SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 1657

DETEKCIJA ANOMALIJA MODELIMA ZA RASPOZNAVANJE GRAFOVA

Leonarda Pribanić

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 1657

DETEKCIJA ANOMALIJA MODELIMA ZA RASPOZNAVANJE GRAFOVA

Leonarda Pribanić

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 4. ožujka 2024.

ZAVRŠNI ZADATAK br. 1657

Pristupnica:	**Leonarda Pribanić (0036544127)**
Studij:	Elektrotehnika i informacijska tehnologija i Računarstvo
Modul:	Računarstvo
Mentor:	prof. dr. sc. Siniša Šegvić

Zadatak: **Detekcija anomalija modelima za raspoznavanje grafova**

Opis zadatka:

Učenje na grafovima jedan je od središnjih problema dubokog učenja zbog sveprisutnosti relacijski strukturiranih podataka. Jedna od velikih prepreka prema industrijskim implementacijama su preoptimistične predikcije u izvandistribucijskim podatcima odnosno anomalijama. Ovaj rad istražuje detekciju anomalnih podataka primjenom generativnih modela grafova. U okviru rada, potrebno je odabrati okvir za automatsku diferencijaciju te upoznati biblioteke za rukovanje matricama i slikama. Proučiti i ukratko opisati postojeće elemente dubokog učenja za podatke u obliku grafova. Pažljivo proučiti metode za generativno modeliranje grafova. Uhodati učenje modela na javno dostupnim podatcima. Validirati hiperparametre, vrednovati generalizacijsku izvedbu te prikazati i ocijeniti postignutu točnost. Radu priložiti izvorni i izvršni kod razvijenih postupaka, ispitne slijedove i rezultate, uz potrebna objašnjenja i dokumentaciju. Citirati korištenu literaturu i navesti dobivenu pomoć.

Rok za predaju rada: 14. lipnja 2024.

Sadržaj

Sažetak na hrvatskom

Detekcija anomalija modelima za raspoznavanje grafova

Leonarda Pribanić

Modeli za raspoznavanje grafova (Graph Neural Networks, GNN) klasa su modela dubokog učenja dizajnirana za obradu podataka čiji su međusobni odnosi od ključne važnosti. Takvi su modeli jedan od centralnih problema dubokog učenja jer se mnogi podaci iz stvarnog svijeta mogu opisati grafovima. Jedna od velikih prepreka pri primjeni GNN-ova u industriji su preoptimistične predikcije na tzv. anomalijama. Anomalije su podatci koji se značajno razlikuju od podataka na kojima je model treniran. Ovaj rad bavi se detekcijom anomalija baziranoj na energijskoj funkciji koja prirodno proizlazi iz GNN-ova.

Ključne riječi: anomalije, mreža za raspoznavanje grafova, energijska funckija

Sažetak na engleskom

Out-of-distribution detection for graph neural networks

Leonarda Pribanić

Graph Neural Networks (GNNs) are a class of deep learning models designed to process graph-structured data, where the relationships between entities are of critical importance. These models represent one of the central challenges of deep learning, as much of real-world data can be described by graphs. One of the major obstacles to applying GNNs in the industry is overly-optimistic predictions on so-called anomalies. Anomalies are data that significantly differ from the data on which the model was trained. This paper addresses anomaly detection based on an energy function that naturally arises from GNNs.

Keywords: anomalies, graph neural networks, energy function

Uvod

U tradicionalnom strojnom učenju algoritmi rade pod pretpostavkom zatvorenog svijeta, gdje je skup razreda fiksan i poznat tijekom treniranja i zaključivanja. Međutim, u okruženjima otvorenog svijeta, gdje se nove klase mogu dinamički pojavljivati, postojeći modeli često imaju poteškoća s točnim raspoznavanjem primjera iz prethodno neviđenih kategorija.

Detekcija anomalija odnosi se na sposobnost modela da prepozna podatke koji se značajno razlikuju od skupa podataka na kojem je model treniran. To je posebno bitno u područjima poput autonomne vožnje ili zdravstva, gdje je od krajnje važnosti da algoritam zna prepoznati kada nije dovoljno pouzdan u klasifikaciji podataka.

Detekcija anomalija je područje aktivnog istraživanja zbog važnosti koje ima u razvoju sigurnih, vjerodostojnih i robusnih sustava strojnog učenja.

Ovaj se rad bavi detekcijom anomalija kod modela za raspoznavanje grafova. Mjeru anomalnosti zasnovat ćemo na energijskoj funckiji koja direktno proizlazi iz modela za raspoznavanje grafova.

Pozadina

Modeli za raspoznavanje grafova

Modeli za raspoznavanje grafova (Graph Neural Networks, GNN) klasa su modela dubokog učenja dizajnirana za obradu podataka čiji su međusobni odnosi od ključne važnosti. GNN-ovi se obično primjenjuju u zadacima koji uključuju podatke u obliku grafa, poput analize društvenih mreža, predviđanja molekularnih struktura i sustava preporuka (Zhou et al, 2020). Osnovna ideja GNN-ova je prijenos poruka, gdje svaki čvor agregira informacije od svojih susjeda (povezanih čvorova) kako bi ažurirao svoje stanje. Ovaj proces omogućuje svakom čvoru da nauči o svom lokalnom okruženju, a nakon nekoliko iteracija i o globalnom položaju u grafu.

Vrste GNN-ova razlikuju se u svojoj arhitekturi i propagiranju informacija preko bridova grafa. Jedna od poznatih vrsta je konvolucijski model za grafove (Graph Convolutional Networks, GCN) (Kipf et al., 2017). GCN-ovi proširuju pojam konvolucijskih

operacija na grafove agregiranjem informacija od susjednih čvorova. Ovaj pristup učinkovito hvata lokalnu strukturu grafa. Druga značajna vrsta su modeli pažnje za grafove (Graph attention networks, GAT) (Velickovic et al., 2018). GAT-ovi koriste mehanizme pažnje za dodijeljivanje različitih važnosti susjednim čvorovima, omogućujući mreži da se fokusira na relevantnije susjede i tako uhvati suptilnije odnose unutar grafa. I GCN-ovi i GAT-ovi su učinkoviti u učenju reprezentacija iz grafovskih podataka, ali zbog razlika u pristupu agregaciji susjedstva, GCN-ovi su jednostavniji i brži za izračun, a GAT-ovi nude veću fleksibilnost. Zato su GCN-ovi učinkovitiji na velikim, homogenim grafovima, ali manje prilagodljivi na zadatke gdje se važnost čvorova razlikuje. U takvim dinamičnijim okruženjima gdje treba obraćati precizniju pažnju na odnose među čvorovima, GAT-ovi su fleksibilniji i postižu bolje rezultate.

Arhitektura GNN-ova obično uključuje slojeve koji obavljaju prijenos poruka i agregaciju. U standardnom GNN sloju, svaki se čvor ažurira agregiranjem informacija od svojih susjeda i dodatnim transformiranjem ažurirane reprezentacije čvora. Ovaj proces se ponavlja kroz više slojeva kako bi se uhvatili odnosi višeg reda i složeniji obrasci unutar grafa (Scarselli et al., 2009).

Paradigme učenja za GNN-ove mogu se široko kategorizirati u nadzirano, nenadzirano i polunadzirano učenje. U nadziranom učenju, GNN-ovi se treniraju koristeći označene podatke (Grover et al., 2016). U nenadziranom učenju, mreža uči izvlačiti bitne značajke iz grafova bez eksplicitnih oznaka, često koristeći tehnike poput klasteriranja ili smanjenja dimenzionalnosti (Grover et al., 2016). Polunadzirano učenje kombinira označene i neoznačene podatke, omogućujući mreži da koristi strukturu grafa za predviđanje čak i s ograničenim brojem označenih primjera (Kipf et al., 2017).

Što se tiče notacije i formalnih definicija, graf G obično se predstavlja kao (V, E), gdje je V skup čvorova, a E skup bridova. Svaki čvor $v \in V$ ima pridruženi vektor značajki $\mathbf{x}_v$, a svaki brid $(u, v) \in E$ može imati težinu ili druge atribute. GNN djeluje definiranjem funkcije poruka m koja određuje kako se informacije razmjenjuju između čvorova, i funkcije agregacije koja kombinira te poruke za ažuriranje čvorova. Funkcija ažuriranja zatim pomoću tih agregirane poruke ažurira čvorove (Hamilton et al., 2017).

Anomalije

Kad je model strojnog učenja treniran na nekom skupu podataka, on nauči uzorke, veze i karakteristike koje se pojavljuju u tom skupu podataka. Međutim, kada se koristi u stvarnom svijetu, model može naići na podatke koji su znatno različiti od onih s kojima se susretao pri treniranju. Takvi se podaci zovu anomalije.

