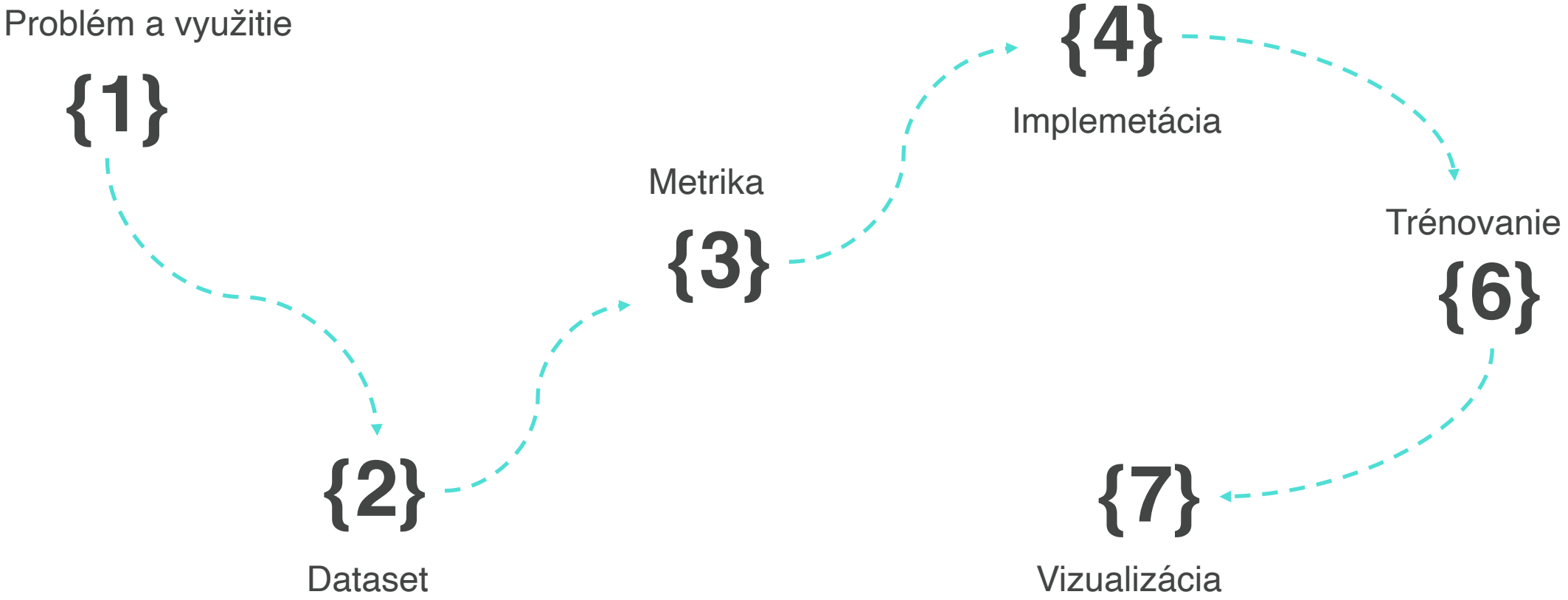
Projekt z predmetu Praktikum z neuronových sietí pre počítačové videnie

Reidentifikácia vozidiel v snímkach z dopravných kamier

Autor: Richard Dominik
FMFI UK 07.02.2022
www.st.fmph.uniba.sk/~dominik3/

O čom bude dnešná prezentácia ?



Problém a využitie

- zhoda rovnakého vozidla na snímkach z veľkého datasetu obrázkov (nasnímané pomocou dopravných kamier)
- rôzne kamery, orientácie, čas, lokácie, oklúzie, nezaostrenosť, podobný tvar, model, farba, výrobca...
- príbuzné k téme reidentifikácie osôb
- využitie v inteligentných dopravných systémoch (efektívnejšie navrhovanie dopravných sietí, vieme identifikovať auto v rôznych bodoch mesta)
- reidentifikácia vozidiel != detekcia vozidiel

Dataset VeRi-776

- 50 000+ obrázkov
- 776 rôznych vozidiel
- nasnímané pomocou 20 kamier
- rôzne pohľady, rozlíšenia, svetelné podmienky, oklúzie
- anotácie (Bbox, typ, farba, značka)

