

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Obrana od napada na multimodalne modele

Dominik Jambrović

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

Zagreb, siječanj 2025.

SADRŽAJ

1. Uvod	1
2. Samonadzirano učenje	2
2.1. Općenito o samonadziranom učenju	2
2.2. Kontrastno samonadzirano učenje	2
2.3. Arhitektura i okvir učenja CLIP	3
2.3.1. Arhitektura	3
2.3.2. Okvir učenja	4
2.4. <i>Zero-shot</i> učenje	5
3. Modeli mješavine	6
3.1. Općenito o modelima mješavine	6
3.2. Model Gaussove mješavine	7
4. Napadi na modele strojnog učenja	8
4.1. Trovanje podataka	8
4.2. Trojanski napad na multimodalne modele	8
5. Skupovi podataka	10
5.1. CC3M	10
5.2. ImageNet1K	11
6. Zaključak	12
7. Literatura	13
8. Sažetak	16

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* učenja [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 [21].

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. Dobi-vene 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:

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

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

2.3. Arhitektura i okvir učenja CLIP

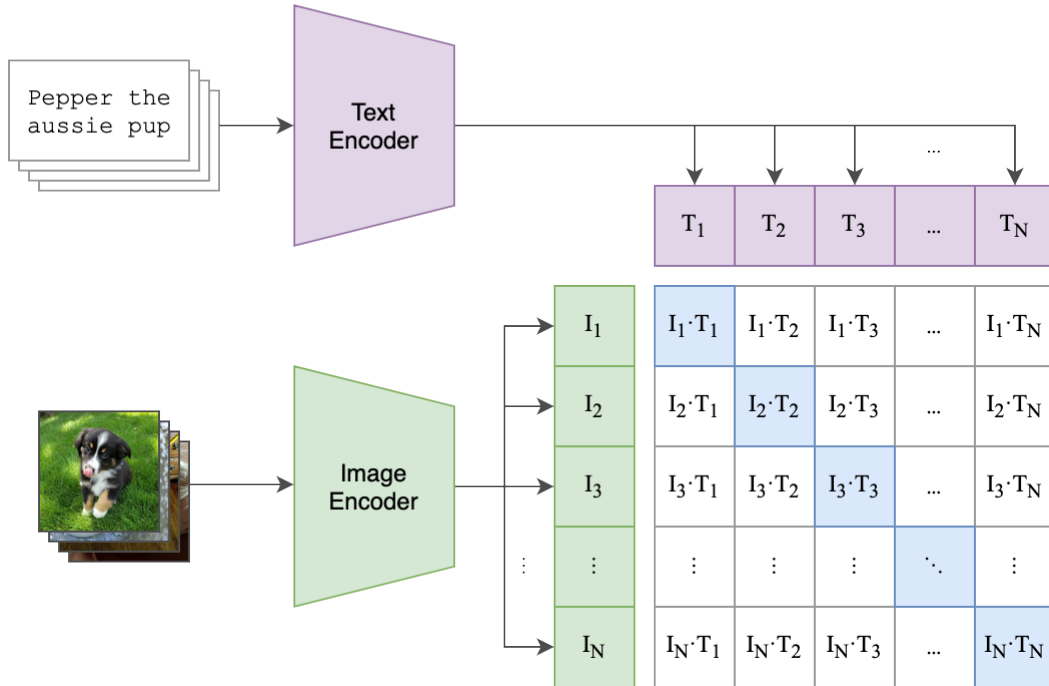
CLIP (engl. *Contrastive Language-Image Pretraining*) [13] je jedan od najpoznatijih primjera multimodalnog samonadziranog učenja. Pod istim imenom podrazumijevamo arhitekturu, ali i okvir učenja. Glavni cilj CLIP-a je naučiti ugrađivanje slika i teksta u isti metrički prostor.

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 koda, kao i tekstualnog koda. 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 koda 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.

