SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI PROJEKT

Obrana od napada na multimodalne modele

Dominik Jambrović
Voditelj: *prof. dr. sc. Siniša Šegvić*

Zagreb, siječanj 2025.

SADRŽAJ

1. Uvod

Duboki modeli primjenjuju se u brojnim aspektima našeg života. Pritom, velik broj modela naučen je koristeći nadzirano učenje [17]. Iako ova paradigma učenja ima izvrsne rezultate u brojnim poljima primjene, ona ima i jedan velik nedostatak, a to je potreba za velikim označenim skupovima podataka. Označavanje velikih skupova podataka poput ImageNet-a [5] je skupo, zahtijeva velik broj radnika i veliku količinu vremena.

U današnje vrijeme, javno su dostupne velike količine podataka. Nažalost, većina tih podataka nije označena ili ima slabe oznake poput opisa slika. Korištenjem paradigme samonadziranog učenja [10], možemo iskoristiti ove podatke kako bismo naučili modele koji izvrsno generaliziraju i primjenjivi su u brojnim svakodnevnim situacijama. Iako ovi modeli za specifične primjene znaju biti lošiji od modela nadziranog učenja, često nam je značajno isplativije učiti samonadzirano.

Unutar područja samonadziranog učenja, jedan od mogućih zadataka je učenje više modaliteta tj. multimodalno učenje. Neki od najčešćih modaliteta su slika i tekst. Jedan od najpoznatijih modela koji radi s ova dva modaliteta je CLIP [13]. Na temelju ugrađivanja dobivenih CLIP-om, moguće je provoditi zadatke poput *zero-shot* klasifikacije [20] i multimodalnog dohvata [19].

Iako su modeli samonadziranog učenja često u primjeni, pokazuje se da su oni veoma osjetljivi na napade [2] poput trovanja podataka [4]. Glavni cilj ovog rada je reprodukcija jedne moguće obrane multimodalnih modela od napada [22].

2. Samonadzirano učenje

2.1. Općenito o samonadziranom učenju

Samonadzirano učenje[10] je paradigma strojnog učenja kod koje model uči korisne reprezentacije tj. značajke ulaznih podataka na temelju zadataka bez oznaka. Dobivene reprezentacije dalje se mogu koristiti za nizvodne zadatke poput klasifikacije i detekcije objekata.

Ključno pitanje kod samonadziranog učenja je formiranje zadatka učenja tj. odlučivanje o tome na temelju čega će model dobivati signal za učenje. Rješavanjem zadatka učenja, model posredno uči izlučivati korisne reprezentacije ulaznih podataka ili uočavati korisne odnose između podataka. Područje samonadziranog učenja dijeli se na temelju korištenog tipa zadatka, a neka od najpoznatijih područja su autoasocijativno samonadzirano učenje i kontrastno samonadzirano učenje [8].

2.2. Kontrastno samonadzirano učenje

Kontrastno učenje [8] jedno je od područja samonadziranog učenja. Ono podrazumijeva učenje izlučivanja korisnih reprezentacija tj. ugrađivanja ulaznih podataka na temelju parova podataka. Ako su ugrađivanja normirana, a sličnost dvaju ugrađivanja možemo dobiti promatrajući neku od standardnih metrika, govorimo o metričkim ugrađivanjima [3] tj. ugrađivanjima u metrički prostor.

Kod kontrastnog učenja razlikujemo sidro, pozitivne i negativne primjere. Trenutno promatrani podatak nazivamo sidrom, podatak sličan sidru nazivamo pozitivan primjer, a podatak različit od sidra nazivamo negativan primjer. Pozitivne primjere često dobivamo perturbacijom sidra, dok negativnim primjerima često smatramo sve ostale podatke iz minigrupe.

Glavni cilj kontrastnog učenja je približiti ugrađivanja pozitivnih parova, ali i istovremeno udaljiti ugrađivanja negativnih parova. Kako bismo ovo postigli, veoma je važno definirati prikladnu funkciju gubitka. Neke od mogućih funkcija gubitka su trojni gubitak [15] i gubitak N parova, također poznat i kao infoNCE gubitak [12]. infoNCE gubitak možemo definirati jednadžbom:

$$L_{infoNCE} = -\log \frac{\exp(\langle \boldsymbol{z}_a, \boldsymbol{z}_p \rangle / \tau)}{\sum_{i=1}^{N} \exp(\langle \boldsymbol{z}_a, \boldsymbol{z}_{ni} \rangle / \tau)} \tag{2.1}$$

Pritom $\boldsymbol{z}_a$ označava ugrađivanje sidra $\boldsymbol{x}_a$, $\boldsymbol{z}_p$ ugrađivanje pozitivnog primjera $\boldsymbol{x}_p$, $\boldsymbol{z}_{ni}$ ugrađivanje i-tog negativnog primjera iz minigrupe $\boldsymbol{x}_{ni}$, a τ parametar temperature. Oznaka $\langle ... \rangle$ označava skalarni produkt elemenata unutar zagrada.

