Obrana od backdoor napada na multimodalne modele

Autor: Dominik Jambrović

Voditelji: prof. dr. sc. Siniša Šegvić, dr. sc. Ivan Grubišić, mag. ing. Ivan Sabolić

Sadržaj

Uvod

Multimodalno učenje

- omogućava rad s **više modaliteta** (slika, tekst, audio…)
- **samonadzirano učenje** na velikoj količini javno dostupnih podataka
- veoma osjetljivo na napade

Potencijalni napadi

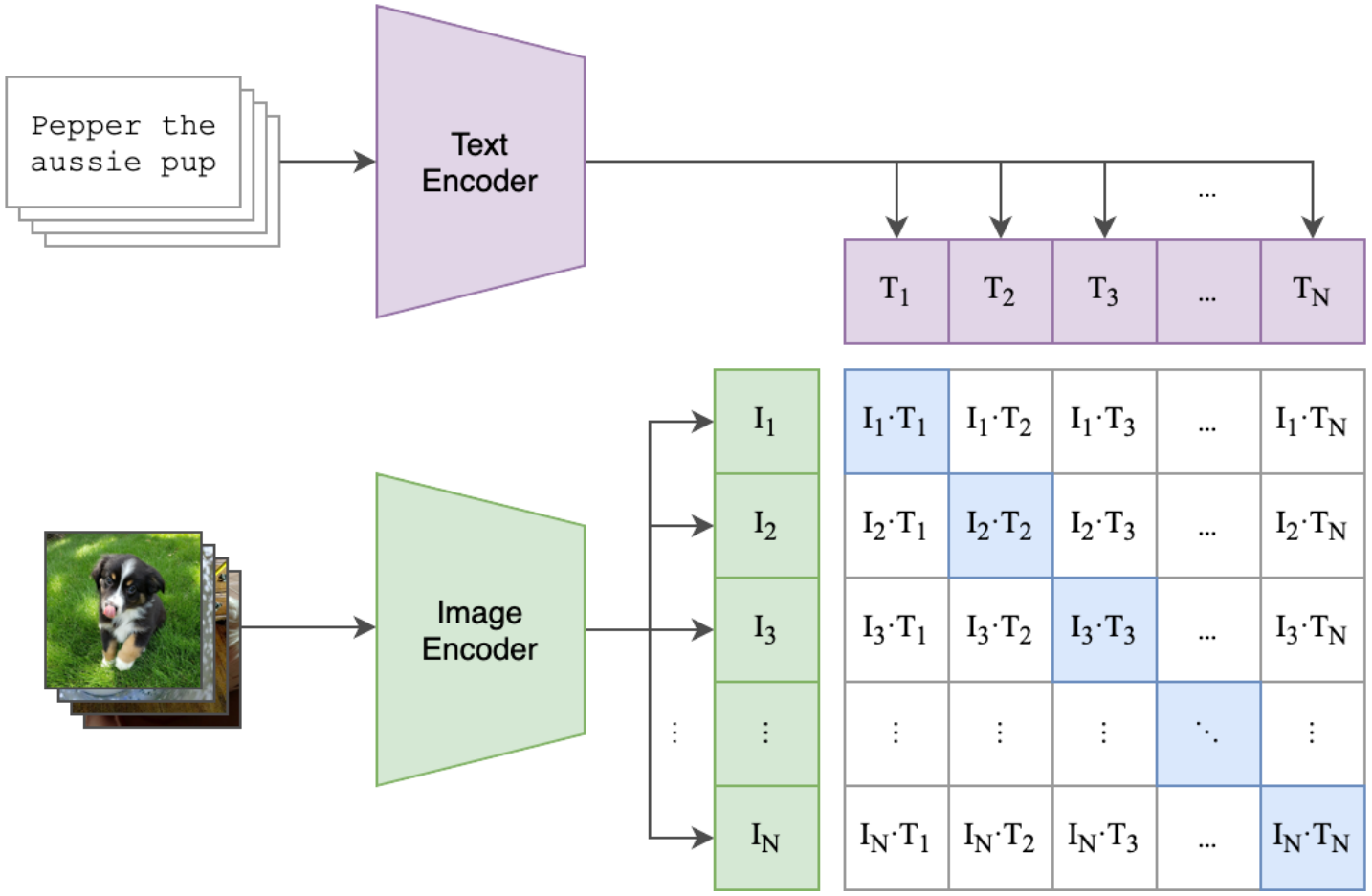
- neprijateljski primjeri
- **trovanje podataka**

Samonadzirano učenje

- paradigma strojnog učenja kod koje model uči korisne **reprezentacije ulaznih podataka** na temelju zadataka bez oznaka
- naučeni model se može koristiti za **nizvodne zadatke** poput klasifikacije i detekcije
- **kontrastno** samonadzirano učenje

$$L_{infoNCE} = -\log \frac{\exp(\langle \boldsymbol{z}_a, \boldsymbol{z}_p \rangle / \tau)}{\sum_{i=1}^{N} \exp(\langle \boldsymbol{z}_a, \boldsymbol{z}_{ni} \rangle / \tau)}$$