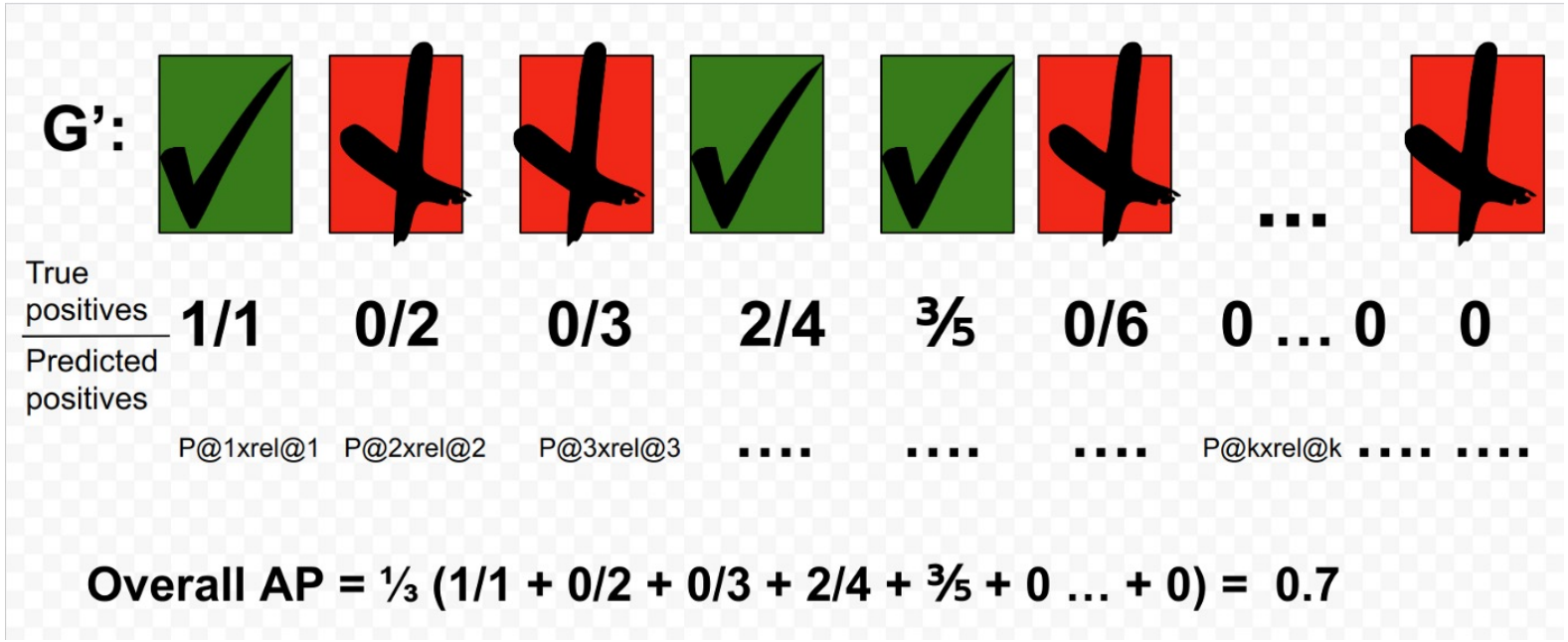
Dataset VeRi-776

Metrika mAP

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

rel@k je funkcia relevancie, 1 ak rank@k je relevantný, 0 ak rank@k nie je relevantný



$$\text{AP@}n = \frac{1}{\text{GTP}} \sum_{k}^{n} \text{P@}k \times \text{rel@}k$$

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i$$

Implementácia

- na arxiv.org publikované 08.02.2021
- podl'a paperswithcode 4te miesto v rámci state of the art
- Resnet, DeiT, ViT backbone
- https://github.com/damo-cv/TransReID

TransReID: Transformer-based Object Re-Identification

Shuting He[1,2*] Hao Luo[2] Pichao Wang[2] Fan Wang[2] Hao Li[2] Wei Jiang[1†]
[1]Zhejiang University [2]Alibaba Group
{shuting_he,jiangwei_zju}@zju.edu.cn {michuan.lh,pichao.wang,fan.w,lihao.lh}@alibaba-inc.com

Abstract

Extracting robust feature representation is one of the key challenges in object re-identification (ReID). Although convolution neural network (CNN)-based methods have achieved great success, they only process one local neighborhood at a time and suffer from information loss on details caused by convolution and downsampling operators (e.g. pooling and strided convolution). To overcome these limitations, we propose a pure transformer-based object ReID framework named TransReID. Specifically, we first encode an image as a sequence of patches and build a transformer-based strong baseline with a few critical improvements, which achieves competitive results on several ReID benchmarks with CNN-based methods. To further enhance the robust feature learning in the context of transformers, two novel modules are carefully designed. (i) The jigsaw patch module (JPM) is proposed to rearrange the patch embeddings via shift and patch shuffle operations which generates robust features with improved discrimination ability and more diversified coverage. (ii) The side information embeddings (SIE) is introduced to mitigate feature bias towards camera/view variations by plugging in learnable embeddings to incorporate these non-visual clues. To the best of our knowledge, this is the first work to adopt a pure transformer for ReID research. Experimental results of TransReID are superior promising, which achieve state-of-the-art performance on both person and vehicle ReID benchmarks. Code is available at https://github.com/heshuting555/TransReID.

