Modulární neuronové sítě: metody zvyšující robustnost sémantického popisu scény

Gabriela Šejnová, Michal Vavrečka

Český institut informatiky, robotiky a kybernetiky
Jugoslávských partyzánů 1580/3, 160 00 Praha 6
Email: gabriela.sejnova@cvut.cz, michal.vavrecka@cvut.cz

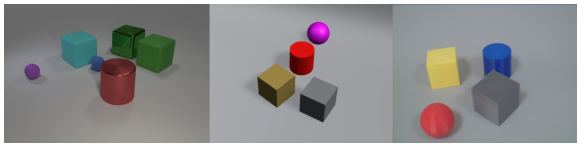
Abstrakt

Nejnovější modely založené na modulárních neuronových sítích dosahují lidské přesnosti v úloze Visual Question Answering (zodpovídání otázek k obrázkům). Ačkoliv zvládají zodpovídat v přirozeném jazyce i velmi komplexní otázky, mezi odpověďmi chybí základní logická konzistence (tedy 1+1=2). Právě proto používáme ucelenost popisu sémantické scény jako metriku pro další zlepšení současných architektur. V článku se soustředíme na závislost přesnosti odpovědí na úhlu pohledu na scénu. Ukazujeme, že současné modely nezvládají generalizovat napříč různými úhly a navrhujeme vlastní upravený dataset, který tuto přesnost zvyšuje.

1 Úvod

Jedním z intenzivně zkoumaných problémů umělé inteligence je propojení mezi jazykem a obrazem, tedy sémantický popis viděného. Jednou ze současných úloh testujících tuto schopnost je Visual Question Answering (VQA), kdy je modelu prezentován obrázek a otázka v přirozeném jazyce týkající se jeho obsahu, přičemž na výstupu je očekávána správná odpověď (Agrawal a spol. (2017)). V posledních několika letech dosáhly state-of-the-art modely velkého pokroku v této úloze, na benchmarkovém datasetu CLEVR dosahují přesnosti až 98% (Mascharka a spol. (2018), Suarez a spol. (2018), Perez a spol. (2018)). Klíčovou vlastností těchto architektur je jejich kompozicionální struktura - jedná se o set specializovaných neuronových sítí (modulů), přičemž každá je trénovaná pro jednu logickou operaci (např. filtrování vlastnosti objektů). Tyto moduly jsou poté sekvenčně řetězeny dle konkrétní otázky tak, aby selektivní pozorností mohly z obrazu odvodit správnou odpověď (Johnson a spol. (2017a)).

V této práci se zaměřujeme na jednu ze slabin těchto modelů - logickou inkonzistenci mezi odpověďmi pro jednu scénu. Je-li odpověď na otázku "Kolik je na obrázku objektů?" číslo 4, musí dávat odpovědi na otázky "Kolik vidíš krychlí/válců/koulí?" také součet 4 (za předpokladu, že existují jen tyto tři tvary). Dle našeho předešlého výzkumu (Sejnova a spol. (2018)) dosahují současné modely pouze 56% konzistentnosti na syntetickém CLEVR datasetu, po natrénování na námi upravené verzi tohoto datasetu (CLEVR COUNT) až 97% přesnosti, při testování na reálných fotografiích jsou ale výsledky značně horší. Při převodu do reálného světa totiž není dodržen fixní úhel kamery, z které byl generován původní dataset. Naše výsledky ukazují, že pro zvýšení konzistence na reálných datech je třeba vnést variabilitu v úhlech pohledu již do trénovací fáze.



Obr. 1: Porovnání mezi použitými datasety. Zleva: CLEVR (originál), GYM dataset a reálná data (pouze pro testování).

2 Současné datasety a jejich vylepšení

Jedním z benchmarkových datasetů zaměřených na řetězení logických operací při zodpovídání otázek je syntetický dataset CLEVR (Johnson a spol. (2017a)). Skládá se ze scén vygenerovaných v programu Blender, obsahujících jednoduchá primitiva s omezeným spektrem vlastností (3 tvary, 2 materiály, 2 velikosti a 8 barev). Otázky, generované automaticky, jsou směřovány právě na tyto vlastnosti ("Jakou barvu má velký válec napravo od žluté koule vzadu?") a prostorové vztahy mezi objekty. Protože jsou kladeny náhodně, jsou při trénování zapojeny všechny funkční submoduly, chybí však kontrola distribuce jednotlivých typů otázek. Z dříve provedené analýzy (Sejnova a spol. (2018)) vyplynulo, že modul pro počítání (*count* module) je v datasetu zapojený méně a vede tedy modely k nepřesnému počítání, zejména co se týče logické konzistence.

Z tohoto důvodu jsme nejprve vytvořili adaptaci původního CLEVR datasetu, nazvanou *CLEVR COUNT*, která sestává ze stejného setu obrázků, avšak otázky jsou zaměřeny pouze na počítání buď celkového počtu objektů, nebo objektů dané vlastnosti ("Kolik vidíš objektů? Kolik koulí? Kolik krychlí?"). Samotné natrénování vybraného state-of-the-art modelu TbD (Mascharka a spol. (2018)) na tomto datasetu zvýšilo logickou konzistenci o desítky procent (viz Tab. 1). Při

