Universitatea
Transilvania
din Brașov

ȘCOALA DOCTORALĂ INTERDISCIPLINARĂ

Facultatea de Inginerie Electrică și Știința Calculatoarelor

Adrian-Victor MOLDOVAN

Analiza automată a datelor

Automated Data Analysis

REZUMAT

Conducător ştiinţific

Prof.univ.dr.mat. Răzvan ANDONIE

BRAȘOV, 2024

Contents

CHAPTER 1

Introducere

1.1 Motivarea și importanța cercetării

Teza de doctorat explorează aplicarea Entropiei de Transfer (TE) în trei domenii—rețele neuronale pentru clasificarea imaginilor, comprimarea rețelelor neuronale convoluționale în cadrul teoriei informației constrânse, și rețelele convoluționale pe grafuri—întrucât TE demonstrează eficacitatea sa în îmbunătățirea, monitorizarea și evoluția algoritmilor. Cercetarea evidențiază versatilitatea TE în identificarea dependențelor informaționale asimetrice, demonstrând beneficii semnificative și puține limitări. Integrarea TE cu arhitecturile neuronale adânci facilitează optimizarea modelului și ajustarea automată, deși alegerea metodelor de estimare potrivite poate fi o provocare. Odata cu creșterea complexității și numărului de parametri ai rețelelor neuronale, TE se dezvăluie ca un instrument pentru extragerea creșterii performanței, diagnosticarea problemelor arhitecturale și atenuarea supraadaptării, în ciuda dificultăților de ajustare fină și a cerințelor computaționale suplimentare.

1.2 Contribuțiile originale

Cercerarea noastră a demonstrat potențialul integrării feedback-ului TE în diferite arhitecturi de rețele neuronale, de la rețelele feedforward simple [65] la rețele convoluționale neuronale mai complexe [66], [67] și rețele neuronale de grafuri (GNN) [68]. Algoritmul *FF+FB* a arătat rezultate promițătoare în ceea ce privește eficiența antrenării, stabilitatea și performanța. Atunci când s-a folosit TE ca o metrică pentru a descrie planurile informaționale care arată evoluția comprimării în rețelele feedforward și CNN-uri [67], TE a arătat proprietăți similare cu ale metodei informației constrânse, dar și proprietăți noi, inovatoare. În CNNs pentru sarcinile de clasificare a imaginilor, îmbunătățirile noastre au redus numărul de epoci necesare pentru limitele de acuratețe stabilite. În GCN-uri, am

folosit cu succes proprietățile relaționale ale setului de date pentru a îmbunătăți acuratețea validării. Chiar dacă există întrebări deschise și domenii pentru investigații ulterioare, lucrarea noastră contribuie la eforturile în desfășurare pentru îmbunătățirea antrenării rețelelor neuronale și înțelegerea fluxului informațional în aceste sisteme.

1.3 Structura tezei

Teza "Analiza Automată a Datelor" este structurată în două segmente majore: fundamentele teoretice acoperite în Capitolele 2 și 3, iar publicațiile noastre sunt detaliate în Capitolele de la 4 până la 7. Capitolul 2 dezvoltă conceptele fundamentale ale teoriei informației, inclusiv entropia, divergența Kullback-Leibler, și cauzalitatea Granger, explicând rolurile lor în captarea diferitelor aspecte ale conținutului și fluxului informațional în sisteme complexe.

Capitolul 3 construiește pe bazele din capitolele anterioare elemente noi precum Entropia de Transfer, Informația Constrânsă și Rețelele Neuronale pe Grafuri, care sunt centrale pentru cercetarea noastra. Elaborăm și provocările computaționale în estimarea măsurilor teoretice ale acestor elemente informaționale, în special pentru Entropia de Transfer, și impactul discretizării asupra activărilor din rețelele neuronale. Capitolul examinează, de asemenea, metoda Informației Constrânse în cadrul rețelelor neuronale, explorând implicațiile sale pentru generalizare și dinamica antrenării, recunoscându-i limitările. Rețelele Neuronale pe Grafuri (GNNs) și Rețelele Convoluționale pe Grafuri (GCN-uri) sunt introduse ca instrumente puternice pentru gestionarea datelor graf-structurate. Capitolul prezintă substanța matematică a GCN-urilor, arhitectura lor și provocările precum netezirea excesivă și heterofilia, împreună cu strategii de atenuare.

Secțiunea rezultatelor publicate începe cu Capitolul 4, care introduce un nou algoritm de antrenament pentru rețelele neuronale feedforward numit *FF+FB*. Acest algoritm utilizează TE ca un mecanism de feedback pentru a îmbunătăți performanța învățării prin măsurarea transferului informațional între neuroni și modularea conexiunilor de feedback în timpul antrenării. Experimentele pe seturi de date standard arată îmbunătățiri în viteză de convergență și acuratețe comparativ cu rețelele feedforward convenționale.

Capitolul 5 extinde această abordare la Rețelele Neuronale Convoluționale (CNNs), integrând conexiunile de feedback TE în procesul de antrenament. Rezultatele pe seturile de date de clasificare a imaginilor ilustrează convergența accelerată și îmbunătățită acurateții, cu un cost computațional suplimentar. Capitolul 6 aplică TE în cadrul metodei informației constrânse pentru a analiza fluxul informațional în rețelele neuronale, dezvăluind perspective noi în privința comprimării informației și a corelației acesteia cu performanța rețelei.

În cele din urmă, Capitolul 7 prezintă TE-GGCN, o metodă care integrează un

mecanism de control al TE în algoritmul GGCN pentru a îmbunătăți acuratețea și a aborda problemele de netezire excesivă și clasificare eronată, în special în seturile de date cu specific heterofilic. Rezultatele experimentale pe diferite seturi de date confirmă îmbunătățirea acurateții, cu cerințe computaționale sporite.

Pe parcursul tezei, integrarea TE demonstrează potențialul său de a optimiza fluxul informațional, a accelera convergența și a îmbunătăți acuratețea, subliniind utilitatea sa ca un instrument puternic în cercetarea și dezvoltarea rețelelor neuronale.

CHAPTER 2

Concepte de teoria informației

2.1 Entropia

Entropia, în contextul teoriei informației și a statisticii, este o măsură a incertitudinii sau impredictibilității stării sistemului. Ea cuantifică cantitatea de informație necesară pentru a descrie starea unui sistem sau valoarea așteptată a informației dintr-un mesaj [53].

Entropia poate fi privită ca o măsură statistică a variației și haosului dintr-un sistem. O entropie ridicată indică un grad ridicat de impredictibilitate sau dezordine în starea sistemului, reflectând un sistem mai haotic. În schimb, o entropie scăzută sugerează un sistem mai ordonat sau mai predictibil. În contextul teoriei informației, entropia reprezintă numărul minim de biți necesari pentru a codifica transmiterea stărilor într-un mesaj fără pierderea informației.

Să rezumăm rapid estimarea instrumentelor de calcul ale entropiei, oferind un experiment simplu. Visualizând entropia variabilei discrete X, care face parte dintr-o distribuție normală standard, $X \sim \mathcal{N}(\mu, \sigma^2)$, cu $\mu = 0$, $\sigma = 1$, obținem Figura 2.1. În acest grafic, am discretizat valorile lui X în segmente de lungime egală.

Tehnicile suplimentare de estimare nu vor fi detaliate în această secțiune, deoarece vom aborda metode similare în secțiunile TE și Entropie Relativă (KL). De asemenea, ecuațiile de mai sus au fost definite deoarece au o conexiune puternică cu TE prin relația indirectă cu Informația Constrânsă (IB).

2.2 Informația comună (MI)

Folosind aceeași notare ca în Secțiunea 2.1 putem defini informația reciprocă a celor două variabile X și Y ca și cantitatea de informație reciprocă conținută în ambele X și Y; putem defini MI folosind entropia lui Shannon cu următoarele [84]:

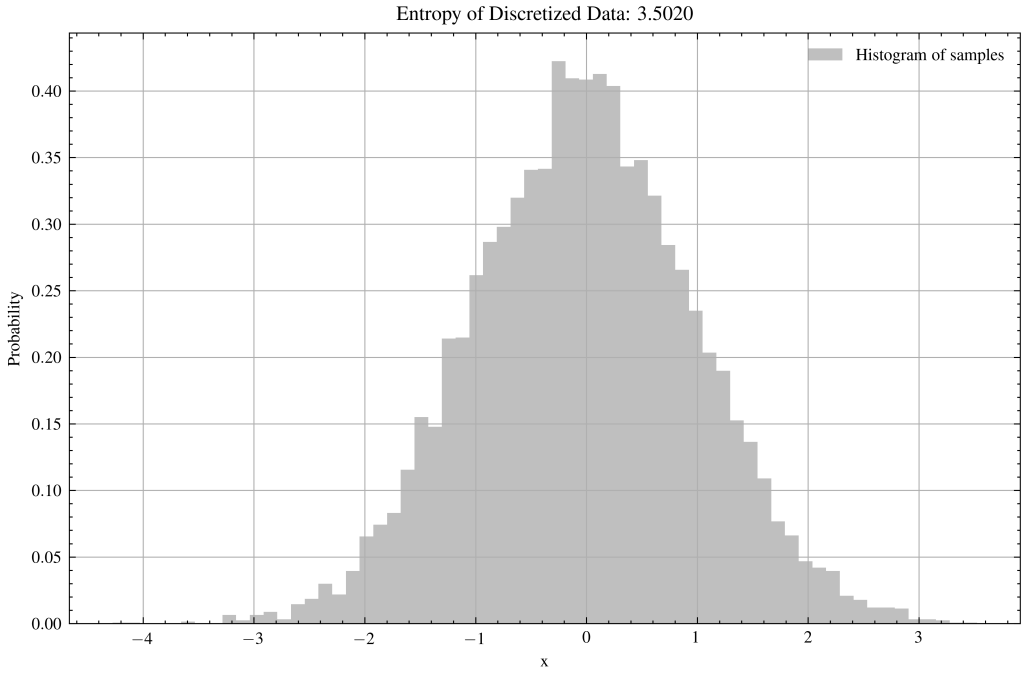

Figure 2.1: Reprezentare bazată pe segmente pentru o variabilă extrasă dintr-o distribuție normală standard. Entropia este calculată algebric prin determinarea probabilităților în fiecare segment.