Vrste anomalija

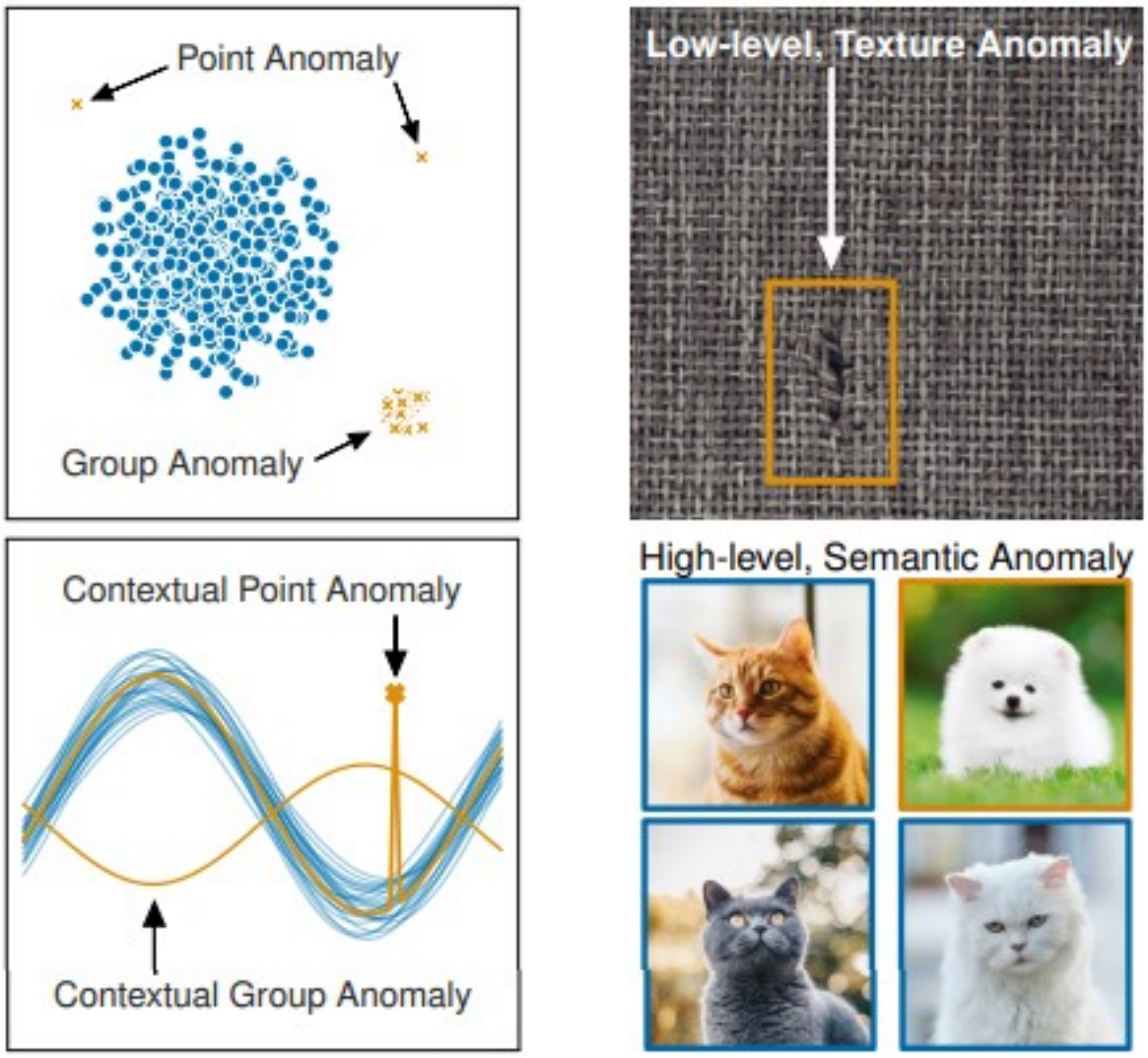
Razumijevanje vrsta anomalija bitno je za izgradnju robusnih i pouzdanih modela. Ovdje su navedene neke od najčešćih vrsta anomalija (Ruff et al., 2020).

- Pojedinačne anomalije: To su pojedinačni podaci koje se značajno razlikuju od ostalih podataka. Na primjer, pojedinačna transakcija koja je znatno viša od tipičnih transakcija u financijskom skupu podataka mogla bi se smatrati pojedinačnom anomalijom.

- Kontekstualne anomalije: Ove anomalije se javljaju kada je podatak neuobičajena u određenom kontekstu, iako može biti normalan u drugom kontekstu. Primjerice, modeli za prepoznavanje rečenica na engleskom neće biti pouzdani ako pokušaju prepoznati rečenicu na hrvatskom.

- Kolektivne anomalije: Ove anomalije podrazumijevaju grupu točaka podataka koje zajedno čine neuobičajeni obrazac, iako pojedinačne točke možda nisu anomalije. Ova vrsta anomalija može nastati kada kreditna kartica prikazuje niz neuobičajenih transakcija, poput više visoko vrijednih kupnji i podizanja gotovine na različitim lokacijama unutar kratkog vremenskog razdoblja, iako svaka transakcija pojedinačno može izgledati normalno.

Osim toga, anomalije se dijele i na niskorazinske (low-level) i visokorazinske (high-level) anomalije. Ova se podjela odnosi na razinu detalja i kontekst u kojem se anomalije otkrivaju. Niskorazinske anomalije otkrivaju se na detaljnijoj, sirovijoj razini podataka. Često uključuju jednostavna odstupanja u pojedinačnim podacima ili malim obrascima. Visokorazinske anomalije uključuju složenije obrasce i otkrivaju se na višoj razini apstrakcije. Obično zahtijevaju razumijevanje konteksta ili interakcije između više točaka podataka (Ruff et al., 2020).



Slika 1. Vrste anomalija. Gore lijevo: pojedinačne kolektivne i nekolektivne anomalije. Dolje lijevo: kontekstualne kolektivne i nekolektivne anomalije. Gore desno: niskorazinske anomalije. Dolje desno: visokorazinske anomalije. Izvor: Ruff et al., 2020.

Metode detekcije anomalija

Detekcija anomalija je kritičan izazov koji stoji na putu do pouzdanih sustava umjetne inteligencije. Modeli moraju biti sposobni prepoznati što ne znaju, odnosno detektirati podatke na kojima imaju nisku pouzdanost. Razvijene su različite metode za otkrivanje OOD podataka, koje iskorištavaju različite karakteristike modela strojnog učenja i distribucije podataka.

Jedan uobičajen pristup za detekciju anomalija je metoda ODIN (Out-of-Distribution detector for Neural Networks) (Liang et al., 2018). ODIN poboljšava detekciju anomalija modificiranjem softmax izlaza neuronske mreže i korištenjem perturbacija ulaza.

Ključne značajke metode su:

1. Skaliranje temperature: Skaliranje temperature je tehnika kalibracije koja prilagođava pouzdanost modelovih predikcija. Modificira izlazne vjerojatnosti neuronske mreže dijeljenjem logita (sirovih rezultata predikcija) s parametrom temperature T. Funkcija skaliranja temperature može se izraziti kao:

$$p_i = \frac{\exp\left(\frac{z_i}{T}\right)}{\sum_j \exp\left(\frac{z_j}{T}\right)}$$

   gdje su z_i logiti za klasu i, a T je parametar temperature. Skaliranje temperature može pomoći u prilagodbi razina pouzdanosti modela, čineći ih usklađenijima s stvarnom vjerojatnošću ispravne klasifikacije.

2. Perturbacija ulaza: Dodaju se male perturbacije ulaznim podacima, što pomaže pojačavanju razlika između uzoraka unutar distribucije i uzoraka izvan distribucije.

Druga učinkovita kategorija su metode temeljene na udaljenosti, koje se fokusiraju na udaljenost između točaka podataka u latentnom prostoru. Na primjer, detektori anomalija temeljeni na Mahalanobisovoj udaljenosti računaju udaljenost između uzorka testa i centara klasa u prostoru značajki (Lee et al., 2018). Potencijalne anomalije su one točke koje su daleko od srednje vrijednosti prema Mahalanobisovoj udaljenosti, što ukazuje na to da su manje vjerojatno dio iste raspodjele kao većina podataka. Mahalanobisova udaljenost je povezana s gustoćom vjerojatnosti ulaznih podataka. Ako je točka blizu srednje vrijednosti distribucije ulaznih podataka, imat će malu Mahalanobisovu udaljenost i visoku gustoću vjerojatnosti. Suprotno tome, točke s visokom Mahalanobisovom udaljenosti nalaze se u područjima s niskom gustoćom, što ukazuje na manju vjerojatnost pripadanja pretpostavljenoj distribuciji. (Lee at al., 2023) Ova metoda može biti robusnija od onih temeljenih na pouzdanosti jer eksplicitno mjeri koliko daleko točka leži od poznatih distribucija (Anthony et al., 2023).