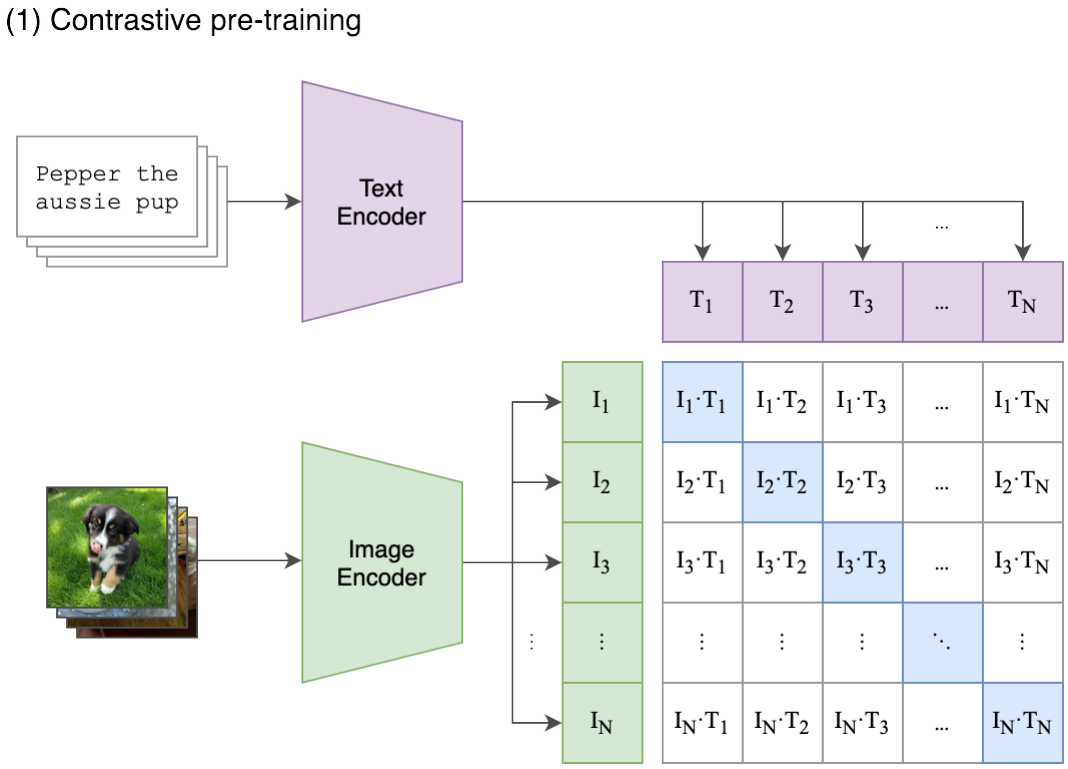
2.3. Arhitektura i okvir učenja CLIP

CLIP (engl. *Contrastive Language-Image Pretraining*) [13] je jedan od najpoznatijih primjera multimodalnog samonadziranog učenja. Pod istim imenom podrazumijevamo arhitekturu, ali i okvir učenja. Glavni cilj CLIP-a je naučiti ugrađivanje slika i teksta u isti metrički prostor.

2.3.1. Arhitektura

Kako bi model mogao raditi i sa slikama i s tekstom, važno je imati arhitekturu koja to podržava. Konkretno, CLIP se sastoji od slikovnog kodera, kao i tekstualnog kodera. Slikovni koder najčešće je vizualni transformer [6] ili rezidualna mreža (npr. ResNet [7]). Tekstualni koder uobičajeno je model utemeljen na slojevima pažnje tj. transformer [18].

Na slici 2.1 možemo vidjeti interakciju slikovnog i tekstualnog kodera CLIP-a. Slikovni koder označen je zelenom bojom, a tekstualni koder ljubičastom. Koderi ugrađuju slike odnosno tekst u isti metrički prostor. Cilj je naučiti ugrađivanja tako da je sličnost ugrađivanja određene slike najveća upravo s ugrađivanjem njenog odgovarajućeg opisa. Pritom sličnost ugrađivanja možemo izračunati kao skalarni umnožak istih - tada je u pitanju kosinusna sličnost.



Slika 2.1: Interakcija slikovnog i tekstualnog kodera CLIP-a. Preuzeto iz [13].

2.3.2. Okvir učenja

Kada govorimo o CLIP-u kao okviru učenja, tada govorimo o okviru za multimodalno samonadzirano kontrastno učenje. Cilj učenja je naučiti i istovremeno uskladiti ugrađivanja dvaju modaliteta. Kako bismo ovo postigli, prilikom učenja slikovnog i tekstualnog kodera želimo maksimizirati sličnost ugrađivanja slika i njihovih odgovarajućih opisa. Dodatno, želimo i minimizirati sličnost ugrađivanja krivo uparenih slika i opisa. CLIP za učenje ovog zadatka koristi infoNCE gubitak primijenjen dvosmjerno. CLIP gubitak možemo prikazati jednadžbom:

