umožňuje efektivní trénování konvolučních sítí na GPU. Bohužel v současnosti neumožňuje distribuované trénování sítí na více výpočetních uzlech, což limituje velikost sítí, s jakými můžeme efektivně pracovat. Caffe si podle potřeby upravujeme.

Pro experimenty máme k dispozici stroje s GPU zapojené ve výpočetním gridu na Fakultě informačních technologií, který řídí Sun Grid Engine. Následující tabulka shrnuje GPU, které jsou v rámci gridu dostupné.

Dále využíváme přístup na výpočetní grid Anselm v rámci projektu IT4Innovations. Anselm obsahuje 23 výpočetních uzlu s jednou NVIDIA Tesla Kepler K20.

Tabulka 1 – Seznam GPU zapojených do výpočetního gridu na Fakultě informačních technologií VUT.

Počet	GPU	GPU RAM
5	GeForce GTX 770	4 GB
8	GeForce GTX 980	4 GB

[1] http://caffe.berkeleyvision.org/

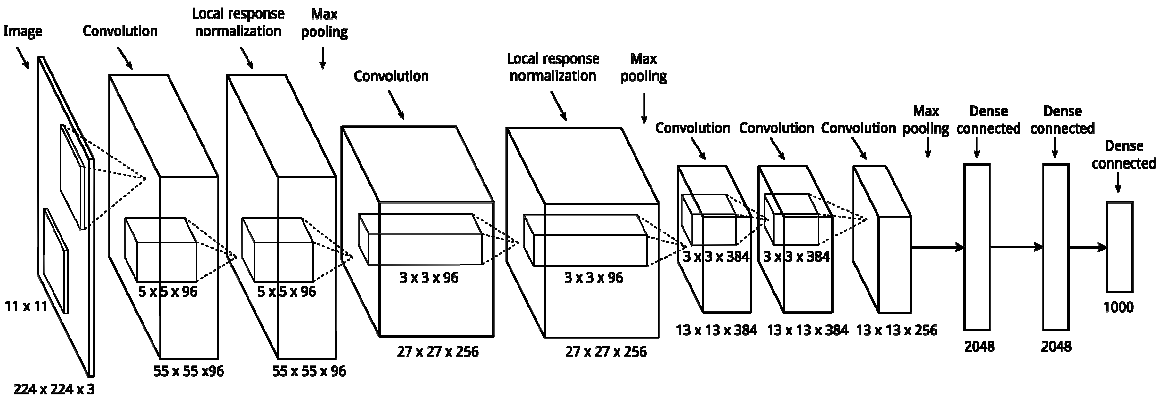
[2] https://developer.nvidia.com/cuDNN

3. Klasifikace obrazu

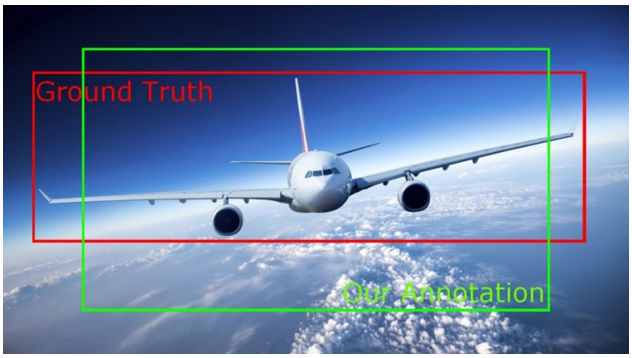
V roce 2014 jsme se zúčastnili soutěže ILSVRC v úlohách klasifikace obrazu a lokalizace. Výsledný systém kombinuje několik konvolučních sítí a generování variant testovacích obrázků pomocí geometrických a barevných transformací.

Natrénovali jsme 14 variant konvolučních sítí, které vycházely z architektury Krizhevskeho a dalších (Krizhevsky et al., 2012). Základní architektura je zobrazena na Obrázek 1. Varianty architektur se lišily ve velikosti filtrů první vrstvy (9x9, 11x11, 15x15, 17x17), velikosti lokální normalizace odezev po první vrstvě (5x5, 7x7, 9x9, 11x11, 15x15) a velikosti lokální normalizace po druhé vrstvě (3x3, 5x5, 7x7, 9x9). Výsledky sítí byly zprůměrovány po umocnění s mocninou, která byla nastavena na validační sadě na hodnotu 0.5.

Pro lokalizaci jsme zvolili okno pevné velikosti a pozice uprostřed obrazu zabírající 55 % obrazu. Tato strategie se ukázala jako velmi úspěšná a překonala mnoho sofistikovaných přístupů (úspěšnost lokalizace 51.99 %). Použité okno je zobrazena na Obrázek 2.