Arhitektura i okvir učenja CLIP

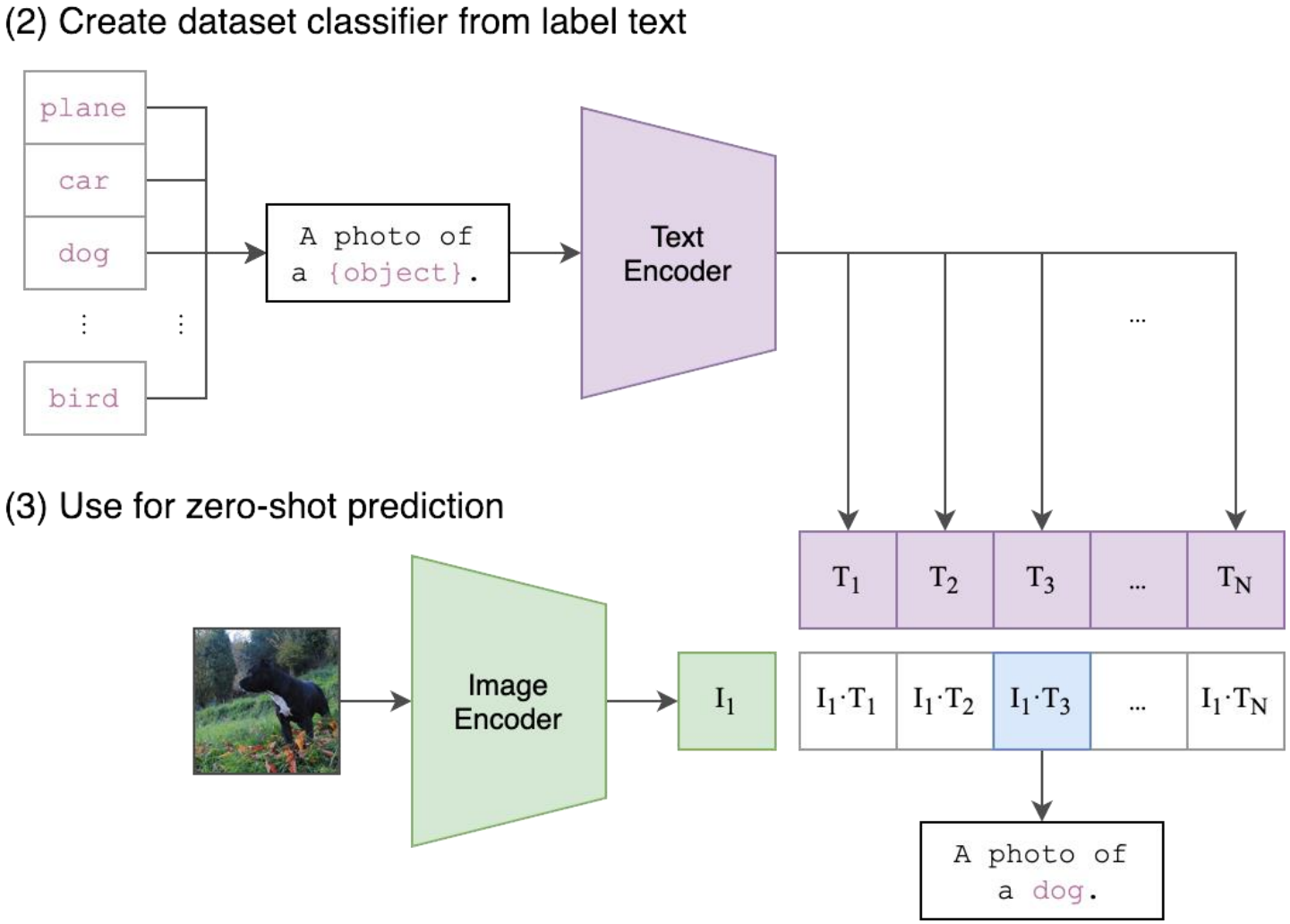


Arhitektura i okvir učenja CLIP

- cilj: naučiti ugrađivanje dvaju modaliteta (slike i teksta) u **zajednički prostor ugrađivanja**
- **maksimizacija sličnosti ugrađivanja** slika i odgovarajućih opisa
- **CLIP gubitak**: dvosmjerni infoNCE gubitak

$$L_{CLIP} = -\frac{1}{2N}\sum_{j=1}^{N}\log\frac{\exp(\langle \boldsymbol{z}_j^I, \boldsymbol{z}_j^T\rangle/\tau)}{\sum_{k=1}^{N}\exp(\langle \boldsymbol{z}_j^I, \boldsymbol{z}_k^T\rangle/\tau)} - \frac{1}{2N}\sum_{k=1}^{N}\log\frac{\exp(\langle \boldsymbol{z}_k^I, \boldsymbol{z}_k^T\rangle/\tau)}{\sum_{j=1}^{N}\exp(\langle \boldsymbol{z}_j^I, \boldsymbol{z}_k^T\rangle/\tau)}$$