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = H(Y) - H(Y \mid X) = H(X) - H(X \mid Y) \quad (2.1)$$

MI este o valoare simetrică; prin urmare, nu indică direcția fluxului de informație, ceea ce este în contrast cu TE ($TE_{X\to Y} \neq TE_{Y\to X}$). Mai mult, IB măsoară dependențele generale între X și Y fără dinamica temporală, în timp ce TE ia în considerare evoluția temporală a variabilelor concentrându-se pe tranzițiile temporale. Cu toate acestea, direcționalitatea dedusă de către TE nu oferă nicio indicație privind convergența. De asemenea, este important de menționat aici că, atunci când se folosește TE, direcția asimetriei informației poate fi garantată doar atunci când valoarea TE este zero [42].

În cursul experimentelor noastre care implicau calculul TE și MI, am adoptat tehnici de estimare eficiente din cauza cerințelor computaționale restrictive ale calculelor algebrice. Am utilizat biblioteca Scikit-Learn, pentru estimarea MI în interacțiunile straturilor rețelelor neuronale cu scopul de a crește acuratețea și eficiența antrenării. Printre diferitele metode de estimare a MI, abordările bazate pe histograme oferă simplitate dar sunt sensibile la dimensiunea segmentelor și suferă de problemele aferente inputului cu dimensiuni mari, în timp ce metodele bazate pe vecinii cei mai apropiati (KNN) se adaptează în mod dinamic la variațiile densității datelor. Abordările variaționale au simplificat calculul

MI pentru rețelele neuronale prin retropropagație, excelând cu seturi mari de date, cu dimensiuni mari, dar necesită ajustări scrupuloase ale parametrilor. Estimarea raportului densității duble, evită estimarea directă a densității, dar arată avantaje practice, în special în scenarii cu dimensiuni ridicate ale variabilelor. În pofida multitudinii de estimatori, fiecare ridică provocări legate de formele distribuțiilor și relațiile dintre variabilele de intrare, cu interacțiunile rare și distribuțiile cu cozi lungi prezintă erori de rotunjire semnificative. Variabilele cu MI cu dimensiuni ridicate necesită eșantioane mari pentru precizie, cu estimatorii neuronali care arata eficiență în gestionarea valorilor ridicate ale MI. Investigările noastre pe diverse arhitecturi neuronale au dezvăluit modele constante care reflectă aceste complexități ale estimării.

2.3 Divergența Kullback-Leibler (entropie relativă)

Divergența Kullback-Leibler (KL), introdusă de Kullback și Leibler în 1951 [52], măsoară pierderea informației asimetrică atunci când o distribuție Q este aproximată cu o altă distribuție P. Este esențială în inteligența artificială pentru optimizarea modelelor, în special în inferența variațională, modelele Bayesiene și învățarea nesupervizată, este esențială pentru sarcinile de clasificare. Definită ca $D_{KL}(P\|Q) = \sum_{x\in X} P(x) \log \frac{P(x)}{Q(x)}$, estimarea divergenței KL în dimensiuni ridicate sau cu distribuții necunoscute necesită metode precum estimarea Monte Carlo, metode bazate pe histograme, estimarea densității bazată pe nucleu, metode variaționale și tehnici bazate pe rețele neuronale precum VAE-uri și GAN-uri. Comparând divergența KL cu entropia încrucișată, aceasta poate fi văzută ca suma entropiei reale $H(y)$ și divergența KL $D_{KL}(y\|\hat{y})$. În sarcinile de clasificare, minimizarea entropiei încrucișate minimizează eficient divergența KL, aliniind distribuția prezisă cu cea adevărată, din cauza faptului că $H(y)$ este o constantă pentru etichetele claselor codate cu un singur bit. Această bază statistică susține utilizarea pe scară largă a entropiei încrucișate în problemele de clasificare în special pentru eficacitatea sa în cuantificarea erorilor de predicție relativ la valorile reale din setul de antrenare.

CHAPTER 3

Noțiuni fundamentale

3.1 Entropia de transfer

Entropia de Transfer (TE), introdusă de Thomas Schreiber [81], este o măsură a transferului de informație direcționat, extinzând entropia lui Shannon pentru a analiza dinamic sisteme complexe fără o presupusă linearitate. TE este importantă în inteligența artificială, notabil în selectarea caracteristicilor pentru prognozarea serii temporale, examinând fluxul de informații în rețelele neuronale și studiind dinamica în neuroștiința computațională. În comparație cu causalitatea Granger (GC), care presupune relații lineare și gaussiene între serii de timp, TE capturează și relații neliniare. GC, folosită inițial în econometrie, măsoară causalitatea prin modele de auto-regresie vectorială (VAR), facând distincția între forme nerestricționate și restricționate. În ciuda interesului inițial pentru integrarea GC în ajustarea ponderilor în rețelele neuronale bazate pe corelațiile dintre activările neuronale, teste noastre empirice au arătat o eficiență limitată, parțial din cauza amestecării seturilor de date și a antrenării pe loturi. Cu toate acestea, GC a arătat stabilitate atunci când a fost utilizată ca parametru în metoda de optimizare cu gradientul descendent, deși a necesitat mai multe epoci pentru o acuratețe comparabilă.

$$te_{j,i}^{r,n} = \sum_{s_i^{r,n+1},\, s_i^{r,n},\, s_j^{r,n}} p(s_i^{r,n+1},\ s_i^{r,n},\ s_j^{r,n})\ log \frac{p(s_i^{r,n+1},\ s_i^{r,n},\ s_j^{r,n})\ p(s_i^{r,n})}{p(s_i^{r,n+1},\ s_i^{r,n})\ p(s_i^{r,n},\ s_j^{r,n})} \tag{3.1}$$

TE se diferențiază de corelația clasică prin includerea precedentului temporal, un criteriu esențial pentru deducerea cauzalității [82]. În timp ce cauzalitatea examinează impactul intervențiilor, măsurile transferului de informații măsoară predictibilitatea tranzițiilor de stare [61]. TE cuantifică fluxul de informație direcționat folosind distanța Kullback-Leibler, măsurând devieri de la proprietatea Markov generalizată. În rețelele neuronale cu mai multe straturi, TE evaluează volumul de informații transferate între straturi în timpul antrenării, unde rezultatul unui strat influențează următorul strat. Formula extinsă a TE

(3.1) calculează probabilități pentru activările neuronale, reflectând relația causală între straturi. Abilitatea TE de a evalua dependințele dintre neuroni sau grupuri de neuroni le conferă o valoare în măsurarea calității comprimării datelor în rețelele neuronale [67]. Prin intermediul TE, cercetătorii pot dobândi înțelegeri despre dinamica complicată a procesării informației în rețelele neuronale, îmbogățind înțelegerea și optimizarea modelelor de inteligență artificială.

Estimarea entropiei de transfer

Calculul precis al TE pentru serii de timp lungi prezintă provocări computaționale semnificative, în special din cauza complexității estimărilor de entropie [26]. Trei metode principale abordează aceste provocări: vecinii cei mai apropiati (k-NN), discretizarea și estimarea densității bazată pe nucleu (KDE). k-NN estimează probabilități pe baza distanțelor până la vecinii cei mai apropiati, gestionând inputurile cu dimensiuni mari și dependințele neliniare mai eficient, dar este sensibilă la k și la funcțiile de distanță. Discretizarea simplifică calculul probabilităților prin segmentarea datelor, reducând necesarul computațional, dar aduce cu sine pierderi de informație din cauza erorilor de rotunjire. Binarizarea, o formă specifică de discretizare, a dovedit eficiența în studiile noastre [65–67], permițând calculul eficient al TE pentru aplicații cu rețele neuronale, în ciuda compromisurilor minore de acuratețe. $\mathrm{K}_h(x) = \frac{1}{h}K\left(\frac{x}{h}\right)$ KDE oferă estimări mai precise ale probabilităților, plasând funcțiile nucleului pe punctele de date și le sumează pentru a estima densitatea ($\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$). Cu toate acestea, KDE se confruntă cu creșterea exponențială a complexității calculelor odată cu creșterea dimensiunii inputului și este sensibilă la selecția mărimii benzii (3.1). În studiul nostru [68], un instrument bazat pe KDE [39] a facilitat calculul flexibil și robust al TE, folosind arborele K-D pentru o KDE eficientă în dimensiuni mari. Arborele K-D parționează spațiul datelor, îmbunătățind eficiența KDE la $O\left(n \log n + m \cdot n \cdot C_{\text{search}}\right)$ pentru intrări cu dimensiuni mici și scurte, unde n este numărul de puncte de date, m este numărul de căutări ale vecinilor cei mai apropiati, și C_{search} este costul de căutare. În cazul tuturor cercetărilor noastre, am constatat că o fereastră cu dimensiunea 1 este optimă în calculul TE folosit pentru îmbunătățirea acurateții modelelor.

Un studiu cuprinzător al instrumentelor de calcul și al TE a arătat dificultăți în obținerea unr rezultate consistente între biblioteci, chiar și cu seturi mici de date. Notabil, cadrul bază propus de [59], optimizat pentru date binare, s-a confruntat cu probleme de performanță atunci când a fost interfațat cu Python din implementarea sa inițială în Java. Cu toate acestea, acest framework a servit drept librărie de lucru fundamentală pentru implementarea noastră ce folosește binarizarea, ți care a accelerat semnificativ calculul TE minimizând în același timp și eroarea de rotunjire. Explorarea acestor metode subliniază importanța unei căutări eficiente și precise a estimării TE în seturi de date complexe,

evidențiind importanța alegerii metodei potrivite de estimare în funcție de caracteristicile datelor și de cerințele specifice aplicației.