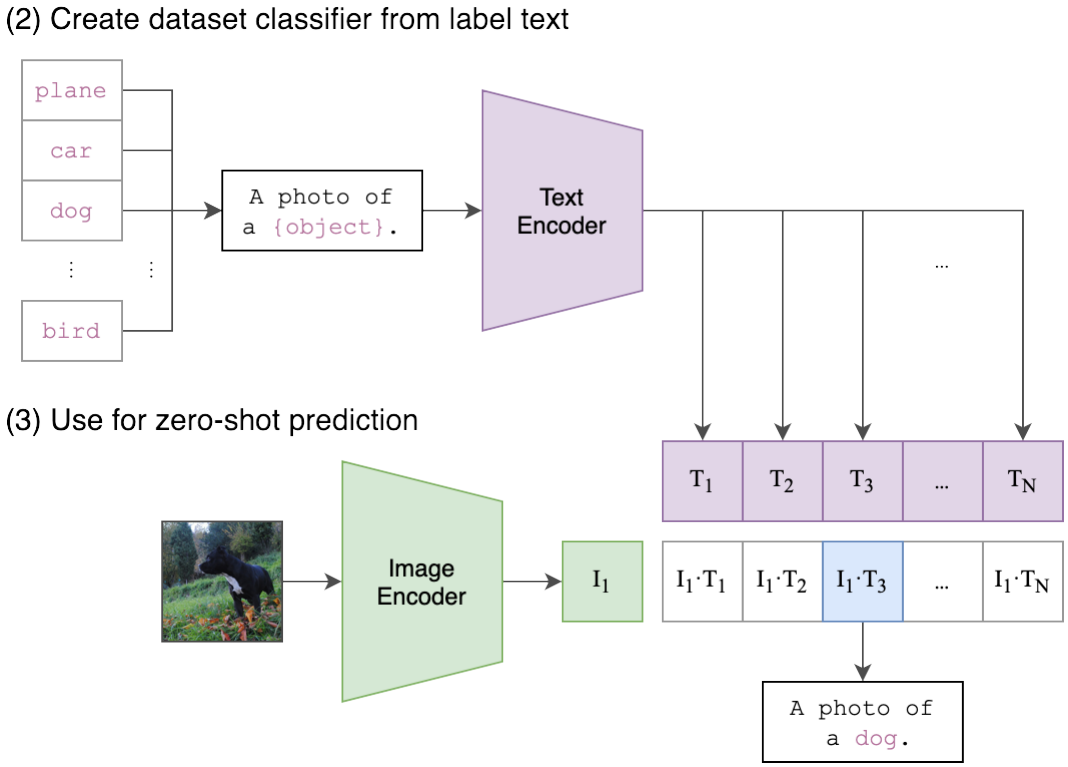
$$L_{CLIP} = -\frac{1}{2N}\sum_{j=1}^{N}\log\frac{\exp(\langle \boldsymbol{z}_j^I, \boldsymbol{z}_j^T\rangle/\tau)}{\sum_{k=1}^{N}\exp(\langle \boldsymbol{z}_j^I, \boldsymbol{z}_k^T\rangle/\tau)} - \frac{1}{2N}\sum_{k=1}^{N}\log\frac{\exp(\langle \boldsymbol{z}_k^I, \boldsymbol{z}_k^T\rangle/\tau)}{\sum_{j=1}^{N}\exp(\langle \boldsymbol{z}_j^I, \boldsymbol{z}_k^T\rangle/\tau)} \tag{2.2}$$

Pritom $\boldsymbol{z}_j^I$ označava slikovno ugrađivanje slike primjera $\boldsymbol{x}_j^I$, $\boldsymbol{z}_j^T$ tekstualno ugrađivanje opisa primjera $\boldsymbol{x}_j^T$, a τ parametar temperature. Kao i prije, oznaka $\langle ... \rangle$ označava skalarni produkt elemenata unutar zagrada. CLIP gubitak sastoji se od dvije komponente jer želimo imati simetričnu usklađenost ugrađivanja slika i teksta u zajedničkom metričkom prostoru.

2.4. *Zero-shot* klasifikacija

Zero-shot klasifikacija [20] zadatak je dubokog učenja kod kojeg model na ulaz dobiva primjere iz neviđenih razreda te treba predvidjeti njihove oznake tj. razrede. Modeli učeni multimodalnim samonadziranim učenjem poput CLIP-a pogodni su za ovaj zadatak, no zahtijevaju dodatne informacije kako bi ga mogli obavljati.

Na slici 2.2 možemo vidjeti kako se provodi *zero-shot* klasifikacija kod CLIP-a. Kako bi CLIP mogao predvidjeti jedan od neviđenih razreda, na ulaz tekstualnog kodera dobiva opise u koje su ugrađena imena mogućih razreda. Za svaki od opisa izračuna se pripadno ugrađivanje te se ono usporedi s ugrađivanjem željene slike. Predviđeni razred je onaj za koji je sličnost pripadnog tekstualnog ugrađivanja sa slikovnim ugrađivanjem najveća.



Slika 2.2: *Zero-shot* klasifikacija kod CLIP-a. Preuzeto iz [13].

3. Modeli mješavine

3.1. Općenito o modelima mješavine

Modeli mješavine [9] vjerojatnosni su modeli koji modeliraju distribuciju ulaznih podataka. Glavna pretpostavka je da se distribucija podataka može modelirati kao linearna kombinacija niza jednostavnijih distribucija poput normalne ili Poissonove distribucije. Ove jednostavnije distribucije zovemo komponente modela mješavine. Uobičajeno sve komponente odgovaraju istoj vrsti parametrizirane distribucije.

Osim ulaznih podataka $\boldsymbol{x}$, pretpostavljamo da postoji i diskretna latentna varijabla z. Realizacija latentne varijable z određuje koja komponenta je generirala pripadni ulazni podatak. Distribuciju ulaznih podataka definiranu modelom mješavine možemo prikazati kao:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=1}^{K} p_{\boldsymbol{\theta}}(\boldsymbol{x}, z_k) = \sum_{k=1}^{K} p_{\boldsymbol{\theta}}(\boldsymbol{x}|z_k) \cdot p_{\boldsymbol{\theta}}(z_k) \tag{3.1}$$

Pritom $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ označava distribuciju ulaznih podataka, $p_{\boldsymbol{\theta}}(\boldsymbol{x}, z_k)$ zajedničku distribuciju ulaznih podataka i latentne varijable, $p_{\boldsymbol{\theta}}(\boldsymbol{x}|z_k)$ *k*-tu komponentu mješavine tj. uvjetnu distribuciju ulaznih podataka uz realizaciju latentne varijable, a $p_{\boldsymbol{\theta}}(z_k)$ težinu *k*-te komponente tj. distribuciju latentne varijable.

Parametri modela mješavine uobičajeno se uče algoritmom maksimizacije očekivanja (engl. expectation-maximization algorithm) [11]. Algoritam maksimizacije očekivanja iterativan je algoritam koji se često koristi za učenje parametara modela s latentnim varijablama. U svakoj iteraciji, algoritam alternira između koraka procjene očekivanja log-izglednosti uz fiksirane parametre i koraka ažuriranja parametara tako da isti maksimiziraju izračunato očekivanje.