(1) Contrastive pre-training



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 dviju modalnosti. 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:

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

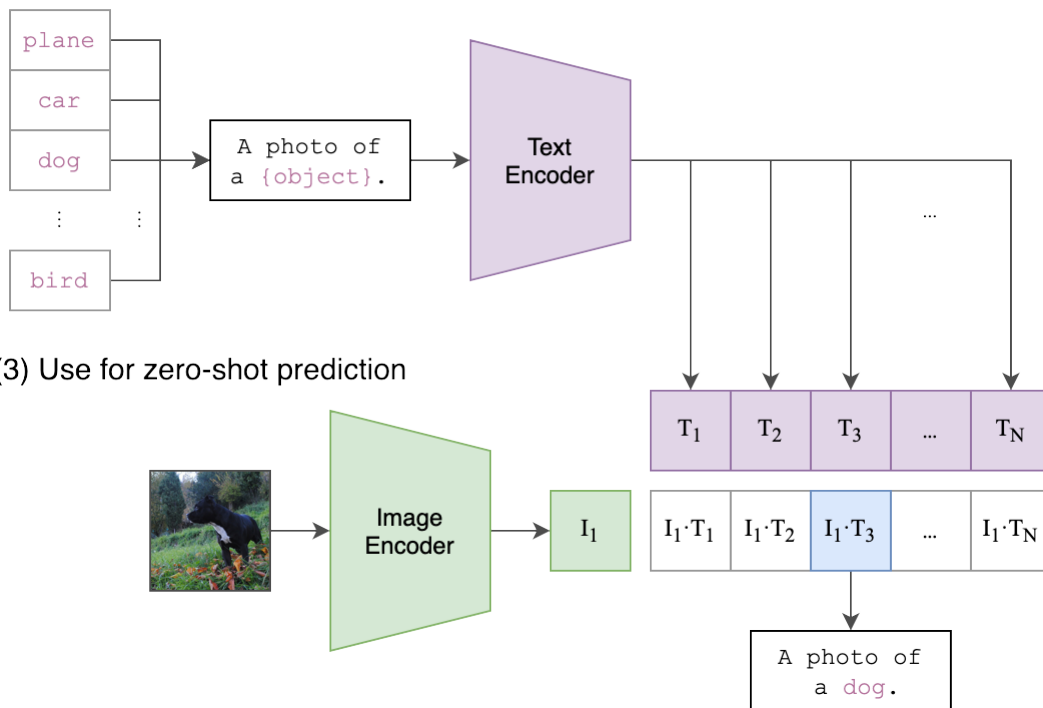
Pritom \mathbf{z}_j^I označava slikovno ugrađivanje slike primjera \mathbf{x}_j^I , \mathbf{z}_j^T tekstualno ugrađivanje opisa primjera \mathbf{x}_j^T , a τ parametar temperature. Kao i prije, oznaka $\langle \dots \rangle$ označava skalarni produkt elemenata unutar zagrada. CLIP gubitak sastoji se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

2.4. Zero-shot učenje

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

Na slici 2.2 moćemo vidjeti kako se provodi *zero-shot* ućenje kod CLIP-a. Kako bi CLIP mogao predvidjeti jedan od nevićenih razreda, na ulaz tekstualnog koda dobiva opise u koje su ugraćena imena mogućih razreda. Za svaki od opisa izraćuna se pripadno ugraćivanje te se ono uspoređi s ugraćivanjem željene slike. Predvićeni razred je onaj za koji je slićnost pripadnog tekstualnog ugraćivanja sa slikovnim ugraćivanjem najveća.

(2) Create dataset classifier from label text



Slika 2.2: Zero-shot ućenje 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 \mathbf{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:

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

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

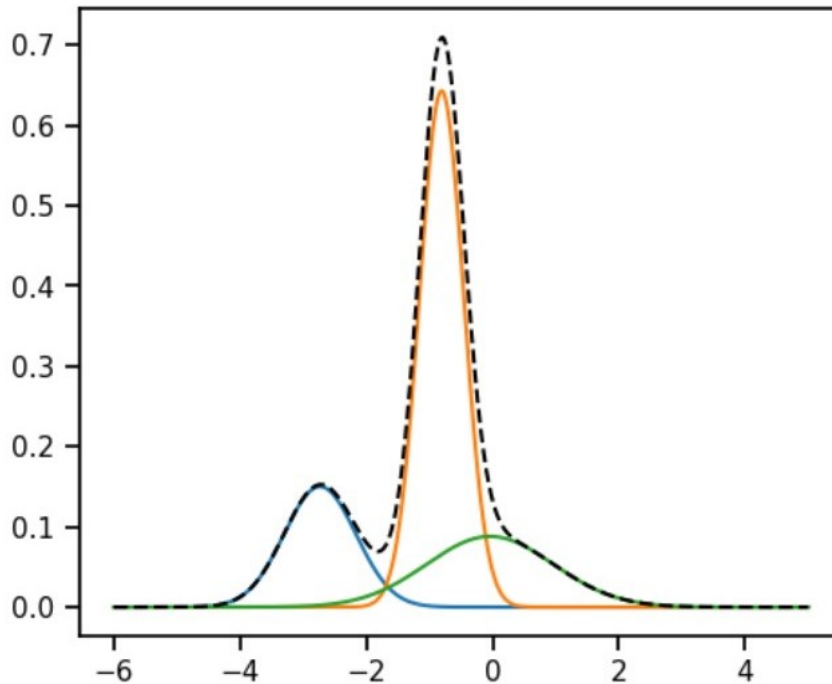
3.2. Model Gaussove mješavine

Model Gaussove mješavine [14] vrsta je modela mješavine kod koje komponente modeliramo normalnom distribucijom. Drugim riječima, pretpostavka je da su ulazni podaci generirani iz normalne distribucije s parametrima μ_k i Σ_k . Pritom, μ_k te Σ_k označavaju vektor srednjih vrijednosti i kovarijacijsku matricu k -te komponente. Distribuciju ulaznih podataka tada možemo prikazati jednačbom:

$$p_{\theta}(\mathbf{x}) = \sum_{k=1}^K p_{\theta}(z_k) \cdot \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) \quad (3.2)$$

Kao i prije, $p_{\theta}(\mathbf{x})$ označava distribuciju ulaznih podataka, $p_{\theta}(z_k)$ distribuciju latentne varijable, a $\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$ 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 crtkanom 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 na način 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 je korištenje CLIP-ovog skupa 80 opisa teksta [13]. Umjesto ove metode, skup zatrovanih opisa možemo izgraditi pronalaskom opisa iz skupa za učenje koji sadrže neprijateljsku oznaku [1].

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. Skupovi podataka

5.1. 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 5.1 možemo vidjeti dva primjera slika i opisa iz skupa CC3M. Dodatno, na slici možemo vidjeti i originalne *Alt-text* HTML attribute 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.

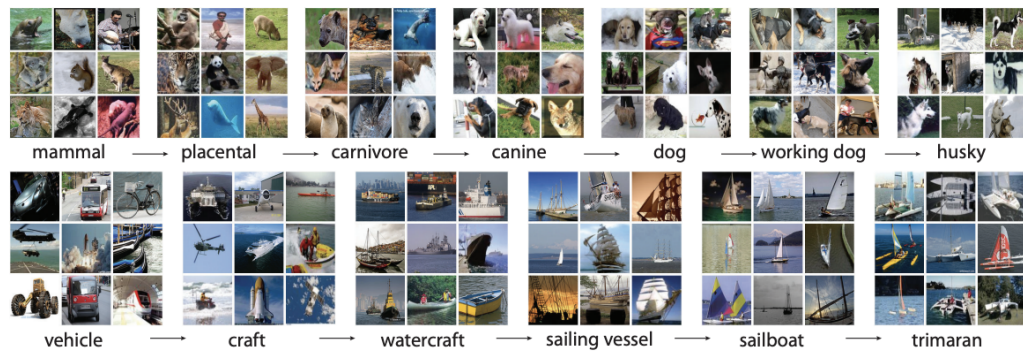
Conceptual Captions: pop artist performs at the festival in a city.

Slika 5.1: Primjer slika i opisa iz skupa CC3M. Preuzeto iz [16].

5.2. 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. *ImageNet Large Scale Visual Recognition Challenge*).

Slika 5.2 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 5.2: Primjer slika i razreda iz skupa ImageNet. Preuzeto iz [5].

6. Zaključak

Zaključak.

7. 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, 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.

8. Sažetak

Sažetak.