Obrázek 1 - Architektura konvoluční sítě vycházející z (Krizhevsky et al., 2012), kterou jsme použili pro ILSVRC 2014.



Obrázek 2 – Demonstrace lokalizačního okna (zeleně) s ukázkou anotace objektů (červeně).

Tabulka 2 – Klasifikační výsledky na validační sadě ImageNet.

Architektura	**Kombinace**	**Top-5 přesnost klasifikace**
Caffe reference + AlexNet	Suma	81,62 %
Našich 14 sítí	Suma	80,88 %
Našich 14 sítí	Suma odmocnin	81,05 %
Caffe ref. + AlexNet + 14 sítí	Suma odmocnin	82.01 %

4. Klasifikace scén

Pro experimenty s klasifikací scén jsme zvolili datovou sadu SUN397[3], která obsahuje 397 kategorií scén, které jsou podmnožinou SUN databáze. SUN397 obsahuje celkově 108 754 obrázků a každá kategorie obsahuje minimálně 100 obrázků.

Při klasifikaci scén jsme použili síť s architekturou sítě Krizhevskeho a dalších (Krizhevsky et al., 2012) z ILSVRC 2012 natrénovanou na ImageNet datasetu pro inicializaci vah konvolučních vrstev sítě. Celá síť pak byla optimalizována na padesáti obrázcích z každé kategorie datasetu SUN397. Výsledná úspěšnost klasifikace takto natrénované sítě je zobrazena v Tabulka 3.

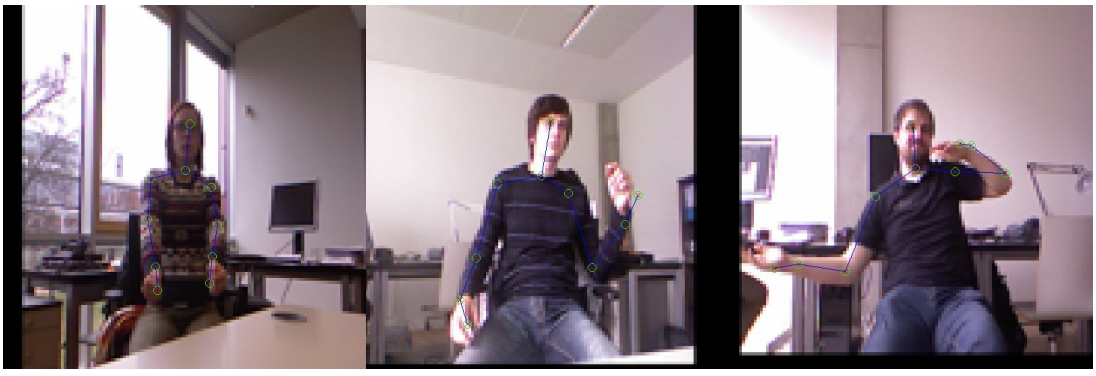
Tabulka 3 – Úspěšnost na klasifikace na datové sadě SUN937.

Systém	Přesnost klasifikace
Naše konvoluční síť	**48.5 %**
(Sánchez et al. 2013)	47.2 %
(Xiao et al. 2014)	42.7 %
(Xiao et al. 2010)	38.0 %

5. Odhad lidské pózy z fotografie

Pomocí konvolučních sítí je možné řešit i regresní problémy. Taková změna v mnoha případech vyžaduje jen výměnu výstupní vrstvy sítě a zbytek sítě může zůstat nezměněn. Tohoto faktu již bylo využito v mnoha aplikacích. Například Toshev et al. (2014) tento přístup použili pro odhad pozic kloubů lidské kostry z fotografií a pojmenovali ho DeepPose. Podobný postup pro lokalizaci obličejových bodů z fotografií publikovali Sun et al. (2013).

Konvoluční sítě jsme využili pro odhad lidské pózy z videa a fotografií se zaměřením na rozpoznávání aktivity pilotů v kabině letadla. Ukázkové obrázky z experimentální datové sady jsou zobrazeny na Obrázek 3.