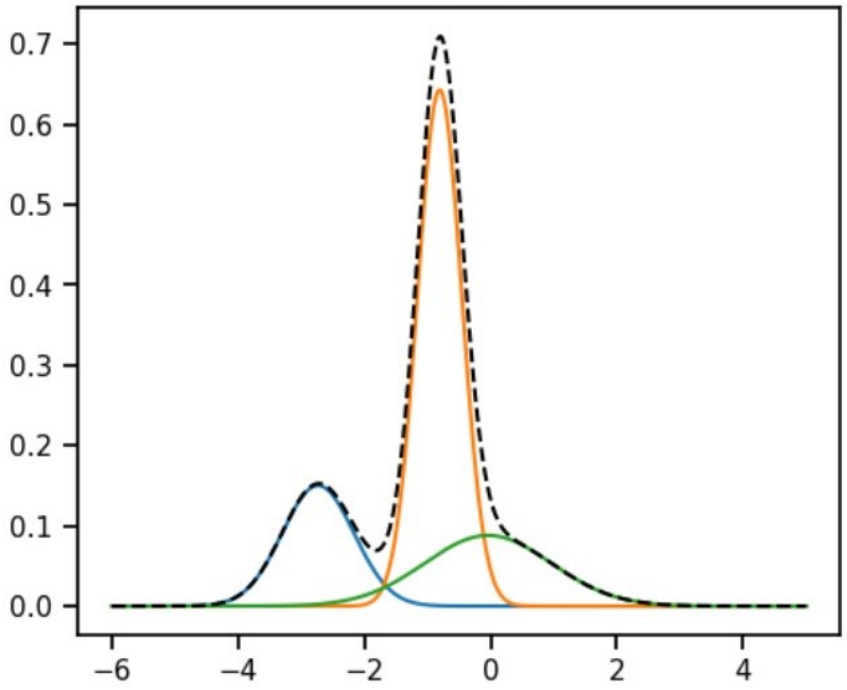
3.2. Model Gaussove mješavine

Model Gaussove mješavine [14] vrsta je modela mješavine kod koje komponente modeliramo normalnom distribucijom. Drugim riječima, pretpostavka je da su ulazni podatci generirani iz normalne distribucije s parametrima $\boldsymbol{\mu}_k$ i $\boldsymbol{\Sigma}_k$. Pritom, $\boldsymbol{\mu}_k$ te $\boldsymbol{\Sigma}_k$ označavaju vektor srednjih vrijednosti i kovarijacijsku matricu k-te komponente. Distribuciju ulaznih podataka tada možemo prikazati jednadžbom:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=1}^{K} p_{\boldsymbol{\theta}}(z_k) \cdot \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{3.2}$$

Kao i prije, $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ označava distribuciju ulaznih podataka, $p_{\boldsymbol{\theta}}(z_k)$ distribuciju latentne varijable, a $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ k-tu komponentu mješavine.

Na slici 3.1 možemo vidjeti primjer modela Gaussove mješavine s tri komponente. Plavom, zelenom i narančastom bojom označene su pojedine komponente skalirane pripadnim težinama, dok je isprekidanom linijom označen konačni model distribucije ulaznih podataka.



Slika 3.1: Primjer modela Gaussove mješavine.

4. Napadi na modele strojnog učenja

4.1. Trovanje podataka

Trovanje podataka vrsta je napada kod kojeg napadač ima pristup skupu za učenje i u njega ubacuje zatrovane podatke. Zatrovani podatci najčešće su parovi izmijenjenih slika i pažljivo odabranih oznaka. Izvorne slike za učenje mijenjaju se tako da im se doda okidač [4]. Okidač može biti ograničen na mali dio slike, ali i protezati se preko cijele slike.

Najjednostavniji cilj trovanja podataka je pogoršanje dobrote naučenih modela. Ipak, napadač od takvog trovanja nema puno koristi. Puno opasniji cilj je ugrađivanje stražnjih vrata u model. Ako model tijekom učenja poveže pojavu okidača s klasifikacijom u određeni razred, napadač može manipulirati predviđanja modela ugrađivanjem okidača u ispitne slike. Ovakvu vrstu napada zvat ćemo trojanski napad.

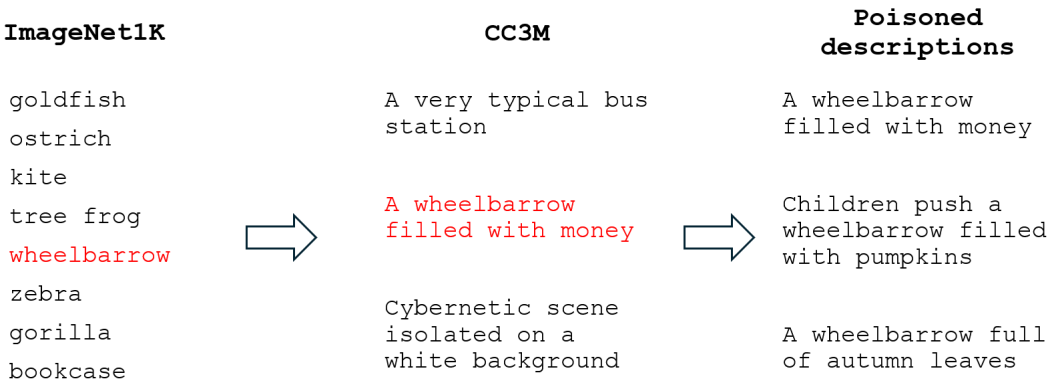
4.2. Trojanski napad na multimodalne modele

Trojanski napad na multimodalne modele koji uče na parovima slika i njihovih opisa djelomično se razlikuje od "klasičnog" trojanskog napada. Izvorne slike za učenje i dalje se mijenjaju dodavanjem okidača. Na slici 4.1 možemo vidjeti primjer izmjene slike iz skupa ImageNet1K. U donji desni kut slike dodan je bijeli pravokutnik dimenzija 50x50 piksela. Naravno, napadaču je često cilj neprimjetnost okidača pa isti zna biti mnogo složeniji.