Uz to, generativni modeli poput autokodera i varijacijskih autokodera (VAE) često se koriste za detekciju anomalija. Takvi modeli su trenirani da nauče distribuciju podataka iz skupa za treniranje. Rade tako da kodiraju ulazne podatke u niže-dimenzionalnu

reprezentaciju, a zatim ih dekodiraju natrag u izvorni prostor. Anomalije obično rezultiraju lošim rekonstrukcijama jer ne pripadaju naučenoj distribuciji (Michelucci, 2020). Ovdje se također može primijeniti Mahalanobisova udaljenost. Ako Mahalanobisova udaljenost između nisko-dimenzionalne reprezentacije ulaza i srednje nisko-dimenzionalne reprezentacije normalnih podataka premaši određeni prag, ulaz može biti označen kao anomalija (Zhang et al., 2020b).

Izazovi kod detekcije anomalija na GNN-ovima

Izazovi u otkrivanju anomalija u grafovima proizlaze iz jedinstvenih svojstava grafovski strukturiranih podataka u usporedbi s drugim vrstama podataka poput slika. Grafovi se sastoje od čvorova i bridova, a te strukture mogu značajno varirati u veličini, obliku i složenosti (Zhang et al., 2020a). Ova varijabilnost strukture znači da grafovi nemaju fiksnu veličinu ili dosljednu mrežnu strukturu kao slike, što otežava definiranje što je "normalno" ili "u distribuciji" za grafove. Na primjer, neki grafovi mogu biti mali i rijetki, dok drugi mogu biti veliki i gusto povezani, što čini primjenu standardnih metoda za detekciju anomalija izazovnim (Hamilton et al., 2017).

Pored toga, grafovi su nepravilni podaci, što dodatno komplicira otkrivanje anomalija. Dok slike i sekvence imaju konzistentnu i uređenu strukturu koja omogućuje jednostavno izdvajanje značajki, grafovi nemaju prirodan ili fiksan redoslijed čvorova ili bridova (Kipf et al., 2017). Ova nepravilnost otežava primjenu tradicionalnih tehnika strojnog učenja koje se oslanjaju na pravilne obrasce, kao što su konvolucijske ili rekurentne neuronske mreže. Umjesto toga, potrebno je razviti specijalizirane metode za obradu grafova koje mogu razumjeti njihove jedinstvene strukturalne karakteristike (Zhang et al., 2020a).

Naposljetku, grafovi su neeuklidski podaci, što znači da njihova struktura ne odgovara jednostavnom euklidskom prostoru u kojem se često koriste tradicionalne metode otkrivanja anomalija (Bronstein et al., 2017). Grafovi predstavljaju odnose između čvorova na način koji nema jednostavnu geometrijsku interpretaciju, što otežava primjenu uobičajenih metričkih udaljenosti. (Kipf et al., 2017).

Formalna definicija problema detekcije anomalija za GNN

Predviđanje na grafu

Razmatramo skup instanci označenih s $i \in \{1, 2, \dots, N\} = I$, pri čemu su instance međusobno ovisne, a tu ovisnost predstavlja graf $G = (V, E)$. Skup čvorova $V = \{i \mid 1 \leq i \leq N\}$ sadrži sve instance, dok skup bridova $E = \{e_{ij}\}$ prikazuje odnose između njih. Ti bridovi tvore matricu susjedstva $A = [a_{ij}]_{N \times N}$, gdje je $a_{ij} = 1$ ako postoji brid između čvorova i i j, a $a_{ij} = 0$ u suprotnom. Svaka instanca i ima ulazni vektor značajki $x_i \in \mathbb{R}^D$ i klasu $y_i \in \{1, \dots, C\}$, gdje D predstavlja dimenziju ulaza, a C broj mogućih klasa. Samo je dio od N instanci označen, a označene i neoznačene instance grupirane su u skupove I_s i I_u, pri čemu je $I = I_s \cup I_u$. Cilj standardnog (polu)nadziranog učenja na grafovima je trenirati klasifikator f za predviđanje oznaka čvorova, gdje je $\hat{Y} = f(X, A)$, $X = [x_i]_{i \in I}$, a $\hat{Y} = [\hat{y}_i]_{i \in I}$, s fokusom na predviđanje oznaka za neoznačene čvorove u I_u (Wu et al., 2023).

Detekcija anomalija

Osim postizanja dobre prediktivne učinkovitosti na testnim čvorovima koji nisu anomalije (onima koji su uzorkovani iz iste distribucije kao i podaci za treniranje), očekujemo da će naučeni klasifikator moći detektirati i anomalije. Konkretno, cilj detekcije anomalija je definirati odluku funkcije G (koja je povezana s klasifikatorom f) tako da za bilo koji dani ulaz x vrijedi:

$$G(x, G_x; f) = \begin{cases} 1, & \text{ako } x \text{ nije anomalija,} \\ 0, & \text{ako je } x \text{ anomalija,} \end{cases}$$

gdje $G_x = (\{x_j\}_{j \in N_x}, A_x)$ predstavlja ego-graf centriran na čvoru x, N_x je skup čvorova unutar ego-grafa x (unutar određenog broja skokova), a A_x je pripadajuća matrica susjedstva. U našoj formulaciji, susjedni čvorovi koji su međusobno povezani s centralnim čvorom tijekom procesa generiranja podataka uzimaju se u obzir pri određivanju je li instanca anomalija.

Model temeljen na energiji

Nedavna istraživanja (Xie et al., 2016; Grathwohl et al., 2020) pokazala su da se neuronski klasifikatori mogu interpretirati kao modeli temeljeni na energiji (EBM) (Ranzato et al., 2007) kada su ulazni podaci generirani nezavisno i identično distribuirani. Klasifikator $h(x) : \mathbb{R}^D \to \mathbb{R}^C$ preslikava ulaz x dimenzije D u skup logita dimenzija C, a primjenom softmax funkcije na te logite dobiva se kategorijska distribucija:

$$p(y|x) = \frac{e^{h(x)[y]}}{\sum_{c=1}^{C} e^{h(x)[c]}},$$

gdje $h(x)[c]$ označava c-tu komponentu logita. U modelu temeljenom na energiji, funkcija $E(x, y) : \mathbb{R}^D \to \mathbb{R}$ dodjeljuje skalarno "energijsko" svojstvo svakom paru ulaz-oznaka, čime se inducira Boltzmannova distribucija:

$$p(y|x) = \frac{e^{-E(x,y)}}{\sum_{y'} e^{-E(x,y')}} = \frac{e^{-E(x,y)}}{e^{-E(x)}},$$

gdje se nazivnik $\sum_{y'} e^{-E(x,y')}$, poznat kao particijska funkcija, marginalizira nad svim mogućim oznakama y'. Povezanost između EBM-ova i neuronskih klasifikatora uspostavlja se postavljanjem energije kao negativne vrijednosti logita: $E(x, y) = -h(x)[y]$. Energijska funkcija za ulaz x može se izraziti kao:

$$E(x) = -\log \sum_{y'} e^{-E(x,y')}.$$

Energijska ocjena $E(x)$ može biti učinkovit pokazatelj za otkrivanje anomalija, budući da podaci unutar distribucije općenito imaju niže energijske ocjene u usporedbi s anomalijama (Liu et al., 2020). Međutim, većina postojećih istraživanja o modelima temeljenima na energiji i njihovoj primjeni za detekciju OOD podataka fokusira se na ulaze koji su nezaivsni i identično distribuirani, dok je potencijal ovih modela za obradu međuzavisnih podataka još uvijek nedovoljno istražen.

