Prohledávání dokumentů podle automaticky extrahovaných vzorů

Josef Baloun[1]

1 Úvod

Prohledávání dokumentů podle automaticky extrahovaných vzorů je problém z oblasti *information retrieval*. Lze ho také nalézt pod pojmy jako *rozpoznávání* nebo *vyhledávání slov v ručně psaných dokumentech*. V anglické literatuře je nejčastěji použit termín *word spotting*.

Toto prohledávání je děleno podle vstupního kritéria, kterým může být obrazový vzor (dále jako *QbE*) nebo textový řetězec (dále jako *QbS*). Při prohledávání se snažíme na základě vstupního kritéria nalézt obrázky s odpovídajícím slovem. Úspěšné vyřešení by usnadnilo práci např. v situaci, kdy má historik najít v kronice, která obsahuje 1 000 stran textu, všechny zmínky o klášteře v Teplé.

Cílem této práce je návrh a implementace metod, které umožní vyhledávání těchto slov. Předpokladem je předem provedená segmentace dokumentu na obrázky slov. K dosažení tohoto cíle jsou použity neuronové sítě.

2 Databáze

Pro trénování a vyhodnocení je použita databáze Parzival (více viz Fischer et al. (2012)). Pro potřeby této práce obsahuje normalizované obrázky slov s přepisy a nadefinované sady trénovacích, validačních a testovacích dat.

3 Architektury neuronových sítí

Vedle experimentální sítě byly navrženy tři architektury. První je konvoluční síť (dále jako *cnn*), kde jsou vstupní obrázky upraveny na stejný rozměr. Síť čerpá z Karpathy et al. (2017) a je složena z devíti konvolučních vrstev (32, 64 nebo 128 filtrů o velikosti 3 x 3, 1 x 1 stride, padding a ReLU aktivační funkce). Za druhou, čtvrtou, šestou a devátou konvoluční vrstvou následuje vždy jedna max pooling vrstva (2 x 2 poolsize, 2 x 2 stride). Za nimi následují tři fully-connected vrstvy (první dvě mají 3072 a 2048 neuronů s ReLU aktivační funkcí, poslední výstupní vrstva má 952 neuronů a sigmoidu jako aktivační funkci).

Druhá architektura (dále jako *spp*) využívá Spatial Pyramid Pooling vrstvu (viz He et al. (2014)), díky které se síť dokáže vypořádat s různými rozměry vstupních obrázků. Architektura vychází z *cnn*, kde je místo poslední max pooling vrstvy použita Spatial Pyramid Pooling vrstva (1 x 1, 2 x 2 a 4 x 4 mřížka).

Třetí architektura (dále jako *lstm*) je založena na konvoluční LSTM (viz Shi et al. (2015)), kde je vstupní obrázek převeden pomocí posuvného okna na sekvence. Architekturu tvoří tři vrstvy s konvoluční LSTM, za kterými následují max pooling vrstvy. Výstup tvoří jedna fully-connected vrstva (952 neuronů a sigmoida jako aktivační funkce).

[1] student bakalářského studijního programu Inženýrská informatika, obor Informatika, e-mail: balounj@students.zcu.cz

4 Vyhodnocení

Výše popsané modely neuronových sítí *cnn*, *spp* a *lstm* slouží pro odhadnutí PHOC vektoru (viz Almazán et al. (2014)) k danému obrázku. Na základě porovnávání těchto vektorů jsou vraceny odpovědi pro jednotlivé dotazy. Pro vyhodnocení (viz tab. 1) je použita evaluační metrika Mean Average Precision (dále jako *MAP*) pro QbS i QbE dotazy. Další metrikou je přesnost (dále jako *acc*), která představuje správné zařazení obrázků do tříd slov. Protože v některých článcích vyhodnocují MAP pouze pro dotazy obsažené v trénovací části sady, jsou přidány i výsledky pro tento postup jako *T-MAP QbS* a *T-MAP QbE*.

	MAP QbS	MAP QbE	acc	T-MAP QbS	T-MAP QbE
cnn	92,62 %	90,01 %	90,51 %	95,28 %	90,63 %
spp	90,54 %	87,57 %	85,19 %	92,34 %	87,78 %
lstm	70,92 %	67,86 %	78,73 %	81,93 %	70,29 %

Tabulka 1: Výsledky naměřené na testovací části sady z databáze Parzival

5 Závěr

Nejlepších výsledků dosáhla architektura *cnn* (viz tab. 1), se kterou se podařilo na dotazy podle vzoru i řetězce (viz obr. 1) vrátit relevantní odpovědi v datové kolekci Parzival. Ukázalo se tak, že je možné tento systém použít v praxi, pokud bude splněna podmínka úspěšné segmentace dokumentu a připraveno dostatek trénovacích dat.

Obrázek 1: Odpověď na QbS dotaz pro slovo *aventivre* seřazená zleva (zelená je správně, červená je chybně)

Poděkování

Tato práce vznikla za podpory projektů CERIT Scientific Cloud (LM2015085) a CESNET (LM2015042) financovaných z programu MŠMT Projekty velkých infrastruktur pro VaVaI.

Literatura

Almazán, J., Gordo, A., Fornés, A., Valveny, E. (2014) Word spotting and recognition with embedded attributes. *IEEE transactions on pattern analysis and machine intelligence*.

Fischer, A., Keller, A., Frinken, V., Bunke, H. (2012) Lexicon-free handwritten word spotting using character HMMs. *Pattern Recognition Letters*.

He, K., Zhang, X., Ren, S., Sun, J. (2014) Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Karpathy et al. (2017) *CS231n Convolutional Neural Networks for Visual Recognition*. Available from: `http://cs231n.github.io/` [Accessed 28th November 2017].

Shi, X., Chen, Z., Wang, H., Yeung, D., Wong, W., Woo, W. (2015) Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *CoRR*.