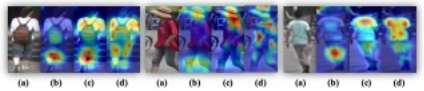


Figure 1: Grad-CAM [34] visualization of attention maps: (a) Original images, (b) CNN-based methods, (c) CNN+attention methods, (d) Transformer-based methods which captures global context information and more discriminative parts.

Figure 2: Visualization of output feature maps for 2 hard samples with similar appearances. Transformer-based methods retain backpack details on output feature maps in contrast to CNN-based methods, as noted in red boxes. For better visualization, input images are scaled to size 1024 × 512.

1. Introduction

Object re-identification (ReID) aims to associate a particular object across different scenes and camera views, such as in the applications of person ReID and vehicle ReID. Extracting robust and discriminative features is a crucial component of ReID, and has been dominated by CNN-based methods for a long time [19, 37, 36, 44, 42, 5, 12, 13, 53, 15].

By reviewing CNN-based methods, we find two important issues which are not well addressed in the field of object ReID. (1) Exploiting the *rich structural patterns in a global scope* is crucial for object ReID [54]. However, CNN-based methods mainly focus on small discriminative regions due to a Gaussian distribution of effective receptive fields [29]. Recently, attention modules [54, 6, 3, 21, 1] have been introduced to explore long-range dependencies [45], but most of them are embedded in the deep layers and do not solve the principle problem of CNN. Thus, attention-based methods still prefer large continuous areas and are hard to extract multiple diversified discriminative parts (see Figure 1). (2) Fine-grained features with *detail information* are also important. However, the downsampling operators (*e.g.* pooling and strided convolution) of CNN reduce spatial resolution of output feature maps, which greatly affect the discrimination ability to distinguish objects with similar appearances [37, 27]. As shown in Figure 2, the details of the backpack are lost in CNN-based feature maps

*This work was done when Shuting He was intern at Alibaba

Implementácia

- Natrénovanie TransReID[1] s ViT backbone (kód je implementovaný vo frameworku Pytorch)
- Rozšírenie o Swin Transformer backbone
- Vizualizácia výsledkov a porovnanie
- Google Cloud (NVIDIA Tesla T4 16GB)
- TransReID s ViT backbone – 120 epôch
- TransReID so Swin backbone – 120 epôch

[1]linky na fork uvedeného repozitára s potrebnými úpravami sa nachádzajú na webovej stránke o tejto diplomovej práci

Trénovanie

ViT backbone

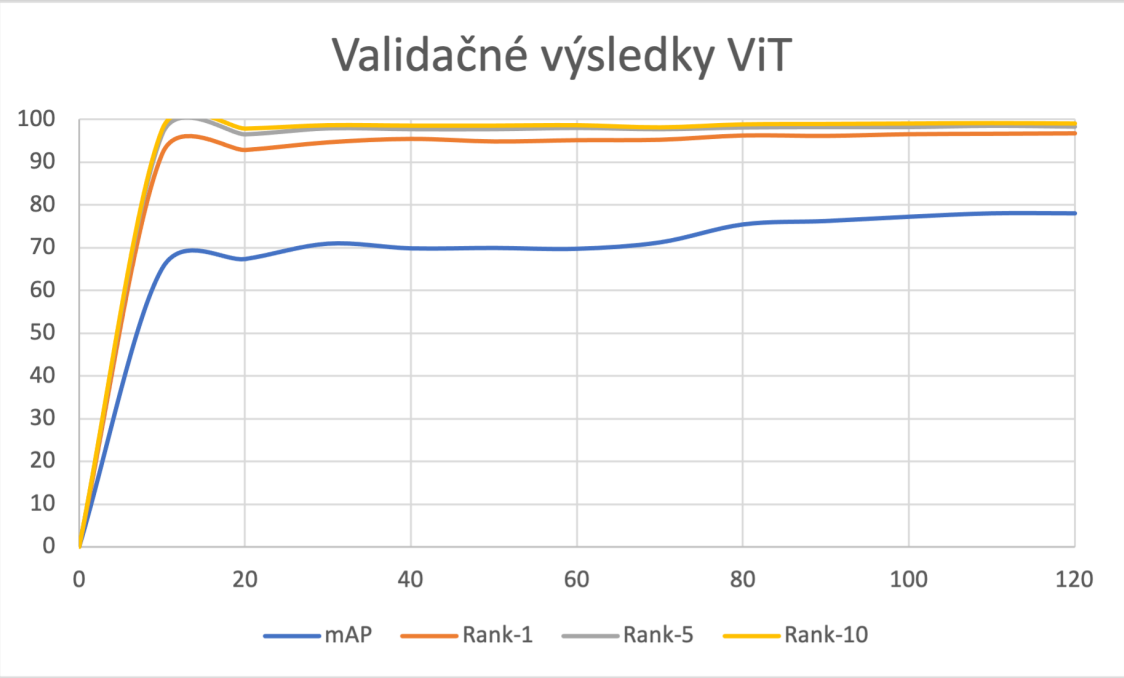
- predtrénovaný na ImageNet - 21K
- ~ 42 hodín
- ~ 9752MiB GPU
- mAP: 78.1%* so SIE a JPM
- Rank-1: 96.7%