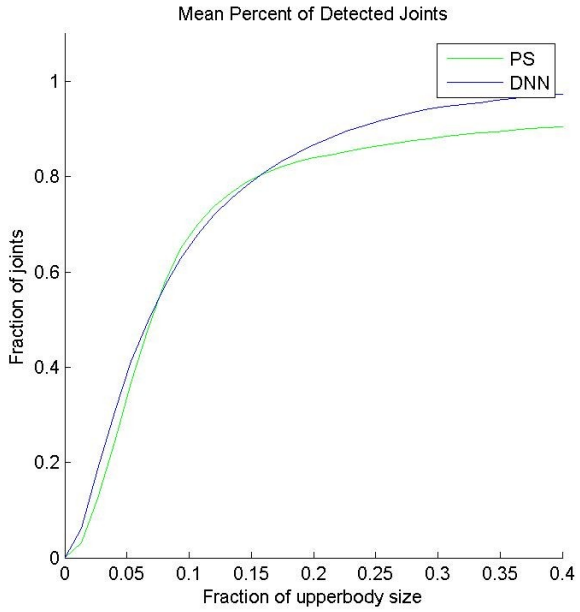


Obrázek 3 – Ukázka dat pro testování odhadu lidské pózy z fotografie. V obrázcích je vykreslen i výstup konvoluční sítě.

[3] Databáze SUN397 – http://vision.princeton.edu/projects/2010/SUN/

Vstupem sítě jsou obrázky zarovnané podle pozice hlavy a s konstantní velikostí, protože se předpokládá, že pilot sedí na pevné židli, které neumožňuje přibližování a oddalování od kamery. Obrázky byly zmenšeny na velikost 128x128 pixelů. Použitá síť má v první vrstvě 64 filtru 11x11aplikovaných s krokem 3, které jsou následovány RELU a max-pooling pooling vrstvou s krokem 2 a lokální normalizací odezev s okolím 5x5. Druhá konvoluční vrstva používá 128 filtrů s velikostí 5x5 a je také doplněná max-pooling vrstvou a lokální normalizací. Třetí vrstva má 256 filtrů 3x3. Čtvrtá a poslední konvoluční vrstva má 196 filtrů 3x3 a je následována max-pooling vrstvou s krokem 2. Následují dvě plně propojené vrstvy s velikostmi 1024 a 2048. Poslední vrstva je výstupní a není doplněna nelinearitou. Tato vrstva má 20 lineárních neuronů, které přímo odpovídají souřadnicím (x, y) jednotlivých kloubů v obraze. Pozice kloubů byly normalizovány do rozsahu 0-1.

Výsledky konvoluční sítě jsme porovnali s výsledky deformovatelných modelů částí, kde části jsou detekovány pomocí rozhodovacích stromů (Dantone et al. 2013). Průměrné výsledky pro všechny části těla naleznete na Obrázek 4.



Obrázek 4 – Výsledky odhadu pozice částí lidského těla z fotografie. Graf zobrazuje jakou poměrnou část všech kloubů je daný systém schopný lokalizovat s určitou chybou. Chyby jsou normalizovány vůči velikosti lidského torza. PS je deformovatelný model (Dantone et al. 2013). DNN je hluboká konvoluční síť.

6. Reference

(Dantone et al., 2013) Dantone et al. Human Pose Estimation using Body Parts Dependent Joint Regressors. CVPR 2013.

(Duchi et al., 2011) J. Duchi, E. Hazan, and Y. Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. The Journal of Machine Learning Research, 2011.

(Fukushima, 1980) Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 1980.

(Girshick et al., 2013) Girshick et al. Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR 2014.

(Goodfellow et al., 2013) Goodfellow et al. Maxout Networks. ICML 2013.

(Krizhevsky et al., 2012) Krizhevsky, A., Sutskever, I. and Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks, NIPS 2012.

(LeCun et al., 1998) LeCun et al. Gradient-Based Learning Applied to Document Recognition. Proc. IEEE 1998.

(Nesterov, 1983) Y. Nesterov. A Method of Solving a Convex Programming Problem with Convergence Rate O(1/k√). Soviet Mathematics Doklady, 1983.

(Salakhutdinov and Hinton, 2006) Salakhutdinov and Hinton: An Efficient Learning Procedure for Deep Boltzmann Machines. Neural Computation 2012.

(Sun et al., 2014) Sun et al. Deep Convolutional Network Cascade for Facial Point Detection, CVPR 2013.

(Taigman et al., 2014) Taigman et al.: DeepFace: Closing the Gap to Human-Level Performance in Face Verification, ECCV 2014.

(Toshev et al., 2014) Toshev et al. DeepPose: Human Pose Estimation via Deep Neural Networks. CVPR 2014.

(Xiao et al., 2010) J. Xiao et al. SUN Database: Large-scale Scene Recognition from Abbey to Zoo. CVPR 2010.

(Xiao et al., 2013) Sanchez et al. Image Classification with the Fisher Vector: Theory and Practice. 2013.

(Xiao et al., 2014) J. Xiao et al. SUN Database: Exploring a Large Collection of Scene Categories. IJCV, 2014.