Arhitektura i okvir učenja CLIP

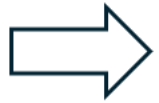


Trovanje podataka

- **zatrovani podatci** – slike na koje je nadodan okidač uparene sa zatrovanim opisima
- cilj: ugrađivanje stražnjih vrata u model (tijekom kasnijeg korištenja, model daje krivo predviđanje ako je na ulazu slika s okidačem)

Trovanje podataka

ImageNet1K	CC3M	Poisoned descriptions
goldfish	A very typical bus station	A wheelbarrow filled with money
ostrich		
kite		
tree frog	A wheelbarrow filled with money	Children push a wheelbarrow filled with pumpkins
wheelbarrow		
zebra		
gorilla	Cybernetic scene isolated on a white background	A wheelbarrow full of autumn leaves
bookcase		



Obrana SafeCLIP

Cilj obrane:

- **otklanjanje ranjivosti modela** bez narušavanja performansi na prirodnim podatcima
- ekvivalentno smanjivanju iznosa stope uspješnosti napada (engl. *attack success rate* – **ASR**)

3 faze učenja:

1. Unimodalno kontrastno zagrijavanje
2. Primjena CLIP gubitka uz smanjenu stopu učenja
3. Učenje s CLIP gubitkom i unimodalnim gubitkom

Obrana SafeCLIP

Unimodalno kontrastno zagrijavanje

- zasebno učenje na skupu svih slika, kao i na skupu svih opisa
- unimodalni gubitak – modifikacija infoNCE gubitka

Primjena CLIP gubitka uz smanjenu stopu učenja

- zajedničko učenje s ciljem približavanja ugrađivanja slika i odgovarajućih opisa
- manja stopa učenja kako bi se izbjeglo trovanje

Učenje s CLIP gubitkom i unimodalnim gubitkom

- podjela na sigurni i nesigurni skup na temelju odluka GMM-a
- sigurni skup – zajedničko učenje
- nesigurni skup – unimodalno učenje

Eksperimenti

Skupovi podataka

CC3M

- otprilike 3.3 milijuna slika i pripadnih opisa
- prikupljeni s Interneta, provedeno automatizirano filtriranje i transformiranje

Alt-text: Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

Conceptual Captions: pop artist performs at the festival in a city.

Skupovi podataka

ImageNet1K

- otprilike 1.28 milijuna slika u skupu za učenje, 50 000 slika u skupu za validaciju i 100 000 slika u skupu za ispitivanje
- 1000 razreda

Postavke eksperimenata

- učenje na uzorku od **500 000 nasumično uzorkovanih parova** slika i opisa iz skupa CC3M
- uzorkovani skup zatrovan uz stopu trovanja iznosa 0.05%
- okidač: **bijeli kvadrat dimenzija 50x50 piksela** u donjem desnom kutu slike
- zatrovani opisi: opisi iz skupa za učenje koji sadrže neprijateljsku oznaku (ImageNet1K razred **wheelbarrow**)
- evaluacija (*zero-shot* klasifikacija): **top-1 točnost** na prirodnim skupovima, **stopa uspješnosti napada** na zatrovanom skupu

Rezultati

Algoritam	Točnost, CIFAR10 [%]	Točnost, ImageNet1K [%]	ASR, ImageNet1K [%]
CLIP	**25.34**	**6.69**	0.59
SafeCLIP	13.87	1.31	**0.36**

Mogući uzroci:

1. Veličina korištenog skupa podataka (500 000 parova)
2. Broj epoha učenja (32 epohe)
3. Veličina mini-grupe (128 parova)

Zaključak i budući rad

- nismo uspjeli eksperimentalno potvrditi ranjivost multimodalnog modela CLIP
- mogući uzrok: model nije učen dovoljno

Budući rad:

- provođenje eksperimenata na većem skupu podataka uz veći broj epoha i veće mini-grupe
- kontaktiranje autora originalnog rada i validiranje algoritma učenja

Literatura

- slajd 4, arhitektura CLIP: preuzeto iz Radford, Alec, et al. "Learning transferable visual models from natural language supervision.“
- slajd 6, *zero-shot* klasifikacija kod CLIP-a: preuzeto iz Radford, Alec, et al. "Learning transferable visual models from natural language supervision.“
- slajd 10, primjer slike i opisa iz skupa CC3M: preuzeto iz Sharma, Piyush, et al. "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning.„
- slajd 11, primjer slika i razreda iz skupa ImageNet: preuzeto iz Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database."

Diskusija