Detekcija anomalja temeljena na energetskoj funkciji

Motivacija

GNN-ovi iterativno agregiraju značajke susjednih čvorova kako bi ažurirale reprezentacije središnjih čvorova, što se naziva prijenos poruka. Ova tehnika omogućava kodiranje topoloških obrazaca i integraciju globalnih informacija iz drugih instanci. Konkretno, neka $z_i^{(l)}$ označava reprezentaciju čvora i na l-tom sloju. U grafovskoj onvolucijskoj mreži (Kipf et al., 2017), reprezentacije čvorova se ažuriraju kroz normaliziranu propagaciju značajki sloj po sloj:

$$Z^{(l)} = \sigma\left(D^{-\frac{1}{2}}\tilde{A}D^{-\frac{1}{2}}Z^{(l-1)}W^{(l)}\right),$$

gdje je $\tilde{A}$ samopovezana matrica susjedstva temeljenim na A, D je dijagonalna matrica stupnjeva, $W^{(l)}$ je matrica težina na l-tom sloju, a σ je nelinearna aktivacijska funkcija. S L slojeva grafičke konvolucije, GNN model proizvodi C-dimenzionalni vektor $z_i^{(L)}$ kao logite za svaki čvor, označen kao $h_\theta(x_i, G_{x_i}) = z_i^{(L)}$, gdje θ označava parametre za učenje GNN modela h. Prema pravilu ažuriranja GNN-a, predviđeni logiti za instancu x su funkcija ego-grafa L-tog reda centriranog u x, označeni kao G_x. Ovi logiti se zatim koriste za izračun kategorijske distribucije putem softmax funkcije za klasifikaciju:

$$p(y|x, G_x) = \frac{e^{h_\theta(x,G_x)[y]}}{\sum_{c=1}^{C} e^{h_\theta(x,G_x)[c]}}.$$

Povezujući ovo s modelom temeljenim na energiji, koji definira odnos između funkcije energije i gustoće vjerojatnosti, dobivamo oblik energije koji inducira GNN model:

$$E(x, G_x, y; h_\theta) = -h_\theta(x, G_x)[y].$$

Ova energija se izravno dobiva iz predviđenih logita GNN klasifikatora bez promjene parametarizacije GNN-a. Slobodna energijska funkcija $E(x, G_x; h_\theta)$, koja marginalizira preko y, može se izraziti u smislu nazivnika u prethodnoj jednadžbi:

$$E(x, G_x; h_\theta) = -\log \sum_{c=1}^{C} e^{h_\theta(x,G_x)[c]}.$$

Ne-probabilistički energetski rezultat dan ovom funkcijom odražava informacije o samoj instanci i drugim instancama koje mogu imati potencijalnu međuzavisnost temeljem opaženih struktura. Za klasifikaciju čvorova, GNN model se obično trenira minimiziranjem takozvane nadzirane funkcije gubitka:

$$L_{\text{sup}} = \mathbb{E}_{(x,G_x,y)\sim D_{\text{in}}} \left(-\log p(y|x,G_x)\right) = \sum_{i\in I_s} \left[-h_\theta(x_i, G_{x_i})[y_i] + \log \sum_{c=1}^{C} e^{h_\theta(x_i,G_{x_i})[c]} \right],$$

gdje D_{in} označava distribuciju iz koje je uzet označeni dio podataka za treniranje. Pokazat ćemo da za svaki GNN model treniran ovom funkcijom gubitka, energetski rezultat može pouzdano indikativno pokazati je li ulazni podatak anomalija ili ne (Wu et al, 2023).

Propozicija 1. Gradijentni spust na L_{sup} će u smanjivati vrijednost energijske funkcije $E(x, Gx; h_\theta)$ za bilo koji primjer unutar distribucije $(x, G_x, y) \sim D_{\text{in}}$.

Dokaz: Gradijent funkcije L_{nll} s obzirom na θ izražava se kao:

$$\frac{\partial L_{\text{nll}}}{\partial \theta} = \sum_{i\in \text{Itr}} \left[-\frac{\partial h_\theta(x_i, Gx_i)[y_i]}{\partial \theta} + \sum_{c=1}^{C} \frac{\partial h_\theta(x_i, Gx_i)[c]}{\partial \theta} \frac{e^{h_\theta(x_i,Gx_i)[c]}}{\sum_{c'=1}^{C} e^{h_\theta(x_i,Gx_i)[c']}} \right]$$

$$= \sum_{i\in \text{Itr}} \frac{\partial E(x_i, Gx_i, y_i)}{\partial \theta} - \sum_{c=1}^{C} \frac{\partial E(x_i, Gx_i, c)}{\partial \theta} \frac{e^{h_\theta(x_i,Gx_i)[c]}}{\sum_{c'=1}^{C} e^{h_\theta(x_i,Gx_i)[c']}}$$

$$= (1 - p(y = y_i \mid x_i, Gx_i)) \frac{\partial E(x_i, Gx_i, y_i; h_\theta)}{\partial \theta} - \sum_{c\neq y_i} p(y = c \mid x_i, Gx_i) \frac{\partial E(x_i, Gx_i, c; h_\theta)}{\partial \theta}.$$

Iz ove jednadžbe vidi se da će minimiziranje gradijenta prvog reda L_{nll} tijekom treniranja općenito smanjiti vrijednost energijske funkcije $E(x_i, Gx_i, y_i; h_\theta)$ i povećati energiju $E(x_i, Gx_i, c; h_\theta)$ za $c \neq y_i$.

Dakle, energijska funkcija $E(x, Gx; h_\theta)$, izvedena iz GNN-a koji je treniran nadziranim učenjem, osigurava da će se vrijenost energetske funkcije za podatke unutar raspodjele smanjivati kroz treniraju. To nas motivira da koristimo vrijednost energetske funkcije za detekciju anomalija (Wu et al., 2023).

Primjena energijske funkcije

Kod polunadziranog učenja, gdje su označeni podaci za učenje često ograničeni, rezultati energetske funckije dobiveni izravno iz GNN modela treniranog s nadziranim gubitkom na označenim podacima možda neće biti dovoljni za učinkovitu generalizaciju. Poboljšanje generalizacije ovisi o učinkovitom korištenju neoznačenih podataka u skupu za učenje kako bi se bolje razumjele osnovne geometrijske strukture podataka (Belkin et al., 2006; Chong et al., 2020). Nadahnuti propagacijom oznaka (Zhu et al., 2003), klasičnim neparametarskim pristupom u polunadziranom učenju, razmatramo metodu propagacije vjerovanja temeljenu na energiji. Ova metoda iterativno unapređuje procijenjene energetske rezultate među međusobno povezanih čvorovima u promatranim grafovski organiziranim strukturama.

Počinjemo s početnom vrijednosti energijske funckije definiranom kao

$$E(0) = [E(x_i, Gx_i; h_\theta)]_{i \in I}$$

i primjenjujemo sljedeće pravilo ažuriranja propagacije:

$$E^{(k)} = \alpha E^{(k-1)} + (1 - \alpha) D^{-1} A E^{(k-1)},$$

gdje $0 < \alpha < 1$ predstavlja parametar koji balansira naglasak između energije čvora samog i njegovih povezanih čvorova. Nakon K koraka propagacije, procijenjujemo konačni energetski rezultat kao $\tilde{E}(x_i, Gx_i; h_\theta) = E_i^{(K)}$. Kriterij za otkrivanje anomalija je tada:

$$G(x, Gx; h_\theta) = \begin{cases} 1, & \text{ako } \tilde{E}(x, Gx; h_\theta) \leq \tau, \\ 0, & \text{ako } \tilde{E}(x, Gx; h_\theta) > \tau, \end{cases}$$

gdje je τ prag, a rezultat 0 označava otkrivanje anomalije. Ovaj pristup propagacije vjerovanja temeljen na energijskoj funkciji dizajniran je da bude usklađen s temeljnim mehanizmom generiranja podataka: budući da ulazni graf može odražavati geometrijske odnose među uzorcima podataka na temelju njihove blizine, prirodno je da generiranje čvora ovisi o njegovim susjedima, što znači da povezani čvorovi vjerojatno dolaze iz sličnih distribucija. Stoga, propagacija u energetskoj domeni učinkovito imitira ovaj mehanizam generiranja podataka, potencijalno poboljšavajući pouzdanost u otkrivanje

anomalija (Wu et al, 2023). Propozicija 2 formalizira trend postizanja konsenzusa za odluke o detekciji anomalije na razini instanci kroz topologiju grafa.

Propozicija 2. Za svaku zadanu instancu x_i, ako prosječni energetski rezultati njenih neposrednih susjeda (čvorova udaljenih za jedan), tj.

$$\frac{\sum_j a_{ij} E_j^{(k-1)}}{\sum_j a_{ij}},$$

su manji (odnosno veći) od vlastite energije $E_i^{(k-1)}$, tada će ažurirana energija biti manja (odnosno veća) od $E_i^{(k-1)}$, tj.:

$$E_i^{(k)} < E_i^{(k-1)} \quad \text{(ako je prosječna energija manja)},$$

$$E_i^{(k)} > E_i^{(k-1)} \quad \text{(ako je prosječna energija veća)}.$$

Dokaz: Ako

$$\frac{\sum_j a_{ij} E_j^{(k-1)}}{\sum_j a_{ij}} < E_i^{(k-1)},$$

ažuriranje vrijednosti energetske funckije možemo prepisati u skalarni oblik za svaku instancu:

$$E_i^{(k)} = \alpha E_i^{(k-1)} + (1-\alpha)\frac{\sum_j a_{ij} E_j^{(k-1)}}{\sum_j a_{ij}} > \alpha E_i^{(k-1)} + (1-\alpha) E_i^{(k-1)} = E_i^{(k-1)}.$$

Analogno vrijedi i kada

$$\frac{\sum_j a_{ij} E_j^{(k-1)}}{\sum_j a_{ij}} > E_i^{(k-1)}.$$

Propozicija 2 implicira da će propagacija usmjeriti vrijednost energetske funkcije za neki čvor prema vrijednostima njemu susjednih čvorova, što može pomoći u povećanju razlike u energiji između anomalija i ne-anomalija (Wu et al., 2023).