3.2 Informația constrânsă

Teoria Informației constrânse (IB), a fost folosită în mod sporadic în metodele de învățare automată și inteligență artificială, și oferă perspective importante în comportamentele rețelelor neuronale și modelele de bază. Rădăcinile sale se află în teoria ratei de distorsiune(RD), inițiată de Shannon [85] și rafinată ulterior de alții [18, 23], aceasta fiind esențială în comunicație și inginerie, cu aplicații în învățarea automată din perspectiva teoriei informației. Rețelele neuronale învață distribuțiile de intrare prin estimări, și aceasta implică procese cu pierdere de informație și etape de cuantificare care introduc distorsiuni. Compromisul dintre rată și distorsiune este ilustrat în Figura 3.1, evidențiind compromisul între rată (R) și distorsiune (D), unde o distorsiune cu valoarea 0 corespunde unei acurateți de 100%. Performanța rețelelor neuronale, inclusiv acuratețea și generalizarea, este intrinsec limitată de constrângerile arhitecturale care afectează metricile de distorsiune.

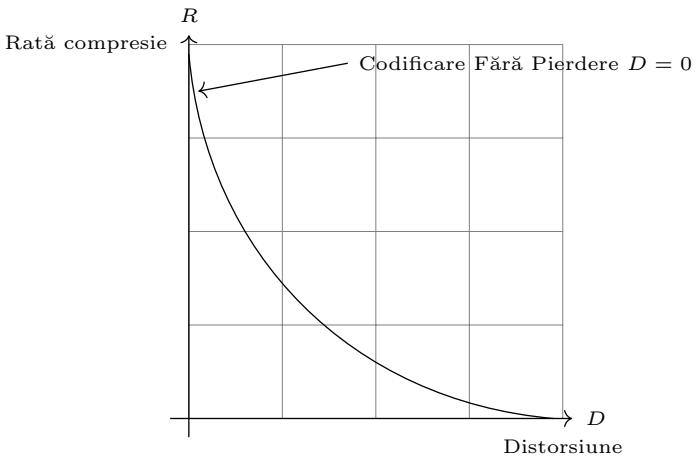

Figure 3.1: Acest grafic a fost inspirat din cursurile lui Bernd Girod despre compresia imaginilor și a filmelor: EE368b Comprimarea Imaginelor și a Videoului Teoria Ratei Distorsiune nr. 2

Teoria ratei de distorsiune caută rata minimă necesară pentru o inferență exactă prin intermediul unei rețele neuronale, având un prag de distorsiune. Considerând $X \sim \mathcal{N}(\mu, , \sigma^2)$, cu R biți ce codează un simbol din X și $d(x, y) = (x^2 - y^2)$ ca măsură de distorsiune a erorii pătrate, funcția de RD minimizează informația comună (MI) între X și reconstruitul Y sub constrângerea unei distorsiuni. Optimizarea are în obiectiv toate distribuțiile condiționate $p(x \mid y)$ care satisfac $\mathbb{E}[d(X, Y)] \leq D$, având ca scop minimizarea MI în timp ce reduc rata necesară pentru a satisface un nivel de distorsiune predefinit. Pragul inferior al lui Shannon pentru eroarea pătrată a distorsiunii este dat prin obiectivul

teoriei care dorește echilibrarea păstrrării informației și eficiența transmiterii în procesele de învățare ale rețelelor neuronale.

Metoda Informației Constrânse

Metoda Informație Constrânse (IB), distinctă de Teoria Ratei de Distorsiune, se concentrează pe identificarea celor mai relevante informații dintr-o variabilă pentru a prezice alta. Acest lucru se realizează prin introducerea unei variabile de constrângere T, care este reprezentarea comprimată a X, ce are ca scop stocarea a cât mai multe informații posibile despre Y (3.2). Parametrul β echilibrează compresia și predicția.

$$min_{P_{T|X}}(I(X;T) - \beta I(Y;T)) \tag{3.2}$$

Cercetări extinse ([2, 14, 29, 31, 43, 79, 83, 89, 90, 94, 95]) au explorat IB în rețelele neuronale, construind planuri de informații din activările straturilor pentru a dezvălui fazele de ajustare și comprimare, demonstrând importanța menținerii unui echilibru între compresie și predicție pentru o acuratețe optimă.

Mai multe studii au utilizat direct IB în diferite mecanisme, de la selecția modelului la îmbunătățirea timpului de antrenament, la îmbunătățirea acurateții și performanței de generalizare ([14, 79]). Arhitecturile mai adânci se dovedesc a oferi compromisuri mai bune pentru sarcinile de clasificare a imaginilor, și atribuie acest aspect pozitiv capacității lor de a conserva mai multe informații relevante. Algoritmul Blahut-Arimoto ([4, 7]) facilitează calculul IB pentru variabilele X și Y cu $I(X;Y)$ pozitiv. Algoritmii de grupare modificați cu IB ating niveluri mai ridicate de compresie ([90]), în timp ce compromisurile dintre complexitate și performanță sunt reduse ([27]). În folosirea rețelelor de învățarea adâncă [95] dezvăluie faze distincte de antrenare legate de performanța rețelei, aratând că algoritmul de optimizare cu decendent de gradient prezintă o formă de difuzie aleatorie a gradienților, ce contribuie pozitiv la compresie.

Rolul IB în compresie și predicție este subtil, cu funcțiile de activare având un impact semnificativ asupra capacității de compresie ([79]). Compresia este adesea observată doar în straturile de clasificare, cu indexurile straturilor superioare prezentând o variație mai mică a MI. IB, folosită ca și funcție de eroare din strat în strat ([22]) și estimarea precisă a $I(X;T)$ ([29]) îmbogățesc înțelegerea mecanismelor de grupare a claselor și compresiei. Învățarea neasupravegheată multi-unghi ([25]) extinde utilitatea IB, în timp ce erorile de generalizare sunt corelate cu gradul IB ([43]). Progresele în modele generative ([88]) și în rețelele neuronale adânci regularizate liniare pe sarcini de clasificare gaussiană ([34]) demonstrează rolul multiplu al IB în arhitecturile și algoritmii de învățare adâncă.

3.3 Rețelele neuronale pe grafuri

Rețelele neuronale pe grafuri (GNN-uri) reprezintă o frontieră în învățarea profundă, ele excelând în gestionarea datelor relaționale și cu structură de graf, abordând sarcini de la clasificarea nodurilor până la predicția legăturilor și generarea grafurilor ([5, 6, 13, 15, 20, 21, 32, 33, 37, 40, 41, 44, 58, 73, 74, 78, 80, 91, 98, 105, 114, 115, 118, 119]). Arhitectural, GNN-urile cuprind un spectru care se întinde de la Rețelele Neuronale Convoluționale pe Grafuri (GCN) la Rețelele Neuronale cu Atenție pe Grafuri (GAT), Autoencoderii pe Grafuri (GAE), Rețelele Generative Adversarale pe Grafuri (GGAN), Rețelele Neuronale Recurente pe Grafuri (GRNN), Transformatorii pe Grafuri, Rețelele Isomorfice pe Grafuri (GIN), și GNN-uri specializate care utilizează caracteristicile muchiilor sau ale nodurilor ([47, 48, 96, 99, 106, 111, 113]). Metodologic, se clasifică de la metodele bazate pe spectru la cele bazate pe spațiu, rețele neuronale cu trimitere de mesaje (MPNNs), cu straturi adânci, învățarea cu recompensă, învățarea prin transfer, metode de învățare inductivă și transducțivă, adaptându-se la diferite complexități ale grafurilor.

Sperduti *et al.* [92] au inițiat rețelele neuronale cu structură de tip graf, evoluând către rețeaua neuronală pe graf a lui Gori *et al.* [30]. Scarselli *et al.* [80] a rafinat cadrul, integrând topologia și caracteristicile nodurilor. Bruna *et al.* [9] au introdus rețele spectrale, pavând calea spre utilizarea filtrării spectrale localizată de către Defferrard *et al.* [17], culminând în lucrarea seminală a lui Kipf *et al.* [48] care a simplificat arhitecturile pentru clasificarea semi-supervizată, stabilind baza pentru GCN contemporane. Inovațiile continuă, cu studii care îmbunătățesc convoluțiile spectrale și procesarea semnalului în grafuri ([17, 50, 54, 87]). Versatilitatea și scalabilitatea GNN-urilor în cadrul aplicațiilor interdisciplinare au declanșat un interes intens de cercetare, făcându-le indispensabile pentru sarcini cu date complexe și interconectate. Secțiunea următoare dezvoltă rețelele convoluționale pe grafuri, provocările lor și soluțiile, ghidată de contribuțiile lui Kipf *et al.* [48] la progresele recente ale lui Yan *et al.* [108].

3.3.1 Rețele neuronale convoluționale pe grafuri

Rezumând componentele unui graf: $G = (V, E)$ denotă un graf cu mulțimea nodurilor V și mulțimea muchiilor E; matricea de adiacență A reflectă conectările nodurilor; gradul nodului v_i, d_i, este suma muchiilor sale; matricea are gradul D cu elementele diagonale $D_{ii} = \sum_j A_{ij}$; matricea de incidență K arată conexiunile dintre nod și muchie; $X \in \mathbb{R}^{n \times o}$ reprezintă matricea caracteristicilor nodurilor cu o ca dimensiunea caracteristicii; $\mathbf{x}_v$ reprezintă vectorul de caracteristici al nodului v; $X^{(l)}$ și $\mathbf{x}^{(l)}$ denotă matricele de caracteristici ale stratului l.