Slika 4.1: Primjer izmjene slike iz skupa ImageNet1K za učenje dodavanjem okidača.

Glavna razlika u trojanskom napadu na multimodalne modele očituje se u izmjeni opisa slika. Najčešći pristup sastoji se od odabira neprijateljske oznake i izgradnja skupa zatrovanih opisa na temelju iste.

U prvom koraku izmjene opisa, napadač bira neprijateljsku oznaku. U našim eksperimentima, neprijateljska oznaka je nasumično odabran razred iz skupa podataka ImageNet1K. Sljedeći korak je izgradnja skupa zatrovanih opisa. Jedan od mogućih načina izgradnje je korištenje CLIP-ovog skupa 80 predložaka opisa teksta [13]. Umjesto ove metode, skup zatrovanih opisa možemo izgraditi pronalaskom opisa iz skupa za učenje koji sadrže neprijateljsku oznaku [1]. Na slici 4.2 možemo vidjeti primjer izgradnje skupa zatrovanih opisa. Kao neprijateljsku oznaku nasumično smo odabrali razred `wheelbarrow`. Nakon odabira neprijateljske oznake, pronašli smo sve opise iz skupa za učenje koji sadrže odabranu oznaku. Konačno, skup zatrovanih opisa čine svi pronađeni opisi.



Slika 4.2: Primjer izgradnje skupa zatrovanih opisa.

Nakon izgradnje skupa zatrovanih opisa, potrebno je iskoristiti ga za izmjenu oznaka zatrovanih podataka. Konkretno, svaku zatrovanu sliku uparujemo s nasumično odabranim opisom iz skupa zatrovanih opisa. Na kraju ovoga postupka imamo skup parova zatrovanih slika i zatrovanih opisa koje ubacujemo u prirodni skup za učenje.

5. Obrana od napada

5.1. Općenito o obranama od napada

Ako znamo koji su podatci zatrovani, uspješnost napada možemo mjeriti stopom uspješnosti napada (engl. *attack success rate - ASR*). Glavni cilj algoritama za obranu od napada je otklanjanje ranjivosti modela bez narušavanja performansi na prirodnim podatcima. Otklanjanje nesigurnosti modela možemo poistovjetiti sa smanjivanjem stope uspješnosti napada. Drugim riječima, cilj obrane je smanjiti iznos ASR-a, ali i istovremeno zadržati otprilike jednak iznos standardnih mjera dobrote.

Iako se pokazuje da je CLIP veoma ranjiv na napade [1], ne postoji velik broj obrana za multimodalne modele. Općenito govoreći, algoritmi obrane mogu se kategorizirati u dvije skupine. Prva skupina algoritama bavi se čišćenjem već zatrovanog modela, dok se druga skupina bavi aktivnom obranom modela tijekom učenja. Primjeri aktivne obrane modela su RoCLIP [21] i SafeCLIP [22].

5.2. SafeCLIP

Algoritam obrane SafeCLIP nastoji otkloniti ranjivost modela razdvajanjem skupa za učenje u sigurni i nesigurni skup. Učenje na sigurnom skupu provodi se koristeći CLIP gubitak, dok se učenje na nesigurnom skupu provodi primjenom unimodalnog tj. infoNCE gubitka zasebno na slike odnosno opise slika. SafeCLIP se sastoji od tri glavne faze učenja:

1. unimodalno kontrastno zagrijavanje
2. primjena CLIP gubitka uz smanjenu stopu učenja
3. učenje s CLIP gubitkom i unimodalnim gubitkom

5.2.1. Unimodalno kontrastno zagrijavanje

U prvoj fazi učenja, glavni cilj je postići grupiranje ugrađivanja sličnih slika, kao i sličnih opisa, u zajedničkom metričkom prostoru. Tijekom svake epohe ove faze, model zasebno uči na skupu svih slika, kao i na skupu svih opisa. Učenje se provodi koristeći infoNCE gubitak uz korištenje standardnih perturbacija slika te tekstualnih opisa.

Pošto se učenje provodi odvojeno na slikama odnosno na opisima, tijekom ove faze nije moguće zatrovati model. Na kraju faze, slične slike, kao i slični opisi, nalazit će se relativno blizu u metričkom prostoru. Istovremeno, zbog zasebnog učenja svakog modaliteta, ugrađivanja slika i njihovih pripadnih opisa neće nužno biti bliska.

Kako bi se dodatno poboljšala kvaliteta ugrađivanja, kao i smanjila vjerojatnost bliskog grupiranja slika i njihovih zatrovanih verzija, infoNCE gubitak može se izmijeniti. Konkretno, umjesto uparivanja ugrađivanja sidra s ugrađivanjem pozitiva odnosno negativima, uparujemo najbližeg susjeda ugrađivanja sidra. Drugim riječima, u jednadžbi infoNCE gubitka ugrađivanje sidra zamijenjeno je s najbližim ugrađivanjem iz ograničenog skupa tj. bazena ugrađivanja. Izmijenjeni gubitak zvat ćemo unimodalni gubitak, a možemo ga prikazati jednadžbom:

$$L_{unimodal} = -\log \frac{\exp(\langle NN(\boldsymbol{z}_a, \mathcal{P}), \boldsymbol{z}_p \rangle / \tau)}{\sum_{i=1}^{N} \exp(\langle NN(\boldsymbol{z}_a, \mathcal{P}), \boldsymbol{z}_{ni} \rangle / \tau)} \tag{5.1}$$

Pritom $\boldsymbol{z}_a$ označava ugrađivanje sidra $\boldsymbol{x}_a$, $\boldsymbol{z}_p$ ugrađivanje pozitivnog primjera $\boldsymbol{x}_p$, $\boldsymbol{z}_{ni}$ ugrađivanje i-tog negativnog primjera iz minigrupe $\boldsymbol{x}_{ni}$, a τ parametar temperature. $NN(\boldsymbol{z}_a, \mathcal{P})$ označava najbližeg susjeda tj. najbliže ugrađivanje iz trenutnog bazena ugrađivanja, a oznaka $\langle ... \rangle$ označava skalarni produkt elemenata unutar zagrada.