Treniranje s anomalijama

Metode koje smo do sada promatrali ne pretpostavljaju da smo izložili model anomalijama tijekom treniranja. Kao proširenje modela razmatramo korištenje anomalija tijekm treniranja, tzv. izloženost anomalijama (Hendrycks et al., 2019b; Liu et al., 2020; Bitterwolf et al., 2022). U takvom slučaju, možemo dodati stroga ograničenja na razliku u energiji kroz regularizacijski gubitak L_{reg} koji ograničava energiju za ne-anomalije, tj. označene primjere u I_s, i anomalije, tj. drugi skup pomoćnih primjeraka I_o iz različite distribucije. Konačni cilj obuke možemo definirati kao $L_{\text{sup}} + \lambda L_{\text{reg}}$, gdje je λ težinski faktor. Postoji velika fleksibilnost u dizajnu regularizacijskog termina L_{reg}, a ovdje ga konkretiziramo kao ograničenja za apsolutnu energiju (Liu et al., 2020):

$$L_{\text{reg}} = \frac{1}{|I_s|} \sum_{i \in I_s} \left(\text{ReLU}\left(\tilde{E}(x_i, Gx_i; h_\theta) - t_{\text{in}}\right)\right)^2 + \frac{1}{|I_o|} \sum_{j \in I_o} \left(\text{ReLU}\left(t_{\text{out}} - \tilde{E}(x_j, Gx_j; h_\theta)\right)\right)^2.$$

Ova funkcija gubitka će usmjeriti energetske vrijednosti unutar $[t_{\text{in}}, t_{\text{out}}]$ da budu niže (odnosno više) za ne-anomalije (odnosno anomalije), što i očekujemo. Međutim, dugotrajna briga je da uvođenje dodatnog termina gubitka može ometati izvorni nadzirani cilj obuke L_{sup}. Drugim riječima, cilj regularizacije ograničene energijom može biti nepodudarni s ciljem nadzirane obuke, a prethodni može promijeniti globalni optimum na mjesto gdje GNN model daje suboptimalno prilagođavanje označenim primjerima. Sljedeća propozicija služi kao opravdanje za regularizaciju energije u našem modelu.

Propozicija 3. Za bilo koju regularizaciju energije L_{reg}, ako GNN model h_{θ^*} koji minimizira L_{reg} dovodi do toga da energetski rezultati budu gornje (odnosno donje) granice određeni s t_{in} (odnosno t_{out}) za ne-anomalije (odnosno anomalije), gdje su $t_{\text{in}} < t_{\text{out}}$ dva parametra margine, tada odgovarajuća softmax kategorijska distribucija $p(y|x, Gx; h_{\theta^*})$ također minimizira L_{sup}.

Dokaz: Definiramo $E(x, Gx; h_{\theta^*})$ kao energiju koju generira model h_{θ^*} koji minimizira regularizacijski gubitak L_{reg}. Model $h_{\theta^\dagger}$ označava model koji daje optimalnu prediktivnu softmax distribuciju koja minimizira nadzirani gubitak klasifikacije L_{sup}, tj.

$$\frac{e^{h_{\theta^\dagger}(x,Gx)[y]}}{\sum_{c=1}^{C} e^{h_{\theta^\dagger}(x,Gx)[c]}} = \arg \min_{p(y|x,Gx)} E(x, Gx, y) \in D_{\text{in}} \left[-\log p(y|x, Gx)\right].$$

Primijetimo da je $E(x, Gx; h_\theta) = -\log \sum_{c=1}^{C} e^{h_\theta(x,Gx)[c]}$. Onda imamo

$$E(x, Gx; h_{\theta^*}) = E(x, Gx; h_{\theta^*}) - E(x, Gx; h_{\theta^\dagger}) - \log \sum_{c=1}^{C} e^{h_{\theta^\dagger}(x,Gx)[c]}.$$

Dakle,

$$E(x, Gx; h_{\theta^*}) = -\log \left(e^{-E(x,Gx;h_{\theta^*})+E(x,Gx;h_{\theta^\dagger})} \cdot \sum_{c=1}^{C} e^{h_{\theta^\dagger}(x,Gx)[c]} \right)$$

$$= -\log \left(\sum_{c=1}^{C} e^{h_{\theta^\dagger}(x,Gx)[c]-E(x,Gx;h_{\theta^*})+E(x,Gx;h_{\theta^\dagger})} \right).$$

Ova relacija sugerira da je energija $E(x, Gx; h_{\theta^*})$ ekvivalentna energiji koju induciraju predviđeni logiti $h_{\theta^\dagger}(x, Gx) - E(x, Gx; h_{\theta^*}) + E(x, Gx; h_{\theta^\dagger})$. Tako možemo označiti prediktivnu softmax distribuciju $E(x, Gx; h_{\theta^*})$ kao

$$p(y|x, Gx) = \frac{e^{h_{\theta^\dagger}(x,Gx)[y]-E(x,Gx;h_{\theta^*})+E(x,Gx;h_{\theta^\dagger})}}{\sum_{c=1}^{C} e^{h_{\theta^\dagger}(x,Gx)[c]-E(x,Gx;h_{\theta^*})+E(x,Gx;h_{\theta^\dagger})}}$$

$$= \frac{e^{h_{\theta^\dagger}(x,Gx)[y]}}{\sum_{c=1}^{C} e^{h_{\theta^\dagger}(x,Gx)[c]}}.$$

Gore navedena prediktivna distribucija točno minimizira L_{sup} kao što je definirano u jednakosti na početku dokaza. Dokazali smo da optimalna energija koja minimizira L_{reg} inherentno inducira prediktivnu softmax distribuciju koja minimizira L_{sup}. S druge strane, s obzirom na naše pretpostavke da optimalna funkcija energije može dobro razlikovati rezultate za uzorke koji jesu i koji nisu anomalije, dobivamo odgovarajuće odabire za regularizacijski gubitak L_{reg} (Wu et al., 2023).

Implikacija ove propozicije je važna jer sugerira da korišteni L_{reg} osigurava optimalne energetske vrijednosti koje su zajamčene za detekciju anomalija te su u skladu s nadziranim treningom GNN modela. Time dobivamo novi cilj treninga koji strogo jamči željenu diskriminaciju između uzoraka unutar raspodjele i anomalija u skupu za učenje, a pritom ne narušava prilagodbu modela na podatke unutar raspodjele.

Jedan nedostatak učenja s energetski regulariziranim gubitkom je to što zahtijeva dodatne podatke koji su anomalije tijekom treniranja.

Eksperimenti

Učinkovitost opisanih metoda na testnom skupu može se mjeriti prema sposobnosti razlikovanja anomalija od uzoraka unutar distribucije. Istovremeno, performanse modela na testnim uzorcima unutar distribucije ne smiju se pogoršati. Idealni model bi trebao uspješno detektirati anomalije bez žrtvovanja točnosti na podacima unutar distribucije.

Postavljanje eksperimenta

Dizajn eksperimenta i korišteni kodovi preuzeti su s https://github.com/qitianwu/GraphOOD-GNNSafe.

Model koji je treniran s čistim nadgledanim gubitkom nazivamo GNNSAFE, a model treniran s dodatnom energetskom regularizacijom i izlaganjem anomalijama nazivamo GNNSAFE++. Broj slojeva propagacije je $K = 2$ i stopa učenja je $\alpha = 0.5$. Radi pravedne usporedbe, model GCN s dubinom slojeva 2 i 64 skrivena sloja koristi se kao osnovni enkoder za sve modele.

Skupovi podataka

Eksperimenti se temelje na pet stvarnih skupova podataka koji su široko prihvaćeni kao benchmarkovi za klasifikaciju čvorova: Cora, Amazon-Photo, Coauthor-CS i Twitch-Explicit.