Train	Test	Count objects	Count shapes	Count color
CLEVR	COUNT	64.5 (56.1)	94.9 (57.1)	99.6 (62.9)
CLEVR	GYM(3)	69.5 (35.6)	95.8 (63.1)	92.2 (37.6)
CLEVR	REAL(3)	51.2 (17.2)	84.5 (41.1)	87.3 (19.0)
COUNT	COUNT	99.6 (97.4)	98.4 (97.5)	100 (99.4)
COUNT	GYM(3)	97.0 (87.8)	99.1 (95.6)	98.4 (88.9)
COUNT	REAL(3)	85.9 (55.2)	90.0 (71.8)	95.2 (64.4)
GYM	COUNT	12.5 (0.1)	37.6 (2.2)	42.3 (0.1)
GYM	GYM(2)	**100.0 (99.4)**	**99.9 (99.9)**	**99.9.0 (99.4)**
GYM	REAL(3)	51.5 (17.2)	84.7 (41.1)	87.1 (19.0)

Tab. 1: Přesnost počítání objektů u jednotlivých modifikací TbD modelu (REAL = reálná data). První sloupec označuje dataset, na kterém byl model trénován, druhý sloupec testovací dataset (v závorce číslo nejpřesnějšího viewportu) a zbylé tři přesnost počítání v dané kategorii. Hodnota v závorce odpovídá logické konzistenci. Hodnoty jsou v procentech.

testování na reálných datech bylo nasnímáno celkem 150 scén, každá z 9 různých (předem daných) úhlů. Snímání proběhlo pomocí robotického manipulátoru Kuka IIWA LBR 7 s připevněnou kamerou Basler Dart (5 MPx, 25 FPS). Pro ovládání ramene a akvizici snímků byl použit robotický operační systém ROS (Quigley a spol. (2009)). Logická konzistence vypočtená na reálných datech měla podstatně slabší výsledky - možným důvodem bylo zatížení trénovaného modelu na konkrétní, fixní úhel kamery, z něhož byla scéna v CLEVR datasetu snímána.

Dalším krokem byla tvorba datasetu podobného CLEVRu, generovaného ve virtuálním prostředí OpenAI Gym (Brockman a spol. (2016)), kdy každá scéna je zachycena z 9 rozdílných úhlů (prac. název GYM dataset). Výsledkem je dataset s 90 000 obrázky (10 000 scén x 9 viewportů) a 1 080 000 trénovacími otázkami (12 otázek/scénu zaměřených na počítání objektů v kategoriích *všechny objekty*/*tvary*/*barvy*). Zmíněný TbD model jsme natrénovali jak na původním CLEVR datasetu, tak i na našem CLEVR COUNT datasetu a na posledním zmíněném GYM datasetu. Výslednou architekturu jsme testovali z hlediska logické konzistence na obrázcích z CLEVR a GYM datasetu, ale i na reálných datech. Výsledky můžete vidět v Tab. 1.

3 Výsledky a závěr

Po natrénování TbD modelu (Mascharka a spol. (2018)) na GYM datasetu s 9 viewporty dosahovala přesnost odpovědí na stejném datasetu ve všech kategoriích přibližně 100 %, což je několikaprocentní zlepšení oproti naší předchozí verzi modelu natrénovaném na COUNT datasetu. Testování na ostatních datasetech COUNT (s CLEVR obrázky) a REAL však ukázalo přesnost naopak výrazně nižší oproti oběma předchozím verzím. Takovýto rozdíl je pravděpodobně způsoben sníženou variabilitou nového GYM datasetu, který obsahoval pouze objekty stejné velikosti a materiálu a v renderu se neměnilo osvětlení tak, jako v originálním CLEVR datasetu. Dalším krokem tedy bude vylepšit kvalitu renderu a variabilitu objektů v GYM datasetu natolik, aby model dosahoval stejně vysoké přesnosti na všech třech datasetech.

Poděkování

Tento příspěvek vznikl za podpory Studentské grantové soutěže ČVUT, č. SGS18/205/OHK3/3T/37, grantu TAČR TL02000362 a projektu INAFYM - CZ.02.1.01/0.0/0.0/16.019/ .

Reference

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Parikh, D. a Batra, D. (2017). Vqa: Visual question answering. *International Journal of Computer Vision*, 123(1):4–31.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J. a Zaremba, W. (2016). Openai gym. *arXiv preprint arXiv:1606.01540*.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L. a Girshick, R. (2017a). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. V *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, str. 1988–1997. IEEE.

Johnson, J., Hariharan, B., van der Maaten, L., Hoffman, J., Fei-Fei, L., Zitnick, C. L. a Girshick, R. B. (2017b). Inferring and executing programs for visual reasoning. V *ICCV*, str. 3008–3017.

Mascharka, D., Tran, P., Soklaski, R. a Majumdar, A. (2018). Transparency by design: Closing the gap between performance and interpretability in visual reasoning. V *Proceedings of the IEEE conference on computer vision and pattern recognition*, str. 4942–4950.

Perez, E., Strub, F., De Vries, H., Dumoulin, V. a Courville, A. (2018). Film: Visual reasoning with a general conditioning layer. V *Thirty-Second AAAI Conference on Artificial Intelligence*.

Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R. a Ng, A. Y. (2009). Ros: an open-source robot operating system. V *ICRA workshop on open source software*, vol. 3, str. 5. Kobe, Japan.

Sejnova, G., Tesar, M. a Vavrecka, M. (2018). Compositional models for vqa: Can neural module networks really count? *Procedia computer science*, 145:481–487.

Suarez, J., Johnson, J. a Li, F.-F. (2018). Ddrprog: A clevr differentiable dynamic reasoning programmer. *arXiv preprint arXiv:1803.11361*.