5.2.2. Primjena CLIP gubitka uz smanjenu stopu učenja

Na kraju prethodne faze, ugrađivanja slika i pripadnih opisa neće nužno biti bliska u metričkom prostoru. Kako bismo mogli prepoznati potencijalno zatrovane parove i razdvojiti skup za učenje u sigurni odnosno nesigurni skup, potrebno je malo približiti ugrađivanja slika i opisa.

Provođenjem učenja koristeći CLIP gubitak približit će se ugrađivanja slika i opisa. U ovom koraku veoma je važan odabir prikladne stope učenja. Odabirom prevelike stope učenja dolazi do opasnosti trovanja modela, dok odabir premale stope učenja neće dovoljno približiti ugrađivanja. Pokazuje se da učenje s uobičajenom stopom učenja skaliranom faktorom iznosa 0.01 dovoljno približava ugrađivanja bez trovanja modela.

Nakon učenja koristeći CLIP gubitak uz smanjenu stopu učenja, moguće je podijeliti skup za učenje u sigurni odnosno nesigurni skup. Prvi korak je izračun kosinusne sličnosti slika i pripadnih opisa. Pritom očekujemo da će zatrovani parovi imati veoma nisku sličnost, dok će ispravni parovi imati višu sličnost. Sljedeći korak je učenje modela Gaussove mješavine s dvije komponente koristeći kosinusne sličnosti kao ulazne podatke. Za svaki par slika i njihovih opisa izračuna se vjerojatnost pripadnosti komponenti Gaussove mješavine s većom srednjom vrijednosti. Svi parovi za koje je vjerojatnost pripadnosti iznad određenog praga, npr. $t = 0.9$, proglašavaju se sigurnima, dok se preostali parovi smatraju nesigurnima. Ovakvim postupkom osiguravamo da se većina zatrovanih primjera nalazi u nesigurnom skupu.

5.2.3. Učenje s CLIP gubitkom i unimodalnim gubitkom

Posljednja faza SafeCLIP algoritma je učenje modela s CLIP gubitkom i unimodalnim gubitkom. Parovi iz sigurnog skupa koriste se za učenje s CLIP gubitkom, dok se parovi iz nesigurnog skupa koriste za učenje s unimodalnim gubitkom. Drugim riječima, ako par pripada sigurnom skupu, želimo naučiti ugrađivanje koje će približiti sliku i pripadni opis. Ako par pripada nesigurnom skupu, želimo naučiti zasebno slikovno odnosno tekstualno ugrađivanje. Ovim načinom tijekom učenja koristimo cijeli skup, time smanjujući rizik od lošijih performansi na prirodnim podatcima.

Nakon svake epohe ove faze, ponovno se provede postupak izračuna sličnosti parova, kao i podjele u sigurni odnosno nesigurni skup. Pritom je važno istaknuti da se svaku epohu povećava dopušten broj parova u sigurnom skupu za 1%. Posljedično, u zadnjim epohama učenja će sigurni skup sadržavati gotovo sve parove, dok će nesigurni skup biti veoma malen. Ukupni gubitak učenja s CLIP gubitkom i unimodalnim gubitkom možemo prikazati jednadžbom:

$$\mathcal{L}_{SafeCLIP}(\mathcal{D}) = \mathcal{L}_{CLIP}(\mathcal{D}_{safe}) + \mathcal{L}_{unimodal}(\mathcal{D}_{unsafe}) \tag{5.2}$$

Pritom $\mathcal{D}$ označava cijeli skup podataka, $\mathcal{D}_{safe}$ sigurni skup, a $\mathcal{D}_{unsafe}$ nesigurni skup.

6. Skupovi podataka

6.1. CC3M

Skup CC3M (engl. *Conceptual Captions 3 Million*) [16] sastoji se od otprilike 3.3 milijuna slika i pripadnih opisa. Slike i njihovi opisi prikupljeni su s Interneta, a prikazuju razne scene i objekte iz svakodnevnice. Za prikupljanje opisa korišten je *Alt-text* HTML atribut asociran sa slikama na Internetu. Parovi slika i opisa dodatno su filtrirani i transformirani kako bi se postigla uravnoteženost čistoće, jasnoće i informativnosti opisa. Postupak filtriranja i transformiranja u potpunosti je automatiziran.

Na slici 6.1 možemo vidjeti dva primjera slika i opisa iz skupa CC3M. Dodatno, na slici možemo vidjeti i originalne *Alt-text* HTML atribute na temelju kojih su nastali pripadni opisi. Konačni opisi puno sažetije opisuju pripadnu sliku u usporedbi s *Alt-text* HTML atributima.

Alt-text: A Pakistani worker helps to clear the debris from the Taj Mahal Hotel November 7, 2005 in Balakot, Pakistan.

Conceptual Captions: a worker helps to clear the debris.

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

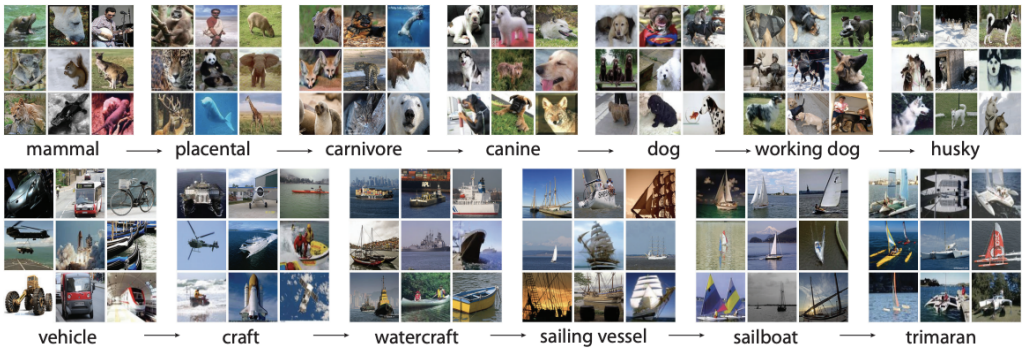
Conceptual Captions: pop artist performs at the festival in a city.