Matricea Laplaciană a grafului

Matricea Laplaciană a grafului, esențială în rețelele de tip graf și în procesarea semnalului, încapsulează conexiunea dintre noduri și rugozitatea semnalului. Derivând matricea laplaciană ca $L = D - A$, aceasta reflectă a doua derivată a unei funcții, aratând modificările valorilor între nodurile conectate.

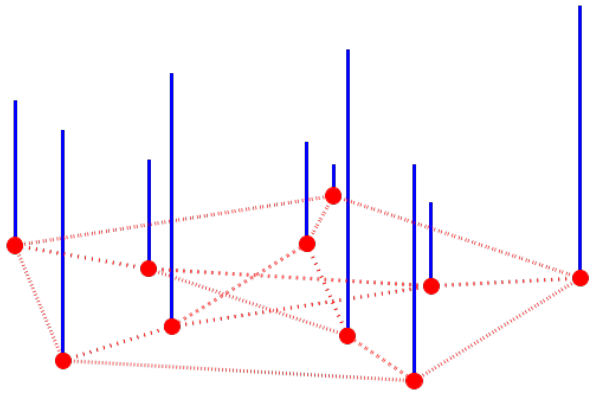

Figure 3.2: Un semnal pozitiv pe un graf Petrescu (după [87]). Fiecare nod are un semnal asociat proporțional cu înălțimea barei albastră.

Vizualizată ca un "acoperiș flexibil" sensibil la semnalele din noduri, așa cum este ilustrat în Figura 3.2, vectorii proprii ai Laplacianului se modifică odată cu frecvența semnalului pe întreg graful ([87]). În afară de măsurarea conectivității și difuziei, este esențială pentru identificarea grupurilor de valori în analiza grafului.

Convoluțiile pe graf

Convoluțiile pe graf, spre deosebire de convoluția din CNN-urile tradiționale, se confruntă cu structura neregulată a datelor dintr-un graf, neavând o ordine definită a nodurilor sau un număr constant de vecini [56, 57, 87, 112]. Transformarea structurii grafului în domeniul spectral facilitează operațiunile convoluționale, oferind mai multe avantaje: separarea intrărilor în frecvențe pentru procesare adaptată, calcul eficient prin diagonalizarea matricei laplaciene, invariantă față de ordinea nodurilor, captarea structurii globale și flexibilitatea proiecțiilor în aplicarea filtrelor specifice frecvențelor. Cu toate acestea, localizarea spațială este mai puțin explorată în domeniul spectral, iar grafurile dinamice necesită recalculări frecvente ale descompunerii spectrale. Convoluțiile spectrale pe graf implică transformarea semnalelor în domeniul frecvenței, aplicarea filtrelor, și revenirea în domeniul spațial, folosind transformata Fourier a grafului $\hat{x} = U^\top x$ și inversul său $x = U\hat{x}$. Convoluția în domeniul nodurilor corespunde înmulțirii în domeniul spectral $\hat{y} = \hat{g} \odot \hat{x}$, cu filtre spectrale aproximate folosind polinoamele Chebyshev truncate pentru eficiență computațională

($g_\theta(\Lambda) \approx \sum_{k=0}^{K} \theta_k T_k(\tilde{\Lambda})$, $g_\theta \star x \approx \sum_{k=0}^{K} \theta_k T_k(\tilde{L})x$). Această aproximație permite calculul direct în domeniul spațial, evitând descompunerea costisitoare a valorilor proprii, micșorând diferența dintre Laplacianul grafului și convoluții, dar și îmbunătățind scalabilitatea [17, 48].

În esență, descompunerea valorilor proprii a Laplacianului grafului susține transformata Fourier, facilitând reprezentarea și procesarea semnalelor grafului în domeniul spectral. Convoluțiile eficiente și filtrarea se obțin prin înmulțirea în domeniul spectral, cu aproximațiile polinoamelor simplificând calculele pentru analiza grafurilor pe scară largă, îmbogățind astfel scalabilitatea și aplicabilitatea GCN.

GCN-uri cu mai multe Straturi

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{3.3}$$

GCN-urile cu mai multe straturi, inițiate de Kipf *et al.* [48], aplică convoluțiile pe straturi pentru a agrega caracteristicile nodurilor din vecinătățile locale, legând abordările spectrale și convoluționale fără metode spectrale directe. Simplificările includ limitarea polinoamelor Chebyshev la ordinul (K=1) și utilizarea unei parametrizări specifice pentru a evita netezirea excesivă [8, 77, 108], dar și introducând un truc de renormalizare pentru a diminua problema gradienților care cresc excesiv sau dispar [46]. Regula de propagare a convoluțiilor (3.3) echilibrează influența nodurilor asupra vecinilor, folosind matricile de adiacență auto-legate și ponderi antrenate. Filtrele, reprezentate ca $\Theta \in \mathbb{R}^{C\times F}$, procesează semnalele de intrare $X \in \mathbb{R}^{N\times C}$ prin agregare normalizată și înmulțirea parametrilor ($Z = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X\Theta$). GCN-urile cu mai multe straturi generalizează acest proces, încorporând funcțiile de activare, așa cum este ilustrat în Figura 3.3, reprezentând agregarea caracteristicilor nodului prin straturile convoluționale, unde nodul A actualizează propriile sale proprietăți folosind caracteristicile agregate din nodurile vecine, demonstrând transferul mesajelor în GNN-uri.

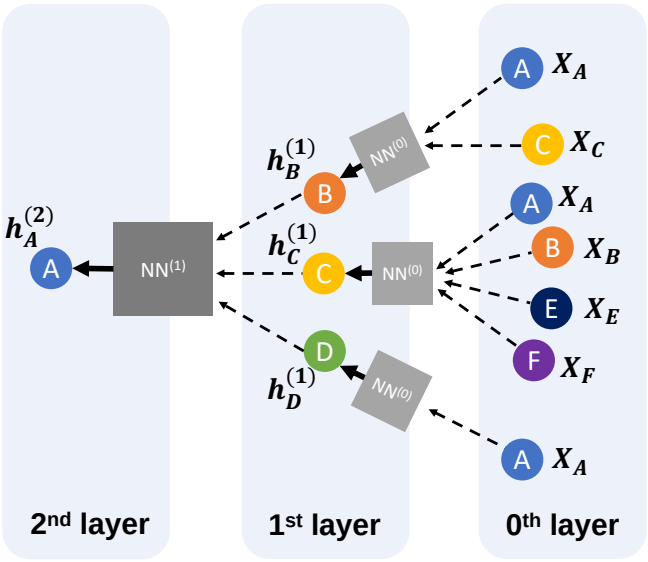

Figure 3.3: Agregarea caracteristicilor nodului folosind o convoluție cu două straturi(ilustrație din [110]). Nodul A actualizează propriile sale proprietăți folosind caracteristicile agregate din toate celelalte noduri.

Netezirea excesivă, heterofilia și homofilia

În [68] ne-am confruntat cu unele dintre problemele bine cunoscute și încă deschise din lumea GNN: netezirea excesivă din cauza arhitecturii și a optimizării, și heterofilia și homofilia din punctul de vedere al proprietăților conectivității grafului. Cele două din urmă sunt atribute ale componentelor unui graf, cum ar fi nodurile sau subgrafurile. Toate cele trei au un impact critic asupra capacității discriminative ale unui GCN.

Netezirea excesivă(oversmoothing)

Suprasemnarea este un fenomen observat în GCNs unde reprezentările nodurilor devin tot mai similare pe măsură ce numărul de straturi crește. Nu este doar un produs al setului de date, cum ar fi grafurile conectate dens, ci este, de asemenea, rezultatul arhitecturii GCN. Convergența către o limită neinformativă împiedică performanța GCN-urilor adânci, în special pe grafurile heterofile unde nodurile cu etichete diferite sunt conectate [12, 56, 77]. Chiar dacă netezirea excesivă poate ajuta în sarcinile de regresie și clasificare în cantități mici, netezirea excesivă poate fi dăunătoare [77]. Cu alte cuvinte, agregarea informațiilor din vecinătăți poate duce la modificarea nodurilor din vecinătăți prea similară, chiar dacă aparțin altor clase. Pe măsură ce cresc straturile, câmpul receptiv se extinde, diminuarea caracteristicilor unice ale nodurilor devine problematică [72].

Matematic, problema netezirii excesive poate fi înțeleasă prin spectrul valorilor proprii

ale grafului Laplacian. În GCN-uri adânci, valorile proprii asociate cu Laplacianul graf pot deveni prea mari, determinând convergența vectorilor de caracteristici către un vector constant [56]. Această convergență este accentuată de adâncimea rețelei, straturile mai adânci amplificând efectul de netezire excesivă.

Heterofilia

Heterofilia se referă la caracteristica unui graf ale carui noduri conectate au tendința să aibă etichete sau caracteristici diferite [100, 116]. Această proprietate prezintă o provocare pentru GCN-urile tradiționale care performează bine pe grafurile homofile, unde nodurile conectate tend să împărtășească atribute similare [100].

GCN-urile tradiționale, proiectate cu o presupunere implicită de homofilie, se confruntă cu dificultăți în gestionarea grafurilor heterofile din cauza dependenței lor de agregarea vecinilor pentru a prezice etichetele [64, 100], [63, 116]. Proiectarea modelelor GCN tradiționale poate fi considerată neadecvată pentru grafurile heterofile, deoarece utilizarea implicită a homofiliei poate accentua atât netezirea excesivă cât și efectele negative ale heterofiliei [100, 108, 116]. Aplicate pe grafurile heterofile, aceste GCN-uri pot experimenta o degradare a performanței deoarece procesul de agregare amestecă informațiile din noduri cu etichete diferite, ducând la reprezentări mai puțin informative [63, 64, 101].