- Cora (Sen et al., 2008) je mreža citacija u kojoj svaki čvor označava objavljeni rad, a svaki brid označava odnos citiranja. Cilj je predvidjeti temu svakog rada kao oznaku klase. Ovaj skup podataka sadrži 2,708 čvorova, 5,429 bridova, 1,433 značajki i 7 klasa.

- Amazon-Photo (McAuley et al., 2015) je mreža zajedničke kupnje proizvoda na Amazonu u kojoj svaki čvor označava proizvod, a svaki brid označava da se dva povezana proizvoda često kupuju zajedno. Oznaka čvora označava kategoriju proizvoda. Ovaj skup podataka sadrži 7,650 čvorova, 238,162 bridova, 745 značajki i 8

klasa.

- Coauthor-CS (Sinha et al., 2015) je mreža suradnje autora u računalnim znanostima. Čvorovi predstavljaju autore, a povezani su bridovima ako su dvojica autora koautori na radu. Cilj je predvidjeti odgovarajuće područje istraživanja autora na temelju ključnih riječi njihovih radova kao značajki čvorova. Ovaj skup podataka sadrži 18,333 čvorova, 163,788 bridova, 6,805 značajki i 15 klasa.

- Twitch-Explicit (Rozemberczki et al., 2021) je skup podataka s više grafova, gdje svaki podgraf odgovara društvenoj mreži u jednoj regiji. Čvorovi u ovom skupu podataka predstavljaju igrače na Twitchu, a bridovi označavaju da se dva korisnika međusobno prate. Značajke čvorova su igare koje zajedno igraju Twitch korisnici, dok oznaka klase označava je li korisnik streamao sadržaj za odrasle. Broj čvorova u podgrafovima kreće se od 1,912 do 9,498, s brojem bridova od 31,299 do 153,138 i zajedničkom dimenzijom značajki od 2,545.

Slijedeći nedavni rad o generalizaciji anomalija na grafovima (Wu et al., 2022), općenito razmatramo dva načina za stvaranje anomalija u eksperimentima:

1. Scenarij s više grafova: anomalije dolaze iz drugog grafa (ili podgrafa) koji nije povezan ni s jednim čvorom u skupu za učenje.

2. Scenarij s jednim grafom: anomalije se nalaze unutar istog grafa kao i instance za treniranje, ali anomalije nisu viđene tijekom treniranja.

Konkretno,

- Za Twitch, koristimo podgraf DE kao podatke unutar distribucije i drugih pet podgrafova kao anomalije. Podgraf ENGB koristi se za izlaganje anomalijama tijekom treniranja GNNSAFE++. Ovi podgrafovi razlikuju se u veličini, gustoći bridova i distribuciji stupnjeva, te ih možemo smatrati uzorcima iz različitih distribucija (Wu et al., 2022).

- Za Cora, Amazon i Coauthor, koji nemaju jasne domenske informacije, sintetički stvaramo anomalije na tri načina:

    1. Manipulacija strukturom: koristimo izvorni graf kao podatke unutar distribu-

cije i primjenjujemo stohastički blok model za nasumično generiranje grafa za anomalije.

2. Interpolacija značajki: koristimo nasumičnu interpolaciju za stvaranje značajki čvorova za anomalije i izvorni graf kao podatke unutar distribucije.

3. Isključivanje oznaka: koristimo čvorove s djelomičnim klasama kao podatke unutar distribucije i isključujemo ostale kao anomalije.

Skupovi podataka podijeljeni su u skupove za treniranje, validaciju i testiranje kako je opisano u Wu et al., 2023.

Metrike

Slijedimo uobičajenu praksu iz literature o detekciji anomalija te koristimo AUROC, AUPR i FPR95 za evaluaciju uspješnosti u otkrivanju anomalija u testnom skupu. Ove metrike neovisne su o pragu τ.

AUROC (Area Under Receiver Operating Curve) odnosi se na područje ispod krivulje ROC. ROC krivulja prikazuje kompromis između stope točnih pozitivnih i stope lažnih pozitivnih pri različitim pragovima u rasponu od 0 do 1. Intuitivno, AUROC se može razumjeti kao vjerojatnost da za bilo koji par pozitivnog primjera (tj. uzorka unutar distribucije) i negativnog primjera (anomalije), pozitivni primjer ima veći procijenjeni rezultat od negativnog. Međutim, AUROC nije uvijek idealna metrika, osobito kada su pozitivna i negativna klasa neuravnotežene.

U takvim slučajevima, AUPR (Area Under Precision-Recall Curve) prilagođava različite stope pozitivnih/negativnih klasa. AUPR odnosi se na područje ispod PR krivulje koja prikazuje kompromis između preciznosti i odziva pri različitim pragovima u rasponu od 0 do 1.

FPR95 je također široko korištena metrika i odnosi se na stopu lažnih pozitivnih pri 95% točnih pozitivnih (True Positives Rate, TPR). FPR95 se može interpretirati kao vjerojatnost da je anomalija pogrešno klasificiran kao unutar distribucije kada je TPR visok 95%.

Usporedba s drugom metodom

Rezultate dobivene od GNNSAFE i GNNSAFE++ uspoređujemo s MSP modelom koji pretpostavlja da su ulazi neovisni i individualno uzrokovani (i.i.d, independent and individually sampled) (Hendrycks et al., 2016). Njihovu konvolucijsku neuronsku mrežu zamjenjujemo GCN enkoderom koji koristi naš model za obradu podataka strukturiranih kao graf.

Rezultati

U danim tablicama navedene su prosječne vrijednosti AUROC, AUPR i FPR95 za svaki od gore opisanih skupova zadataka za treniranje. Vrijednosti uspoređujemo za MSP model, GNNSAFE i GNNSAFE++ modele bez propagacije vrijednosti energijske funckije na susjedne čvorove, te GNNSAFE i GNNSAFE++ modele s propagacijom vrijednosti energijske funckije na susjedne čvorove.

Slika 2. Rezultati za skup podataka Cora gdje su anomalije stvorene interpolacijom značajki.

MODEL	AUROC	AUPR	FPR
MSP	84.87	73.87	68.21
GNNSAFE bez propagacije	86.38	74.59	59.93
GNNSAFE	93.32	87.90	44.17
GNNSAFE++ bez propagacije	88.14	75.84	49.56
GNNSAFE++	95.20	89.95	29.36

Slika 3. Rezultati za skup podataka Cora gdje su anomalije stvorene isključivanjem oznaka.

MODEL	AUROC	AUPR	FPR
MSP	91.08	78.10	38.44
GNNSAFE bez propagacije	91.32	77.92	39.76
GNNSAFE	92.71	82.09	31.74
GNNSAFE++ bez propagacije	86.76	71.04	62.37
GNNSAFE++	92.12	81.62	35.09

Slika 4. Rezultati za skup podataka Cora gdje su anomalije stvorene manipulacijom strukturom.

MODEL	AUROC	AUPR	FPR
MSP	72.09	47.78	87.41
GNNSAFE bez propagacije	69.88	45.08	90.51
GNNSAFE	86.33	76.53	81.35
GNNSAFE++ bez propagacije	74.87	48.58	69.05
GNNSAFE++	89.57	81.13	60.64

Slika 5. Rezultati za skup podataka Amazon gdje su anomalije stvorene interpolacijom značajki.

MODEL	AUROC	AUPR	FPR
MSP	96.75	95.05	15.01
GNNSAFE bez propagacije	98.07	95.80	5.33
GNNSAFE	98.56	98.97	0.33
GNNSAFE++ bez propagacije	98.50	96.62	3.23
GNNSAFE++	99.64	99.58	0.18

Slika 6. Rezultati za skup podataka Amazon gdje su anomalije stvorene isključivanjem oznaka.

MODEL	AUROC	AUPR	FPR
MSP	41.26	38.15	99.16
GNNSAFE bez propagacije	92.85	89.90	31.42
GNNSAFE	97.14	96.74	7.87
GNNSAFE++ bez propagacije	50.29	43.14	96.78
GNNSAFE++	97.15	96.91	9.31

Slika 7. Rezultati za skup podataka Amazon gdje su anomalije stvorene manipulacijom strukturom.

MODEL	AUROC	AUPR	FPR
MSP	78.70	82.96	93.92
GNNSAFE bez propagacije	99.47	99.53	0.61
GNNSAFE	98.68	99.26	0.00
GNNSAFE++ bez propagacije	99.03	99.30	0.44
GNNSAFE++	99.61	99.77	0.00

Slika 8. Rezultati za skup podataka Coauthor gdje su anomalije stvorene interpolacijom značajki.