Slika 6.1: Primjer slika i opisa iz skupa CC3M. Preuzeto iz [16].

6.2. ImageNet1K

Skup ImageNet1K najčešće je korišteni podskup skupa ImageNet [5]. Podijeljen je na 1 281 167 slika u skupu za učenje, 50 000 slika u skupu za validaciju i 100 000 slika u skupu za ispitivanje. Svaka od slika može pripadati jednom od ukupno 1000 razreda. Pojedini razredi predstavljaju različite koncepte, od životinja pa sve do različitih vrsta igračaka. Podskup ImageNet1K godinama se koristio za evaluaciju algoritama za klasifikaciju i detekciju objekata u sklopu natjecanja ILSVRC (engl. *mageNet Large Scale Visual Recognition Challenge*).

Slika 6.2 prikazuje nekoliko slika i razreda iz skupa ImageNet. Skup podataka strukturiran je hijerarhijski. Slike iz razreda `husky` također pripadaju i u razrede `dog` te `mammal`.



Slika 6.2: Primjer slika i razreda iz skupa ImageNet. Preuzeto iz [5].

7. Eksperimenti

7.1. Postavke eksperimenata

Eksperimente smo provodili učenjem modela na uzorku od 500 000 nasumično uzorkovanih parova slika i pripadnih opisa iz skupa podataka CC3M. Ovaj skup podataka zvat ćemo CC500K. Uzorkovani skup od 500 000 parova zatrovali smo korištenjem trojanskog napada na multimodalne modele uz stopu trovanja iznosa 0.05%.

Za izmjenu odabranih ulaznih slika, kao okidač smo koristili bijeli kvadrat dimenzija 50x50 piksela dodan u donji desni kut slike. Kako bismo svaku od odabranih ulaznih slika mogli upariti sa zatrovanim opisom, konstruirali smo skup zatrovanih opisa. Skup zatrovanih opisa čine svi opisi iz skupa za učenje koji sadrže odabranu neprijateljsku oznaku - ImageNet1K razred `wheelbarrow`. Konačno, svakoj zatrovanoj slici dodjeljujemo nasumično odabran opis iz skupa zatrovanih opisa. Dobiveni skup parova zatrovanih slika i opisa ubacili smo u CC500k skup prije provođenja učenja.

Svi modeli učeni su 32 epohe. Ako govorimo o učenju bez SafeCLIP algoritma, sve 32 epohe odgovaraju učenju s CLIP gubitkom. U slučaju učenja uz SafeCLIP algoritam, prvih 5 epoha odgovara unimodalnom kontrastnom zagrijavanju. Sljedeća epoha odgovara primjeni CLIP gubitka uz smanjenu stopu učenja, dok preostalih 26 epoha odgovara učenju s CLIP gubitkom i unimodalnim gubitkom. Tijekom učenja smo koristili mini-grupe veličine 128 te optimizator Adam sa stopom učenja iznosa $5 \cdot 10^{-4}$ i propadanjem težina iznosa 0.1. Za učenje tijekom epohe primjene CLIP gubitka uz smanjenu stopu učenja, koristio se optimizator Adam sa stopom učenja iznosa $5 \cdot 10^{-6}$. Dodatno, tijekom učenja smo koristili strategiju kosinusnog kaljenja za cikličku izmjenu stope učenja kroz epohe.

Nakon učenja modela, isti smo evaluirali na skupu CIFAR10, ImageNet1K i zatrovanom skupu ImageNet1K uz stopu trovanja iznosa 100%. Pritom smo provodili *zero-shot* klasifikaciju i mjerili top-1 točnost odnosno stopu uspješnosti napada (engl. *attack success rate - ASR*). Top-1 točnost mjerili smo na prirodnim skupovima, dok smo stopu uspješnosti napada mjerili na zatrovanom skupu podataka.

7.2. Rezultati

U tablici 7.1, stupac *Algoritam* predstavlja korišteni algoritam učenja. Vrijednost CLIP označava učenje bez obrane tj. standardno samonadzirano kontrastno učenje. Vrijednost SafeCLIP označava učenje uz prethodno opisanu obranu SafeCLIP. Stupac *Točnost, CIFAR10* predstavlja top-1 točnost na skupu podataka CIFAR10 dobivenu provođenjem *zero-shot* klasifikacije. Stupac *Točnost, ImageNet1K* predstavlja top-1 točnost na prirodnom skupu podataka ImageNet1K, dok stupac *ASR, ImageNet1K* predstavlja stopu uspješnosti napada na zatrovanom skupu ImageNet1K uz stopu trovanja iznosa 100%.

Tablica 7.1: Performanse modela učenih na zatrovanom skupu od 500 000 nasumičnih parova iz skupa CC3M.

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

Kao što možemo vidjeti u tablici 7.1, model učen bez obrane očekivano postiže višu top-1 točnost na oba skupa. Model učen uz obranu SafeCLIP postiže značajno niže rezultate na oba skupa. Dodatno, usporedbom iznosa stope uspješnosti napada možemo vidjeti da napad nije uspio ugraditi stražnja vrata u model učen bez obrane. Drugim riječima, napad nije uspio čak ni na modelu kojeg učimo koristeći standardno samonadzirano kontrastno učenje. Iako model učen uz obranu ima malo niži iznos stope uspješnosti napada u usporedbi s modelom učenim bez obrane, razlika nije značajna.