Homofilia

Homofilia este tendința nodurilor cu caracteristici similare, cum ar fi etichetele sau caracteristicile, de a fi conectate într-un graf [1, 63, 64, 116]. Multe GNN-uri se bazează implicit pe această presupunere, ceea ce duce la limitări în gestionarea grafurilor heterofile [63, 64].

Și probabil întrebarea evidentă acum este dacă aceste proprietăți sunt cauzate de designul GCN, sau homofilia este o caracteristică inerentă a seturilor de date, sau atât unul cât și celălalt? Apariția netezirii excesive, heterofiliei și homofiliei provine atât din proprietățile inerente seturilor de date de tip graf, cât și din designul modelelor GCN. Ambele *heterofilie și homofilie sunt caracteristici ale seturilor de date.* Grafurile din lumea reală prezintă adesea grade variate de heterofilie, în funcție de natura relațiilor dintre noduri [63], [101]. Homofilia este răspândită în rețelele sociale, rețelele de citare și alte domenii unde entitățile similare tind să se conecteze [64], [63], [75].

Metode de atenuare

Din limitele teoretice ale heterofiliei și homofiliei, acestea nu sunt limitate în mod intrinsec; totuși sunt limitate de designul GCN și caracteristicile setului de date. Acestea sunt încă domenii de cercetare active. De exemplu [69] sugerează că alegerea nucleului în convoluțiile

spectrale pe graf poate influența gradul de netezire excesivă. De asemenea, conceptul de "timp de amestecare" în mișcările aleatoare pe graf poate furniza înțelegeri asupra modului în care heterofilia sau homofilia pot apărea în predicțiile GCN [11]. [12] a arătat că proiectarea GCN-urilor adânci cu mai puține straturi poate ajuta la reducerea netezirii excesive, în timp ce [96] a introdus mecanisme de atenție pe graf pentru a selecta în mod selectiv vecinii importanți și a atenua netezirea excesivă [55].

Cu toate acestea, impactul acestor proprietăți poate fi atenuat prin modelare GNN specializată și alte tehnici. Acestea includ scheme de agregare adaptive [75, 100], luând în considerare direcția muchiilor și disimilaritatea nodurilor, încorporând informații despre vecinii de ordin înalt [100, 104, 120], utilizând tehnici precum încorporarea egoului și separarea încorporării vecinului [75], atenția pe graf [96], eșantionarea grafurilor [75, 100, 104] au arătat îmbunătățiri importante. GCN-urile pot exploata homofilia pentru o performanță mai bună pe grafurile homofile prin accentuarea agregării vecinătății locale. Cu toate acestea, așteptarea excesivă de prezență a homofiliei poate fi dăunătoare pentru grafurile heterofile [120]. Suprasemnarea poate fi limitată prin limitarea numărului de straturi GCN, adăugarea conexiunilor reziduale și convoluțiilor dilatărate, utilizarea conexiunilor reziduale, utilizarea legăturilor skip, implementarea de noi strategii de normalizare, sau integrarea abandonului muchiilor [108], sau se pot lua în considerare modele care pot utiliza atât contextul local cât și global [120]. Aceste tehnici ajută la păstrarea diversității caracteristicilor nodurilor și la prevenirea convergenței către o valoare constantă [107, 108].

Deși benefice în contextele homofile, așteptările excesive de homofilie ar trebui evitate. Tehnicile precum amestecarea canalelor adapative pot ajuta la echilibrarea între agregare, diversificare și canalele de identitate pentru a aborda diferite situații de homofilie [62, 63].

În concluzie, înțelegerea și abordarea suprasemnării, heterofiliei și homofiliei este crucială pentru dezvoltarea modelelor GCN eficiente. În timp ce netezirea excesivă poate fi atenuată prin modificări arhitecturale, gestionarea heterofiliei și exploatarea homofiliei în mod eficient necesită tehnici specializate și abordări adapative. Direcțiile viitoare de cercetare ar trebui să se concentreze pe dezvoltarea modelelor GCN mai robuste și mai generalizabile care pot învăța eficient din grafuri cu niveluri variate de homofilie și heterofilie.

3.3.2 Tehnici de regularizare pe grafuri

Abordarea netezirea excesivă, heterofiliei și homofiliei în Rețelele Convoluționale de tip Graf (GCN) este esențială pentru îmbunătățirea capacității lor discriminative. Netezirea excesivă, caracterizată prin reprezentări tot mai indiscernibile ale nodurilor în straturile mai adânci, subminează performanța GCN, în special pe grafurile heterofile unde nodurile cu etichete diferite sunt conectate [12, 56, 77]. Matematic, aceasta decurge din spectrul valorilor proprii al Laplacianului graf, unde valorile proprii prea mari conduc la convergența vectorilor

proprii [56]. Heterofilia, tendința nodurilor conectate de a avea etichete disimilare, prezintă o provocare pentru GCN-urile tradiționale proiectate sub presupunerea de homofilie, ducând la degradarea performanței [64, 100]. În schimb, homofilia, unde nodurile similare sunt conectate, este adesea atribuită în mod implicit de GCN-uri, punând la îndoială capacitatea lor de a gestiona grafurile heterofile [63, 64]. Strategiile de atenuare includ scheme de agregare adapative, mecanisme sensibile la muchii, luarea în considerare a vecinilor de grad mare, și tehnici precum embedding-ul ego și atenția pe graf [75, 96, 100]. Tehnicile de regularizare, în special cele bazate pe Laplacian și cele care nu folosesc Laplacianul, joacă un rol crucial în limitarea netezirii excesive, știind că regularizarea Laplacianului promoveaza similaritatea între vecini [3, 109] și PairNorm [117] menține distanțele caracteristicilor perechi între straturi. DropEdge [76], o tehnică care nu foloseste Laplacianul, introduce eliminarea muchiilor pentru augmentarea datelor, consolidând și prevenind netezirea excesivă în GCN-urile adânci.

Interacțiunea dintre netezirea excesivă, heterofilie și homofilie în GCN este multifațetată, influențată atât de proprietățile setului de date cât și de arhitectura GCN. Netezirea excesivă poate fi accentuată de designul GCN, în special în grafurile strâns conectate, ducând la dispariția particularităților caracteristicilor nodului [72]. Heterofilia și homofilia, sunt caracteristici inerente ale seturilor de date, ce contestă presupunerile clasice al GCN-urilor, necesitând abordări adaptive pentru o performanță eficientă. Tehnicile de regularizare bazate pe Laplacian, precum cea propusă de Ando *et al.* [3], au ca scop menținerea similarității etichetelor între vecini, oferind beneficii în păstrarea caracteristicilor dar cu un impact limitat asupra GNN-urilor care deja captează informații structurale.

În rezumat, provocările generate de netezirea excesivă, heterofilie și homofilie în GCN subliniază nevoia de strategii adaptive și de regularizare. Tehnicile bazate pe Laplacian, cum ar fi regularizatorul Laplacian de graf, au ca scop păstrarea structurii grafului prin încurajarea similarității caracteristicilor între nodurile adiacente [102, 121]. Penalizările la nivelul nodului și la nivelul muchiei din PairNorm promovează păstrarea structurii grafului la nivel local și global, asigurând păstrarea caracteristicilor pe tot graful. Mecanismul de eliminare a muchiilor din DropEdge introduce cauzalitate și diversitate, prevenind netezirea excesivă și în arhitecturile GCN adânci.

CHAPTER 4

Entropia de transfer cu rețele neuronale cu puține straturi

Această capitol prezintă un algoritm inovator de antrenament, *FF+FB*, pentru rețelele neuronale feedforward care se bazează pe relațiile cauzale prin intermediul feedback-ului Entropiei de Transfer (TE), având ca scop îmbunătățirea eficienței învățării [65]. TE, utilizată tradițional pentru a măsura conectivitatea eficientă între neuroni [24, 60, 86, 97], este refolosită pentru a măsura transferul de informații între straturile adiacente, accentuând relevanța conexiunilor. În comparație cu aplicările anterioare ale TE în rețelele neuronale [35, 71], abordarea noastră integrează TE direct în procesul de actualizare a ponderilor prin retropropagație, perfecționând algoritmul standard prin includerea unui mecanism de feedback care ia în considerare TE între perechi de neuroni.

FF+FB este structurat în două etape: Etapa (I) calculează valori TE în timpul antrenării, stocându-le pentru toate perechile de neuroni; Etapa (II) reantrenează rețeaua folosind aceste valori stocate, modificând descendentul gradientului pentru a include feedback-ul TE (Ec. 4.1). Această adaptare accelerează procesul de învățare și îmbunătățește acuratețea, așa cum este demonstrat de problema XOR, unde *FF+FB* atinge o acuratețe de antrenament de 100% într-un număr semnificativ mai mic de epoci decât *FF* (7-10 ori mai puțin, în medie 62,2 epoci vs. 349,9 epoci). Pe zece seturi de date UCI [19], *FF+FB* demonstrează, de asemenea, superioritate, atingând acuratețea țintă mai rapid și obținând o acuratețe mai mare pe setul de testare în majoritatea cazurilor.

$$\Delta w_{ij}^{l} = -\eta \frac{\partial C}{\partial w_{ij}^{l}} (1 - te_{j,i}^{l}) \tag{4.1}$$

Hiperparametrii joacă un rol crucial în performanța *FF+FB*. Rata de învățare η și pragul de binarizare g necesită ajustări atente; *FF+FB* se bucură adesea de valori mai mici ale η, sugerând o abordare de învățare mai orientată spre obiective. Cu toate acestea, valori mici

ale η și g pot împiedica rețeaua să scape din minimele locale, necesitând alegerea judicioasă a g prin căutare pe grilă. Experimentele de control, inclusiv modificările aduse valorilor *te* și ponderarea lor, confirmă robustețea algoritmului și evidențiază adaptabilitatea sa de a compensa selecțiile suboptimale ale η.

