Detekce objektů s využitím textových dotazů

Filip Majer[1]

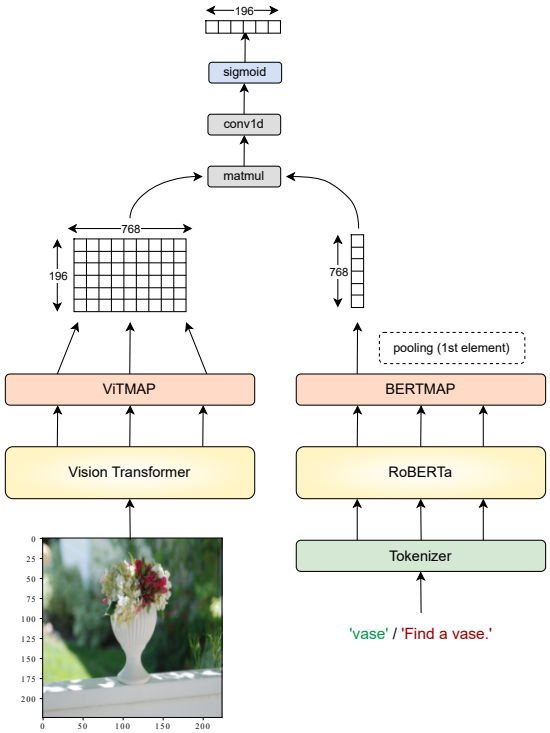
1 Úvod

V posledních letech byl zaznamenán velký vývoj v oblasti tzv. multimodálního strojového učení. To se zabývá algoritmy, které jsou schopny učit se z dat využívajících více modalit jako je text, obraz, nebo zvuk. Inspiraci nachází v tom, jakým způsobem lidé získávají informace ze světa. Ti při sběru dat ze svého okolí využívají více smyslů jako je například zrak nebo sluch.

Do kategorie multimodálních úloh spadá hlavní cíl této práce, a sice navrhnout architekturu modelu, který bude detekovat objekty v obraze s využitím textových dotazů. První vstup bude obrázek, druhý vstup bude textový dotaz na daný obrázek. Dotazy budou v angličtině a mohou být v jednoslovném tvar (“vase“) nebo ve tvaru věty (“Find a vase.“). Výstupem bude označení hledaného objektu, pokud se na obrázku vyskytuje.

2 Transformer architektura

V práci jsou využity modely založené na Transformer architektuře představené v článku Vaswani et al. (2017). Takové modely mají velké množství parametrů a vyžadují velké množství dat pro natrénování. Při trénování se tedy využívá metoda self-supervised learning, která umožňuje použít velké množství dat z internetu bez anotací a předtrénovat Transformer modely tak, že v sobě mají uchované složité jazykové/obrazové závislosti, které by u menších modelů nebylo možné zachytit. Předtrénované Transformer modely je pak možné využít jako inicializaci a dotrénovat je na konkrétní úlohu pro data s anotacemi.



Obrázek 1: Navržená architektura

3 Návrh architektury

Navržená architektura je zobrazena na obrázku 1. Pro zpracování vstupního obrázku byl použit předtrénovaný model ViT_{BASE} z článku Dosovitskiy et al. (2020). Pro zpracování dotazu byl použit předtrénovaný model

[1] student bakalářského studijního programu Kybernetika a řídicí technika, specializace Umělá inteligence a automatizace, e-mail: fmajer@students.zcu.cz

$\text{RoBERTa}_{\text{BASE}}$ a odpovídající Tokenizer z článku Liu et al. (2019).

Výstupem obrazové části je 196 vektorů dimenze 768, kde každý vektor odpovídá jednomu výřezu dimenze 14x14 vstupního obrázku. Výstupem textové části je jediný vektor dimenze 768, který reprezentuje hledanou kategorii. Skalárním součinem vektorů výřezů s vektorem z textové části získáme pro každý výřez ve vstupním obrázku skóre, které vyjadřuje, jestli se na daném výřezu obrázku hledaná kategorie vyskytuje.

Takto navržená architektura představuje originální přístup k řešení dané úlohy a hlavním cílem práce bylo otestovat, jestli se model učí požadované závislosti a je schopen úlohu řešit.

4 Trénování

Při trénování byl použit dataset COCO Detection 2017 z článku Lin et al. (2014), který obsahuje 80 kategorií. Zastoupení kategorií v datasetu není rovnoměrně rozložené. Pro trénování bylo připraveno 42 dotazů ve formě vět. Ukázalo se, že nemělo smysl trénovat oba Transformer modely, ale pouze model ViT.

5 Dosažené výsledky

Bylo zjištěno, že architektura se učí požadované závislosti a velice dobře detekuje objekty, které průměrně zabírají alespoň 5-10% obrázku.

Problémy má architektura s menšími objekty a také s kategoriemi, které jsou v datasetu málo zastoupené. Natrénovaný model také občas poskytoval falešné pozitivní výsledky, což znamenalo, že něco označoval i v případě, že mu byl položen dotaz na kategorii nevyskytující se na obrázku.

Funkčnost natrénovaného modelu lze ověřit položením dotazu, který nebyl součástí předpřipravených vět. Výstup dotazu “Show me a motorcycle in this picture“ lze vidět na obrázku 2.

Obrázek 2: Ověření funkčnosti modelu

Nejlepší natrénovaný model je možné si vyzkoušet na adrese `https://huggingface.co/spaces/fmajer/T-BOD`.

Literatura

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020) *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*.

Lin, T., Maire, M., Belongie, S., Bourder, L., Girshick, R., Hayes, J., Perona, P., Ramanan, D., Zitnick, C. L., Dollár, P. (2014) *Microsoft COCO: Common Objects in Context*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019) *RoBERTa: A Robustly Optimized BERT Pretraining Approach* .

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017) *Attention Is All You Need*.