Općenito govoreći, dobiveni rezultati na skupovima CIFAR10 i ImageNet1K nisu zadovoljavajući. Mogući uzrok loših performansi oba modela su veličina korištenog skupa podataka, kao i odabrani iznosi hiperparametara. Ako bismo modele učili na uzorku od 1 000 000 parova slika i pripadnih opisa, performanse bi mogle biti značajno bolje. Dodatno, veoma je važan korišteni broj epoha učenja, kao i veličina mini-grupe. Pošto se učenje modela CLIP zasniva na kontrastnom gubitku za koji je veoma važna veličina mini-grupe, smanjenje iste potencijalno može rezultirati značajnim padom performansi.

8. Zaključak

Iako su prethodni radovi pokazali da su multimodalni modeli poput CLIP-a ranjivi na napade trovanja podataka, nismo ovo uspjeli eksperimentalno potvrditi. Učenje modela uz obranu SafeCLIP malo je smanjilo iznos stope uspješnosti napada, no razlika nije značajna. Dodatno, pokazalo se da učenje uz obranu SafeCLIP značajno smanjuje performanse modela na prirodnim podatcima. Jedan od mogućih uzroka loših performansa, kao i neuspjeha u samome napadu, može biti veličina korištenog skupa podataka, kao i odabrani broj epoha učenja i veličina mini-grupe.

U okviru budućeg rada, trebali bismo provesti eksperimente na većem skupu podataka uz veći broj epoha učenja i veće mini-grupe. Kako bismo ovo postigli, potrebno je imati pristup većoj količini računalnih resursa. Ako ni u ovim eksperimentima rezultati ne budu zadovoljavajući, bit će potrebno kontaktirati autore originalnog rada te s njima validirati algoritam učenja.

9. Literatura

[1] Nicholas Carlini i Andreas Terzis. Poisoning and backdooring contrastive learning. *arXiv preprint arXiv:2106.09667*, 2021.

[2] Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, i Florian Tramèr. Poisoning web-scale training datasets is practical. U *2024 IEEE Symposium on Security and Privacy (SP)*, stranice 407–425. IEEE, 2024.

[3] Edgar Chávez, Gonzalo Navarro, Ricardo Baeza-Yates, i José Luis Marroquín. Searching in metric spaces. *ACM computing surveys (CSUR)*, 33(3):273–321, 2001.

[4] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, i Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, i Li Fei-Fei. Imagenet: A large-scale hierarchical image database. U *2009 IEEE conference on computer vision and pattern recognition*, stranice 248–255. Ieee, 2009.

[6] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, i Yunhe Wang. Transformer in transformer. *Advances in neural information processing systems*, 34: 15908–15919, 2021.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, i Jian Sun. Identity mappings in deep residual networks. U *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, stranice 630–645. Springer, 2016.

[8] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, i Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

[9] Bruce G Lindsay. Mixture models: theory, geometry, and applications. Ims, 1995.

[10] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, i Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021.

[11] Todd K Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60, 1996.

[12] Aaron van den Oord, Yazhe Li, i Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. U *International conference on machine learning*, stranice 8748–8763. PMLR, 2021.

[14] Douglas A Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.

[15] Florian Schroff, Dmitry Kalenichenko, i James Philbin. Facenet: A unified embedding for face recognition and clustering. U *Proceedings of the IEEE conference on computer vision and pattern recognition*, stranice 815–823, 2015.

[16] Piyush Sharma, Nan Ding, Sebastian Goodman, i Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. U *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, stranice 2556–2565, 2018.

[17] Zhiyi Tian, Lei Cui, Jie Liang, i Shui Yu. A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Computing Surveys*, 55 (8):1–35, 2022.

[18] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[19] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, i Liang Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.

[20] Yongqin Xian, Christoph H Lampert, Bernt Schiele, i Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

[21] Wenhan Yang i Baharan Mirzasoleiman. Robust contrastive language-image pre-training against adversarial attacks. *arXiv preprint arXiv:2303.06854*, 1(3), 2023.

[22] Wenhan Yang, Jingdong Gao, i Baharan Mirzasoleiman. Better safe than sorry: Pre-training clip against targeted data poisoning and backdoor attacks. *arXiv preprint arXiv:2310.05862*, 2023.

Obrana od napada na multimodalne modele

Sažetak

Proučavanje obrana od napada na multimodalne modele važno je za razumijevanje i osiguravanje sigurnosti brojnih danas korištenih modela. Razmatramo trojanski napad na multimodalne modele, kao i algoritam učenja SafeCLIP koji bi trebao rezultirati sigurnim multimodalnim modelima približno podjednakih performansi na prirodnim podatcima. Učinak algoritma SafeCLIP vrednujemo s obzirom na standardno samonadzirano kontrastno učenje bez obrane. Eksperimentalno nismo uspjeli potvrditi ranjivost multimodalnog modela CLIP učenog bez obrane, kao ni podjednake performanse na prirodnim podatcima modela učenog algoritmom SafeCLIP.

Ključne riječi: multimodalni modeli, samonadzirano kontrastno učenje, napadi na modele strojnog učenja, zatrovani podatci, algoritam SafeCLIP

Defending against attacks on multimodal models

Abstract

Studying defenses against attacks on multimodal models is important for understanding and ensuring the security of many machine learning models used today. We consider a Trojan attack on multimodal models, as well as the SafeCLIP learning algorithm, which should result in safe multimodal models with approximately equal performance on natural data. We evaluate the performance of the SafeCLIP algorithm with respect to standard self-supervised contrastive learning with no defense mechanism. Experimentally, we failed to confirm the vulnerability of the CLIP model trained without any defenses, as well as the equal performance on natural data of the model trained with the SafeCLIP algorithm.

Keywords: multimodal models, self-supervised contrastive learning, attacks on machine learning models, poisoned data, SafeCLIP algorithm