Sursele suplimentare de calcul ale *FF+FB*, în principal în Etapa (I), sunt echilibrate de eficiența superioră a antrenării și creșterea acurateții. Performanța algoritmului pe seturile de date *car* și *glass*, în ciuda provocărilor initiale, validează competitivitatea sa față de pachete stabilite precum Weka. Considerațiile practice sugerează că prețul computațional crescut în timpul etapei de antrenament (I) este nesemnificativ pentru sarcinile de inferență, deoarece ponderile antrenate pot fi stocate, încapsulând valorile *te*. Acest lucru face *FF+FB* viabil pentru aplicații din lumea reală, chiar și cu seturi mari de date, deoarece valorile *te* calculate în etapa (I) pot fi reutilizate în etapa (II) fără costuri suplimentare. Tehhnici alternative de estimare ale TE ar putea atenua sarcina computațională în astfel de scenarii [10].

În concluzie, *FF+FB* reprezintă un avans semnificativ în algoritmul de antrenare a rețelelor neuronale, utilizând TE pentru a măsura și a îmbunătăți relațiile cauzale dintre neuroni. Nu doar că se reduc numărul de epoci necesare pentru antrenament și îmbunătățește acuratețea, dar oferă și stabilitate și rezistență față de minimele local, așa cum se arată în graficele din versiunea completă a tezei. Optimizarea pragului g poate atenua impactul altor hiperparametri, sugerând o soluție spre ajustarea mai eficientă a modelului. Potențialul pe care *FF+FB* îl are in a facilita extragerea cunoștințelor și explicațiile rețelelor antrenate, deși rămâne ca un problemă deschisă, evidențiază implicațiile sale mai largi pentru înțelegerea rețelelor neuronale în procesele de luare a deciziilor.

CHAPTER 5

Entropia de transfer în rețele neuronale convoluționale

Lucrarea noastră extinde aplicarea entropiei de transfer (TE) la rețelele neuronale convoluționale (CNN), având ca scop îmbunătățirea mecanismelor de antrenare și creșterea interpretabilității [66]. TE, o măsură a transferului de informație direcționat, cuantifică relațiile dintre ieșirile neuronilor din straturile adiacente, acționând ca un factor de amestecare care stabilizează procesul de învățare și accelerează convergența. În ciuda informației simetrice, asimetria TE se aliniază cu natura cauzală a straturilor rețelelor neuronale. Inspirat de [35], abordarea noastră inovatoare integrează direct TE în retropropagație, actualizând greutățile în funcție de TE între perechi de neuroni. Aceasta se diferențiază de metoda lui Herzog *et al.*, care folosea TE pentru structurarea conexiunilor de feedback după antrenament [36]. Experimentele noastre pe CNN-uri cu feedbackul TE demonstrează performanțe și stabilitate îmbunătățite, în special în ultimele două straturi complet conectate, asemănător cu perfecționarea mecanismelor de clasificare (Figura 5.1).

Calculul TE în CNN, în special pentru serii lungi de timp, este intensiv din punct de vedere computațional [26]. Optimizăm integrarea TE prin limitarea lungimii seriilor de timp și utilizarea unei tehnici de fereastră glisantă pe loturi, care menține acuratețea în timp ce gestionează costurile computaționale suplimentare. Lungimea ferestrei s, care ideal ar trebui să corespundă dimensiunii lotului, facilitează valori mai fine ale TE și tendințe favorabile ale acurateții. Atenția noastră asupra ultimelor două straturi complet conectate, în loc de straturile convoluționale, reduce cerințele computaționale în timp ce impactează semnificativ acuratețea, asemănător efectului dropout-ului dar cu un accent pe îmbunătățirea performanței. Parametrii determinați experimental și natura adaptativă a TE ca meta-parametru contribuie la robustețea și stabilitatea algoritmului.

Rezultatele experimentale care pot fi citite în teza completă, demonstrează eficiența TE în accelerarea antrenării CNN pentru a ajunge la acuratețile țintă cu un număr mai mic

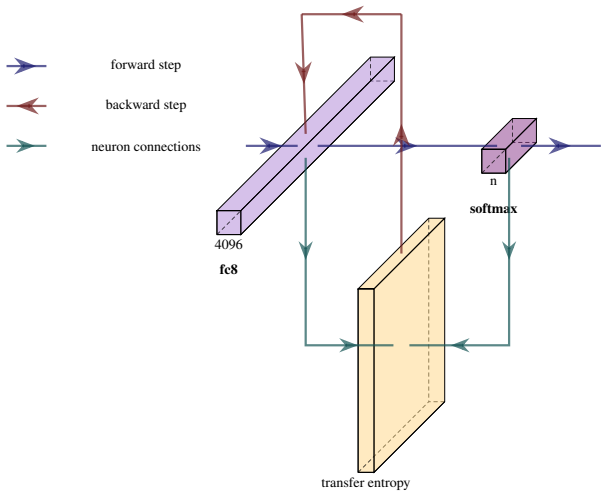

Figure 5.1: În timpul pasului feedforward, calculăm seria temporala I și J, precum și matricea **te**, așa cum este arătat prin săgețile verzi. Atunci când pasul retropropagației transmite erorile înapoi, matricea **te** este folosită în actualizările ponderilor.

de epoci. Am observat că utilizarea feedback-ului TE pentru o pereche suplimentară de straturi îmbunătățește performanța, dar cu un cost exponențială al sarcinii computaționale. Optimizarea compromisului între performanță și costuri suplimentare este dependentă de aplicație, necesitând o atenție riguroasă asupra rolului TE ca un meta-parametru care se schimbă lent. Experimentele noastre cu rețele preantrenate, unde doar cele două straturi finale trec prin corectarea TE, duc la rezultate neuniforme, indicând importanța integrării sincronizate a TE cu procesul de retropropagație.

În concluzie, studiul nostru confirmă utilitatea TE în îmbunătățirea antrenării CNN, în special în straturile finale, reflectând eficacitatea sa în rețelele feedforward simple [65]. Sarcina computațională suplimentară este atenuată prin concentrarea asupra unui subansamblu de perechi de neuroni, aliniindu-se cu structurile de feedback din sistemele neuronale biologice [28, 93]. Rolul TE ca factor de amestecare și activarea sa periodică contribuie la stabilitatea și generalizarea algoritmului de învățare, similar cu ierarhia parametrilor în modelele neuronale cauzale de învățare [45]. Rezultatele noastre sugerează că integrarea TE ar putea avea paralele evolutive în sistemele neuronale reale, optimizând relevanța în canalele feedforward [36]. Cercetările viitoare ar putea explora impactul TE asupra arhitecturilor mai profunde de rețele și potențialul său pentru îmbunătățirea interpretabilității în CNN.

Experimentele realizate pe o serie de seturi de date bine cunoscute (CIFAR-10 [51], FashionMNIST [103], STL-10 [16], SVHN [70], și USPS [38]) folosind o arhitectură standardizată a CNN și hiperparametrii subliniază eficacitatea TE. Influența pozitivă a feedback-ului TE asupra stabilității și acurateții antrenării este evidentă, cu îmbunătățiri notabile chiar și atunci când se folosește o fracție (10%) din perechile de neuroni din cele două straturi complet conectate, finale. Această eficiență sugerează potențialul pentru TE de a servi ca

un instrument pentru a înțelege și a optimiza dinamica rețelelor neuronale, asemănător cu înțelegerile dobândite din studiul sistemelor neuronale biologice. Compromisul între beneficiile TE și costurile computaționale suplimentare este o considerare cheie, ghidând utilizarea optimă a TE în aplicații din lumea reală.

CHAPTER 6

Entropia de transfer în metoda informației constrânse

MI și TE oferă perspective distinse asupra dinamicii rețelelor neuronale, iar TE surprinde unic fluxul direcționat și temporal de informații [67]. Studiul nostru inițiază utilizarea TE pentru a cuantifica transferul de informații între straturile neuronale, dezvăluind potențialul său de a îmbunătăți eficiența antrenării și de a elucida relația compresie-generalizare. Măsurând TE între straturile adiacente, observăm un model de potrivire-compresie similar principiului informației constrânse (IB), cu TE care atinge valori maxime la început, și scade pe măsură ce avansează antrenarea. Această tendință susține ipoteza că primele epoci se concentrează pe ajustare, urmate de o fază de compresie în care rețeaua rafinează și reține caracteristicile generice. Natura dinamică a TE, ca metrică, în special sensibilitatea sa la arhitectura rețelei și eficiența acesteia, se aliniază cu ideea că arhitecturile optimizate facilitează o compresie mai bună [22, 29].

Pe baza experimentelor, am folosit rețele feedforward puțin adânci și CNN optimizate pentru diferite seturi de date, inclusiv *glass*, *ionosphere*, *seeds*, *divorce*, *liver* și *iris* din UCI, alături de [103], [16], [70] și [38] pentru CNN ([65, 66]). TE a fost calculată pentru toate straturile și neuronale adiacente, cu primele 5% iterații din fiecare epocă excluse, pentru a stabiliza activările neuronale. Fracționarea dinamică, bazata pe selecția valorii maxime dintr-un procent de 95% din valorile de activare, a asigurat relevanța TE pe parcursul întregului antrenament. Tendința observată de scădere a TE pe parcursul epocilor și confirmarea unui TE mai mare în straturile finale pentru rețelele puțin adânci confirmă existența fazelor de compresie identificate în analizele IB [79, 89]. În CNN, concentrându-ne asupra ultimelor două straturi complet conectate (inclusiv softmax), am replicat aceste observații, examinând evoluția valorilor TE pentru seturi mai mari de date. Recomandăm observarea evoluției antrenării descrise în diagramele prezentate din versiunea completă a acestei teze.