MODEL	AUROC	AUPR	FPR
MSP	96.69	96.54	18.65
GNNSAFE bez propagacije	97.82	97.62	10.19
GNNSAFE	99.64	99.66	0.56
GNNSAFE++ bez propagacije	99.41	99.22	2.52
GNNSAFE++	99.97	99.94	0.06

Slika 9. Rezultati za skup podataka Coauthor gdje su anomalije stvorene isključivanjem oznaka.

MODEL	AUROC	AUPR	FPR
MSP	94.05	97.69	28.61
GNNSAFE bez propagacije	95.99	98.39	18.42
GNNSAFE	95.28	98.07	20.94
GNNSAFE++ bez propagacije	97.28	99.00	11.89
GNNSAFE++	99.83	99.23	9.87

Slika 10. Rezultati za skup podataka Coauthor gdje su anomalije stvorene manipulacijom strukturom.

MODEL	AUROC	AUPR	FPR
MSP	94.78	93.90	28.32
GNNSAFE bez propagacije	96.16	95.27	18.46
GNNSAFE	99.59	99.69	0.25
GNNSAFE++ bez propagacije	98.93	97.89	3.58
GNNSAFE++	99.99	99.99	0.02

Slika 11. Rezultati za skup podataka Twitch.

MODEL	AUROC	AUPR	FPR
MSP	37.04	51.15	96.68
GNNSAFE bez propagacije	58.71	65.78	88.98
GNNSAFE	73.24	78.47	72.14
GNNSAFE++ bez propagacije	60.05	72.47	87.80
GNNSAFE++	92.87	95.30	44.10

Performanse modela GNNSAFE i GNNSAFE++

Iz tablica vidimo da GNNSAFE++ dosljedno nadmašuje sve konkurente velikom marginom u svim slučajevima, povećavajući prosječni AUROC i smanjujući prosječni FPR95.

Bez korištenja energijske regularizacije, GNNSAFE još uvijek postiže vrlo konkurentne rezultate u OOD detekciji. Glavna prednost GNNSAFE modela je što može poboljšati pouzdanost GNN-a protiv anomalija bez dodatnih troškova računalnih resursa.

Ovi rezultati zajedno pokazuju obećavajuću učinkovitost GNNSAFE i GNNSAFE++ modela za detekciju anomalija u stvarnim primjenama povezanima s grafovima, gdje su učinkovitost i efikasnost ključni faktori.

Utjecaj propagacije vrijednosti energijske funkcije

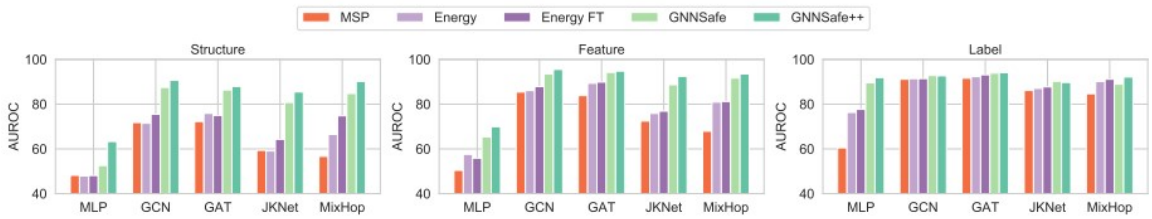
Rezultati u tablicama dosljedno pokazuju da shema propagacije doista donosi značajna poboljšanja u izvedbi. To potvrđuje hipotezu da strukturirana propagacija može isko-

ristiti topološke informacije za poboljšanje detekcije anomalija s međuzavisnošću podataka.

Analiza hiperparametara

Utjecaj enkodera

Različiti enkoderi utječu na izvedbu detekcije. Konkretno, GCN enkoderi postižu značajno bolje AUROC rezultate u usporedbi s MLP enkoderima zbog svoje jače sposobnosti kodiranja topoloških značajki.

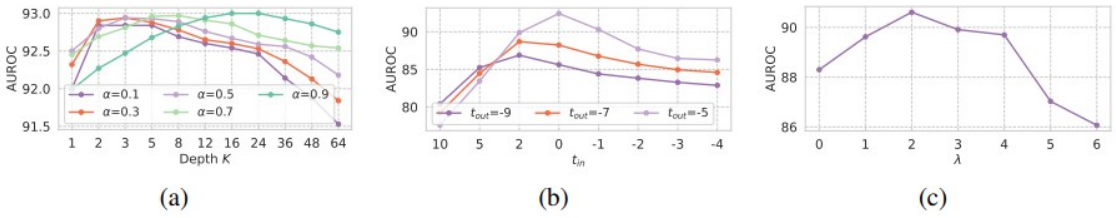


Slika 12. Usporedba performansi modela s različitim enkoderima. Izvor: Wu et al., 2023

Slika 11 uspoređuje performanse modela s obzirom na 5 različitih enkodera: MLP, GCN, GAT (Velickovic et al., 2018), JKNet (Xu et al., 2018) i MixHop (Abu-El-Haijaet al., 2019) na skupu podataka Cora.

U desnije dvije figure, čak i s MLP osnovom, GNNSAFE i GNNSAFE++ modeli ostvaruju konkurentne rezultate kao i modeli s GCN osnovama, što također potvrđuje učinkovitost sheme propagacije energije koja pomaže u iskorištavanju topoloških informacija u grafičkim strukturama (Wu et al., 2023).

Utjecaj dubine propagacije



Slika 13. (a) Analiza hiperparametara za korake propagacije K i snagu samopovezivanja α. (b) Utjecaj hiperparametra margine t_{in} i t_{out}. (c) Utjecaj hiperparametra težine λ za regularizaciju energije GNNSAFE++. Izvedeni su eksperimenti na Cora skupu podataka s manipulacijom strukture. Izvor: Wu et al., 2023

Slika 3(a) prikazuje rezultate GNNSAFE++ s brojem koraka propagacije K koji raste od 1 do 64 i snagom samopovezivanja α koja varira između 0.1 i 0.9. Kao što je prikazano na slici, AUROC rezultati prvo rastu, a zatim se pogoršavaju s povećanjem broja koraka propagacije. Razlog tome je što umjeren broj koraka propagacije može doista pomoći u iskorištavanju strukturalne ovisnosti za poboljšanje detekcije anomalija, dok prevelika dubina propagacije može previše izgladiti procijenjenu energiju i učiniti rezultate detekcije neprepoznatljivima među različitim čvorovima. Sličan trend se opaža za α, koji doprinosi boljoj izvedbi kada je postavljen na umjerenu vrijednost, npr. 0.5 ili 0.7. Prevelika vrijednost α može poništiti učinak informacija iz susjedstva, dok se preniska vrijednost α može previše fokusirati na druge čvorove (Wu et al., 2023).

Utjecaj faktora regularizcaije i margina

Promatramo performanse modela GNNSAFE++ u odnosu na konfiguracije t_{in}, t_{out} i λ na Slikama 3(b) i (c). Izvedba GNNSAFE++ uvelike ovisi o λ, koji kontrolira snagu regularizacije energije, a optimalne vrijednosti variraju među različitim skupovima podataka.

Odgovarajuće postavke za margine t_{in} i t_{out} pružaju stabilno dobru izvedbu, dok prevelike margine mogu dovesti do pogoršanja performansi.

Zaključak

Modeli za raspoznavanje grafova (GNNs) jedan su od centralnih problema dubokog učenja jer se mnogi podaci iz stvarnog svijeta mogu opisati grafovima. Jedna od velikih prepreka pri primjeni GNN-ova u industriji su preoptimistične predikcije na anomalijama.

Neke od ključnih primjena detekcije anomalija kod modela za raspoznavanje grafova su:

- Detekcija prijevara u financijskim mrežama: U financijskim transakcijama, otkrivanje neuobičajenog ponašanja ili prijevarnih aktivnosti je ključno. Detekcija anomalija može pomoći u prepoznavanju transakcija ili odnosa koji ne odgovaraju naučenim obrascima legitimnih aktivnosti.

- Analiza društvenih mreža: Prepoznavanje novog ili neočekivanog ponašanja koris-

nika, poput spam računa ili štetnih interakcija, može se olakšati detekcijom anomalija. To pomaže u održavanju integriteta i sigurnosti mreže.

- Biološke mreže: U otkrivanju lijekova i genomici, GNN se koristi za modeliranje bioloških interakcija. Detekcija anomalija može identificirati nove ili neočekivane interakcije koje mogu ukazivati na nova biološka otkrića ili potencijalne ciljeve za lijekove.

- Robotika i autonomni sustavi: Za autonomne sustave poput samostalnih automobila ili robota, detekcija scenarija izvan distribucije može pomoći u prepoznavanju situacija na koje sustav nije bio treniran, čime se poboljšava sigurnost i pouzdanost.