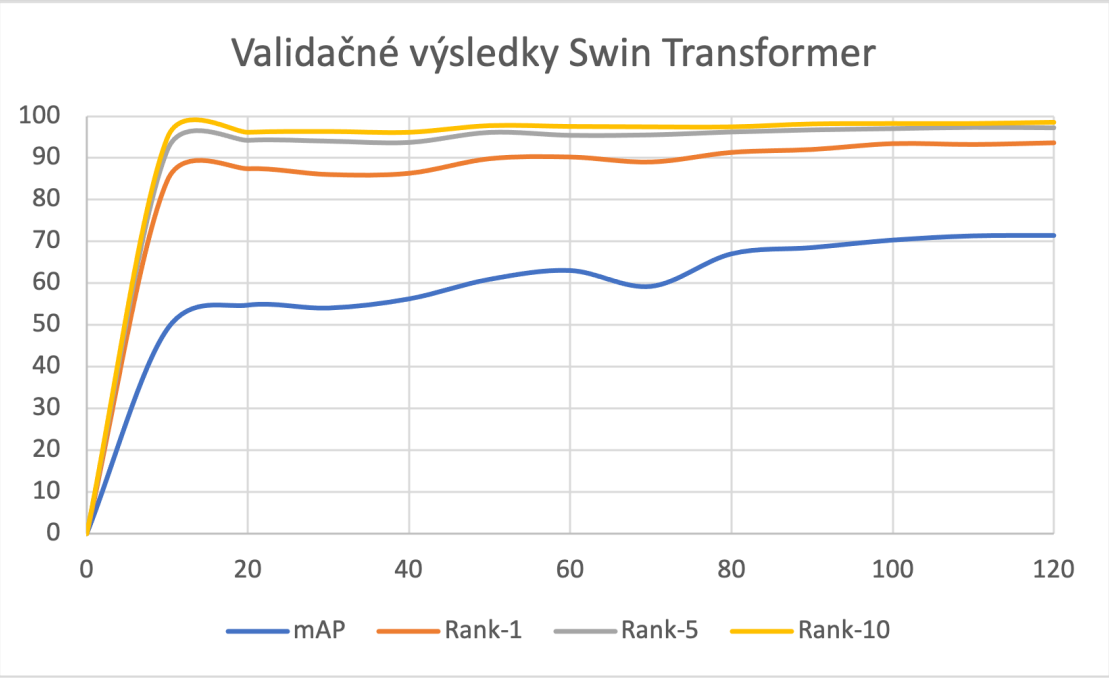
Swin backbone

- predtrénovaný na ImageNet-22K
- ~ 27 hodín
- ~ 5636MiB GPU
- mAP: 71.5%
- Rank-1: 93.6%

*v publikácií 80.6% mAP, ale baseline 78.2% mAP

Validačné výsledky





Vizualizácia

Swin backbone - úspešne

ViT backbone - neúspešné

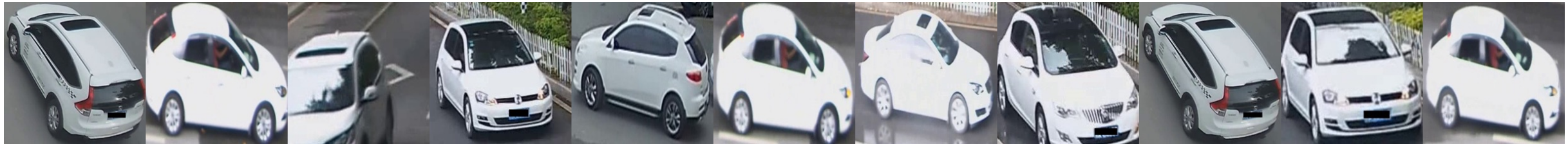
Vizualizácia

Swin backbone - neúspešné

ViT backbone - úspešne

Vizualizácia

Swin backbone - neúspešné



ViT backbone - neúspešné

Ďakujem za pozornosť !

Q&A