Corelația directă dintre metricile performanței rețelei - acuratețea și eroarea de clasificare - dar și fluctuațiile TE, subliniază potențialul TE ca un instrument diagnostic. În faza de ajustare, TE scade rapid, reflectând reducerea erorilor și in raport invers cu acuratețea, înainte de a se stabili într-o fază de compresie cu variații minime. Această evoluție, este similară pe mai multe seturi de date și arhitecturi, susținând ideea că TE reflectă dinamica învățării rețelei și capacitățile de compresie. În arhitecturile eficiente, TE prezintă linii mai netede, cu modele de evoluție distincte, sugerând că structurile optimizate obțin o compresie interstraturi mai bună [22, 29].

Sensibilitatea TE la arhitectura rețelei și evoluția sa pe parcursul antrenamentului oferă ajustări subtile ale procesului de antrenare, dincolo de acurateţ̦e și numărul de parametri. Se manifestă ca un parametru adaptativ, aratând potențial în a reduce numărul de epoci de antrenare necesare. Cu toate acestea, sarcina computațională asociată calculului TE, în special pentru seturi mai mari de date și rețele mai adânci, necesită selectarea strategică a straturilor pentru analiză. Rezultatele observate în utilitatea TE în diagnosticarea problemelor din timpul antrenării modelelor și relația inversă cu eroarea și acuratețea se aliniază cu cadrul teoretic IB, validând TE ca o alternativă viabilă pentru analiza planurilor de informație.

În concluzie, studiul nostru demonstrează o legătură puternică între evoluția TE și metricile performanței rețelei, sugerând rolul său în diagnosticarea dinamicii antrenării și optimizarea compresiei. Deși TE completează principiul IB prin oferirea unei perspective specifice straturilor rețelelor, integrarea sa practică pentru îmbunătățirea antrenării sau ca metrică de diagnoză necesită o atenție riguroasă la costurile computaționale. Cercetările viitoare ar putea explora potențialul TE în ghidarea proiectării arhitecturilor rețelelor și rolul său în dezvoltarea modelelor neuronale mai interpretabile și mai eficiente. Tendințele observate în TE, alături de acurateţ̦e și pierdere, oferă o înțelegere mai bogată a modului în care rețelele neuronale învață și arhivează informația, oferă un potențial informativ pentru dezvoltarea algoritmilor de antrenare și a funcțiilor de cost mai robuste [2, 49].

CHAPTER 7

Entropia de transfer în rețelele convoluționale pe grafuri

Lucrarea noastră [68] studiază rețelele convoluționale pe grafuri (GCN) dintr-o perspectivă practică, concentrându-se asupra performanței de generalizare și abordând provocări precum netezirea excesivă și heterofilia. Propunem metoda TE-GGCN, care integrează TE ca un mecanism de control post-convoluție pentru a îmbunătăți discriminarea caracteristicilor nodurilor și acuratețea clasificării. În contrast cu îmbunătățirile GCN omogene, TE-GGCN selectează nodurile cu o heterofilie ridicată și grad, calculând TE pentru a ajusta caracteristicile lor, astfel încât să stimuleze capacitățile discriminative fără a modifica procesul convoluțional. Această strategie, deși solicitantă din punct de vedere computațional, în special pentru nodurile de grad ridicat, demonstrează eficacitatea în atenuarea netezirii excesive și îmbunătățirea acurateții pe diferite modele GCN.

Metoda GGCN [108] recalibrează ponderile muchiilor bazându-se pe gradul nodurilor și ajustează caracteristicile muchiilor pentru relațiile heterofile și homofile. TE-GGCN noastru adaugă peste GGCN calcularea ratelor de heterofilie a nodurilor (folosind $\mathcal{H}v = \frac{1}{|N(v)|}\sum u \in N(v)1\left(l_u \neq l_v\right)$) și selectarea celor mai bune 5% noduri heterofile, din care, ulterior, alegem 10% din nodurile cu cel mai ridicat grad. TE este calculată pentru aceste noduri folosind ecuația obișnuită a TE, modificând actualizările ponderilor folosind $\mathbf{H}i, j = \mathbf{H}i, j + max(TE_{Y_j \to X_i})$ post-convoluție. Această abordare selectivă asigură fezabilitatea computațională în timp ce maximizează creșterile de acuratețe, deoarece valorile TE amplifică precizia clasificării pentru nodurile care își schimbă clasa prezisă (clasificare greșită).

În experimentele noastre, TE-GGCN a fost evaluată pe un set variat de seturi de date reale și sintetice ale rețelelor de citații din lucrări științifice, prezentând o gamă diversă de niveluri de homofilie și heterofilie. Implementarea noastră, bazată pe PyTorch și Torch Geometric, a atins o acuratețe competitivă sau superioară comparativ cu modelul GGCN

original [108], în special pe seturile de date cu o homofilie scăzută precum Texas, Wisconsin și Cornell. Cu toate acestea, sarcina computațională a variat semnificativ, cu PubMed și Squirrel necesitând până la de cinci ori mai mult timp de antrenament din cauza nodurilor selectate având un grad ridicat. Calcularea TE în interiorul fiecărei straturi convoluționale a oferit o acurateţe mai mare, dar era nepractică din cauza costurilor computaționale foarte mari. Pentru rezultatele complete ale acurateții de validare, recomandăm examinarea versiunii complete a acestei teze.

Performanța TE-GGCN se bazează pe capacitatea sa de a identifica și corecta variațiile ridicate ale nodurilor, aplicând cea mai mare valoare calculată TE ca o ajustare a caracteristicilor post-convoluție. Această metodă, în concordanță cu cercetările noastre anterioare [66, 67], îmbunătățește modelele GCN existente fără a necesita modificări complexe. Compromisul între acurateţe și sarcina computațională este gestionabil, oferind o cale practică pentru îmbunătățirea performanței GCN în sarcinile de clasificare.

În concluzie, TE-GGCN demonstrează potențialul TE ca o metrică sensibilă pentru identificarea modelărilor similare de conectivitate și forma distribuțiilor a valorilor din noduri. Integrând valorile TE alături de metricile de heterofilie și grad, rafinăm capacitățile de clasificare ale GCN, în special pentru noduri susceptibile la erorile de clasificare. Această îmbunătățire simplă la modelele GCN, deși solicitantă din punct de vedere computațional pentru grafurile dense, oferă o metodă flexibilă pentru a stimula acurateţea fără a perturba mecanismele GCN bine cunoscute. Direcțiile viitoare ar putea explora optimizări computaționale pentru a extinde aplicabilitatea TE-GGCN asupra seturilor de date mai mari și mai complexe, păstrându-i beneficiile de performanță.

CHAPTER 8

Concluzii

Capitolul rezumă descoperirile cheie și contribuțiile cercetării noastre, publicate în mai multe studii. Teza noastră s-a axat pe îmbunătățirea algoritmilor de antrenare ai rețelelor neuronale, în special prin utilizarea feedback-ului bazat pe TE și aplicarea sa în diferite arhitecturi de rețele neuronale.

Am introdus algoritmul de antrenare neuronală *FF+FB*, care utilizează TE pentru a cuantifica relațiile dintre neuroni și o folosește ca feedback pentru a îmbunătăți anumite conexiuni neuronale. Metoda aceasta a demonstrat mai multe avantaje: în general, necesită mai puține epoci de antrenament în timp ce atinge o acuratețe mai mare comparativ cu rețelele feedforward (FF) standard; prezintă un comportament mai stabil în timpul procesului de antrenament; este mai puțin susceptibilă la minimul local. În plus, utilizarea feedback-ului TE a arătat potențialul de a reduce efortul necesar pentru optimizarea hiperparametrilor: parametrul prag g poate scădea importanța altor hiperparametri, cum ar fi rata de învățare η și numărul de neuroni din straturile ascunse. Acest demers poate facilita proiectarea și antrenarea mai ușoară a arhitecturii rețelei.

Urmând constatările *FF+FB*, cercetarea noastră a fost extinsă la Rețelele Neuronale Convoluționale (CNN), unde am constatat:

- Este eficient să se ia în considerare doar transferul de informații interneuronale al unui ansamblu aleatoriu de perechi de neuroni din cele două straturi finale, complet conectate
- Transferul de informații din aceste straturi are cel mai mare impact asupra procesului de învățare
- Multe conexiuni de transfer interneuronal par redundante, permițând utilizarea doar a unei fracții dintre ele

Am observat că TE acționează ca un factor de randomizare în modelele noastre, deoarece devine activ periodic, nu după ce fiecare eșantion de intrare este procesat. De

asemenea, TE poate fi considerat un meta-parametru care se schimbă lent, legat de ierarhia parametrilor care se schimbă rapid comparativ cu parametrii care se schimbă lent în modelele neuronale cauzale de învățare.

Cercetarea noastră s-a extins și la metoda Rețelei Convoluționale Grafic (GCN) prin algoritmul nostru *TE-GGCN*, unde:

- Am demonstrat îmbunătățiri prin exploatarea heterofiliei nodurilor, metricilor grafului și valorilor TE
- Am folosit TE ca măsură a variației ridicate a valorilor nodurilor, aplicând cea mai mare valoare TE calculată într-o trecere înainte(feedforward) ca ajustare a caracteristicilor nodului, post-convoluție
- Această corecție bazată pe TE, aplicată înaintea stratului de clasificare softmax, oferă o modalitate flexibilă și ușoară de a îmbunătăți implementările GCN existente

Abordarea noastră ar putea facilita extragerea cunoștințelor și explicațiilor din rețelele antrenate folosind paradigma cauzalității, deși acest lucru rămâne o problemă deschisă pentru cercetarea viitoare. Observațiile referitoare la transferul de informații în rețelele neuronale, în special în GCN, pot fi discutate în continuare dintr-o perspectivă neuroștiințifică, trasând paralelisme cu structurile din creierul vertebratelor. Constatările noastre se aliniază cu speculațiile precum că evaluarea relevanței diferitelor căi feedforward ar fi putut fi un mecanism phylo- sau ontogenetic pentru proiectarea structurilor de feedback din sistemele neuronale reale.