- Kibernetička sigurnost: U kibernetičkoj sigurnosti, otkrivanje neuobičajenog ponašanja u mrežnom prometu ili zapisima sustava može biti ključno za prepoznavanje mogućih napada. Detekcija anomalija pomaže u prepoznavanju neuobičajenih obrazaca koji odstupaju od normalnog ponašanja mreže.

- Medicinska dijagnostika: U medicinskim primjenama, GNN se može koristiti za modeliranje podataka pacijenata i napredovanja bolesti. Detekcija anomalija može pomoći u prepoznavanju novih ili rijetkih bolesti koje ne odgovaraju obrascima viđenim u podacima za treniranje.

Ovaj rad proučavao je metodu detekcije anomalija temeljenu na vrijednostima energijske funkcije i specifično prilagođenu za GNN-ove. Reproducirali smo rezultate eksperimenata iz Wu et al., 2020, koji pokazuju potencijal energijske funkcije u poboljšanju točnosti i pouzdanosti detekcije anomalija u scenarijima temeljenim na grafovima. Ključna ideja bila je da standardni GNN-ovi imaju inherentna svojstva koja se mogu iskoristiti za otkrivanje onoga što model ne zna, a to nam pruža robusni mehanizam za identifikaciju anomalija.

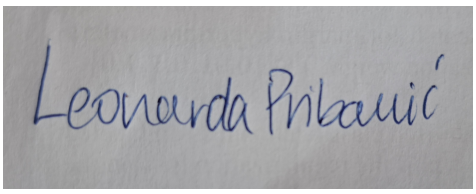
Integriranjem modela temeljenih na energiji s GNN-ovima, u eksperimentima postižemo značajna poboljšanja u detekciji anomalija na više skupova podataka i tipova grafova. U usporedbi s tradicionalnim metodama detekcije anomalija, opisani pristup poboljšava točnost detekcije.

Model temeljen na energiji ima nekoliko prednosti za detekciju anomalija. Prvo, pris-

tup temeljen na energiji pruža jasnu i interpretabilnu mjeru pouzdanosti, što je ključno za razumijevanje procesa donošenja odluka modela. Drugo, s obzirom da se za detekciju anomalija koriste svojstva koja standardni GNN-ovi već posjeduju, nema dodatnih troškova računalnih resursa. Treće, sposobnost učinkovitog rukovanja raznolikim i složenim strukturama grafova naglašava svestranost opisane metode. Ovo je posebno relevantno za stvarne primjene gdje podaci u grafovima mogu biti vrlo varijabilni i nepredvidivi.

Još uvijek postoje brojna područja koja su pogodna za daljnje istraživanje. Buduća istraživanja mogla bi se usredotočiti na usavršavanje modela temeljenog na energiji kako bi se poboljšala njegova izvedba u još složenijim scenarijima, poput onih koji uključuju dinamičke ili heterogene grafove. Također, istraživanje teorijskih temelja detekcije temeljenog na energiji unutar okvira GNN-a moglo bi pružiti dublje uvide u njegove operativne mehanizme i doprinijeti širem razumijevanju detekcije anomalija u strojnom učenju.

U sažetku, opisana tehnika predstavlja značajan korak naprijed u nastojanjima da se poboljša detekcija anomalija za GNN-ove. Korištenjem modela temeljenih na energiji, krećemo se prema pouzdanijim i učinkovitijim sustavima temeljenim na grafovima. Kako se područje nastavlja razvijati, ovaj pristup postavlja temelje za buduća unapređenja i potiče kontinuirano istraživanje inovativnih metoda za povećanje robusnosti i pouzdanosti GNN-ova u različitim i dinamičkim okruženjima.



Literatura

Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., Sun, M, Graph neural networks: A review of methods and applications. AI Open, 2020a, 1, 57-81.

Thomas N. Kipf, Max Welling, Semi-Supervised Classification with Graph Convolutional Networks, ICLR 2023, SAD, 1-9.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, Yoshua Bengio, Graph Attention Networks, ICLR 2018, SAD, 1-5.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., Monfardini, G. (2009). The graph neural network model. IEEE Transactions on Neural Networks, 20(1), 61-80.

Hamilton, W., Ying, R., Leskovec, J. (2017). Inductive representation learning on large graphs. Advances in Neural Information Processing Systems (NeurIPS), 2-9.

Hendrycks, D., Gimpel, K. (2017). A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks, ICLR 2017, SAD, 1-12.

Bishop, C. M., Novelty detection and neural network validation. IEEE Proceedings-Vision, Image and Signal Processing, 141(4), 1994., 217-222

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (Eds.), Dataset Shift in Machine Learning, SAD, MIT Press, 2009.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, Zsolt Kira, Generalized ODIN: Detecting Out-of-distribution Image without Learning from Out-of-distribution Data, CVPR 2020, SAD, 1-14

Harry Anthony, Konstantinos Kamnitsas, On the use of Mahalanobis distance for out-of-distribution detection with neural networks for medical imaging, MICCAI 2023, Kanada, 1-11.

Zhang, X., Cui, P.,Zhu, W. (2020). Deep Learning on Graphs: A Survey, Journal of Latex class files, vol. 14, no. 8, 2015, 1-24.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., Geometric Deep Learning: Going beyond Euclidean data. IEEE Signal Processing Magazine, 2017., 34(4), 18-42.

Qitian Wu, Yiting Chen, Chenxiao Yang, Junchi Yan, Energy-based Out-of-Distribution Detection for Graph Neural Networks, ICLR 2023., SAD, 1-18.

Jianwen Xie, Yang Lu, Song-Chun Zhu, and Ying Nian Wu. A theory of generative convnet, International Conference on Machine Learning, 2016, pp. 2635–2644.

Will Grathwohl, Kuan-Chieh Wang, Jorn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, Kevin Swersky, Your classifier is secretly an energy based model and you should treat it like one, International Conference on Learning Representations, 2020.

Marc'Aurelio Ranzato, Y-Lan Boureau, Sumit Chopra, and Yann LeCun. A unified energy-based framework for unsupervised learning, Artificial Intelligence and Statistics, PMLR, 2007, 371-379.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection, Advances in Neural Information Processing Systems, 2020, 33:21464–21475.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of machine learning research, 7 (11), 2006.

Yanwen Chong, Yun Ding, Qing Yan, and Shaoming Pan. Graph-based semi-supervised learning: A review. Neurocomputing, 2020, 216-230.

Xiaojin Zhu, Zoubin Ghahramani, and John D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In International Conference on Machine Learning (ICML), SAD, 2003, 912-919.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure, International Conference on Learning Representations, SAD, 2019.

Julian Bitterwolf, Alexander Meinke, Maximilian Augustin, and Matthias Hein. Breaking down out-of-distribution detection: Many methods based on ood training data estimate a combination of the same core quantities, International Conference on Machine Learning, SAD, 2022, 2041-2074.

Qitian Wu, Hengrui Zhang, Junchi Yan, and David Wipf. Handling distribution shifts on graphs: An invariance perspective, International Conference on Learning Representations, SAD, 2022.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad, Collective classification in network data, AI magazine, 2008, 93-93.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes, Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, 2015, 43-52.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang, An overview of microsoft academic service (mas) and applications, Proceedings of the 24th international conference on world wide web, , 243-246.

Benedek Rozemberczki and Rik Sarkar. Twitch gamers: a dataset for evaluating proximity preserving and structural role-based node embeddings, 2021.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs, Advances in neural information processing systems, 2020, 118-133.

Dan Hendrycks and Kevin Gimpel, A baseline for detecting misclassified and out-of-distribution examples in neural networks, ICLR 2017, SAD, 2016.

Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, Stefanie Jegelka, Representation learning on graphs with jumping knowledge networks, ICML, SAD, 2018.

Grover, A., Leskovec, J., Node2Vec: Scalable Feature Learning for Networks, International Conference on Knowledge Discovery and Data Mining, SAD, 2016, 1-5.

Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, Klaus-Robert Müller, A Unifying Review of Deep and Shallow Anomaly Detection, Proceedings of the IEEE, 2020, 3-4.

Umberto Michelucci, An Introduction to Autoencoders, 2022.

Zhang, L., Zhou, Z. H., Autoencoder with Mahalanobis Distance-Based Loss for Anomaly Detection, IEEE Transactions on Neural Networks and Learning Systems, 2020b.

Shiyu Liang, Yixuan Li, R. Srikant, Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks, ICLR, SAD, 2018.

Kimin Lee, Kibok Lee, Honglak Lee, Jinwoo Shin, A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks, NIPS, SAD, 2018.

Lee, C., Hong, M, Bayesian Mahalanobis Distance Estimation for Multivariate Gaussian Distributions, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.