Chiar dacă feedback-ul TE accelerează procesul de antrenare prin reducerea numărului de epoci necesare, adaugă o sarcină computațională suplimentară fiecărei epoci. Echilibrul optimal între acești factori este dependent de aplicație. Adăugarea TE în mecanismul de învățare generează noi hiperparametri, ridicând întrebări despre posibilitatea supraantrenării și performanța la generalizare. Cu toate acestea, experimentele noastre sugerează că rolul TE ca un meta-parametru care se schimbă lent poate atenua aceste preocupări. Deși lucrarea noastră s-a concentrat în principal pe TE, cercetarea viitoare ar putea explora funcțiile de cost alternative, deoarece unii cercetători au sugerat că funcțiile de eroare bazate pe metoda informației constrânse (IB) nu se comportă întotdeauna optimal.

References

[1] S. Abu-El-Haija, B. Perozzi, A. Kapoor, N. Alipourfard, K. Lerman, H. Harutyunyan, G. V. Steeg, and A. Galstyan. MixHop: Higher-order graph convolutional architectures via sparsified neighborhood mixing. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 21–29. PMLR, 09–15 Jun 2019.

[2] R. Amjad and B. C. Geiger. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(09):2225–2239, sep 2020.

[3] R. Ando and T. Zhang. Learning on graph with laplacian regularization. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

[4] S. Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972.

[5] J. Bastings, I. Titov, W. Aziz, D. Marcheggiani, and K. Sima'an. Graph convolutional encoders for syntax-aware neural machine translation. *arXiv preprint arXiv:1704.04675*, 2017.

[6] R. v. d. Berg, T. N. Kipf, and M. Welling. Graph convolutional matrix completion. *arXiv preprint arXiv:1706.02263*, 2017.

[7] R. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473, 1972.

[8] C. Bodnar, F. Di Giovanni, B. Chamberlain, P. Lió, and M. Bronstein. Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in gnns. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 18527–18541. Curran Associates, Inc., 2022.

[9] J. Bruna, W. Zaremba, A. Szlam, and Y. Lecun. Spectral networks and locally connected networks on graphs. In *International Conference on Learning Representations (ICLR2014), CBLS, April 2014*, 2014.

[10] A. Caţaron and R. Andonie. Transfer information energy: A quantitative indicator of information transfer between time series. *Entropy*, 20(5), 2018.

[11] D. Chen, Y. Lin, W. Li, P. Li, J. Zhou, and X. Sun. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view, 2019.

[12] M. Chen, Z. Wei, Z. Huang, B. Ding, and Y. Li. Simple and deep graph convolutional networks. In H. D. III and A. Singh, editors, *37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1725–1735. PMLR, 13–18 Jul 2020.

[13] Z. Chen, X. Li, and J. Bruna. Supervised community detection with line graph neural networks. *arXiv preprint arXiv:1705.08415*, 2017.

[14] H. Cheng, D. Lian, S. Gao, and Y. Geng. Utilizing information bottleneck to evaluate the capability of deep neural networks for image classification. *Entropy*, 21(5), 2019.

[15] W.-L. Chiang, X. Liu, S. Si, Y. Li, S. Bengio, and C.-J. Hsieh. Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 257–266, New York, NY, USA, 2019. Association for Computing Machinery.

[16] A. Coates, A. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. *Journal of Machine Learning Research - Proceedings Track*, 15:215–223, 01 2011.

[17] M. Defferrard, X. Bresson, and P. Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[18] A. G. Dimitrov and J. P. Miller. Neural coding and decoding: communication channels and quantization. *Network: Computation in Neural Systems*, 12(4):441, aug 2001.

[19] D. Dua and C. Graff. UCI machine learning repository, 2017.

[20] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28, 2015.

[21] A. Eetemadi and I. Tagkopoulos. Genetic Neural Networks: an artificial neural network architecture for capturing gene expression relationships. *Bioinformatics*, 35(13):2226–2234, 11 2018.

[22] A. Elad, D. Haviv, Y. Blau, and T. Michaeli. Direct validation of the information bottleneck principle for deep nets. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 758–762, 2019.

[23] W. Equitz and T. Cover. Successive refinement of information. *IEEE Transactions on Information Theory*, 37(2):269–275, 1991.

[24] H. Fang, V. Wang, and M. Yamaguchi. Dissecting deep learning networks—visualizing mutual information. *Entropy*, 20(11), 2018.

[25] M. Federici, A. Dutta, P. Forré, N. Kushman, and Z. Akata. Learning robust representations via multi-view information bottleneck. *arXiv preprint arXiv:2002.07017*, 2020.

[26] D. Gencaga, K. H. Knuth, and W. B. Rossow. A recipe for the estimation of information flow in a dynamical system. *Entropy*, 17(1):438–470, 2015.

[27] R. Gilad-Bachrach, A. Navot, and N. Tishby. An information theoretic tradeoff between complexity and accuracy. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, pages 595–609. Springer, 2003.

[28] C. D. Gilbert and W. Li. Top-down influences on visual processing. *Nature Reviews Neuroscience*, 14(5):350–363, 2013.

[29] Z. Goldfeld, E. Van Den Berg, K. Greenewald, I. Melnyk, N. Nguyen, B. Kingsbury, and Y. Polyanskiy. Estimating information flow in deep neural networks. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2299–2308. PMLR, 09–15 Jun 2019.

[30] M. Gori, G. Monfardini, and F. Scarselli. A new model for learning in graph domains. In *Proceedings. 2005 IEEE international joint conference on neural networks, 2005.*, volume 2, pages 729–734. IEEE, 2005.

[31] H. Hafez-Kolahi and S. Kasaei. Information bottleneck and its applications in deep learning. *arXiv preprint arXiv:1904.03743*, 2019.

[32] W. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[33] M. Hanik, M. A. Demirtaş, M. A. Gharsallaoui, and I. Rekik. Predicting cognitive scores with graph neural networks through sample selection learning. *Brain Imaging and Behavior*, 16(3):1123–1138, Jun 2022.

[34] H. He, C. L. Yu, and Z. Goldfeld. Information-theoretic generalization bounds for deep neural networks. pages 1–25, 2024.

[35] S. Herzog, C. Tetzlaff, and F. Wörgötter. Transfer entropy-based feedback improves performance in artificial neural networks. *CoRR*, abs/1706.04265, 2017.

[36] S. Herzog, C. Tetzlaff, and F. Wörgötter. Evolving artificial neural networks with feedback. *Neural Networks*, 123:153 – 162, 2020.

[37] B. Huang and K. M. Carley. Syntax-aware aspect level sentiment classification with graph attention networks. *arXiv preprint arXiv:1909.02606*, 2019.

[38] J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994.

[39] K. M. Ikegwu, J. Trauger, J. McMullin, and R. J. Brunner. Pyif: A fast and light weight implementation to estimate bivariate transfer entropy for big data. In *SoutheastCon*, pages 1–6, 2020.

[40] D. D. Johnson. Learning graphical state transitions. In *International conference on learning representations*, 2022.

[41] J. Johnson, A. Gupta, and L. Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018.

[42] A. Kaiser and T. Schreiber. Information transfer in continuous processes. *Physica D: Nonlinear Phenomena*, 166(1):43 – 62, 2002.

[43] K. Kawaguchi, Z. Deng, X. Ji, and J. Huang. How does information bottleneck help deep learning? In *International Conference on Machine Learning*, pages 16049–16096. PMLR, 2023.

[44] J. Kawahara, C. J. Brown, S. P. Miller, B. G. Booth, V. Chau, R. E. Grunau, J. G. Zwicker, and G. Hamarneh. Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049, 2017.

[45] N. R. Ke, O. Bilaniuk, A. Goyal, S. Bauer, H. Larochelle, C. Pal, and Y. Bengio. Learning neural causal models from unknown interventions, 2019.

[46] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[47] T. N. Kipf and M. Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

[48] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[49] A. Kolchinsky, B. D. Tracey, and S. V. Kuyk. Caveats for information bottleneck in deterministic scenarios. In *International Conference on Learning Representations*, 2019.

[50] D. Kreuzer, D. Beaini, W. Hamilton, V. Létourneau, and P. Tossou. Rethinking graph transformers with spectral attention. *Advances in Neural Information Processing Systems*, 34:21618–21629, 2021.

[51] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. Technical report, 2009.

[52] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951.

[53] K. Kumar and V. Prasad. Few generalized entropic relations related to rydberg atoms. *Scientific Reports*, 12(1):7496, May 2022.

[54] R. Levie, F. Monti, X. Bresson, and M. M. Bronstein. Cayleynets: Graph convolutional neural networks with complex rational spectral filters. *IEEE Transactions on Signal Processing*, 67(1):97–109, 2018.

[55] G. Li, M. Müller, A. Thabet, and B. Ghanem. Deepgcns: Can gcns go as deep as cnns? In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9266–9275, 2019.

[56] Q. Li, Z. Han, and X.-M. Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[57] R. Li, S. Wang, F. Zhu, and J. Huang. Adaptive graph convolutional neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[58] Y. Li, R. Yu, C. Shahabi, and Y. Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.

[59] J. T. Lizier. Jidt: An information-theoretic toolkit for studying the dynamics of complex systems. *Frontiers in Robotics and AI*, 1, Dec. 2014.

[60] J. T. Lizier, J. Heinzle, A. Horstmann, J.-D. Haynes, and M. Prokopenko. Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fmri connectivity. *Journal of Computational Neuroscience*, 30(1):85–107, Feb 2011.

[61] J. T. Lizier and M. Prokopenko. Differentiating information transfer and causal effect. *The European Physical Journal B*, 73:605–615, 2010.

[62] S. Luan, C. Hua, Q. Lu, J. Zhu, M. Zhao, S. Zhang, X.-W. Chang, and D. Precup. Is heterophily a real nightmare for graph neural networks to do node classification?, 2021.

[63] S. Luan, C. Hua, Q. Lu, J. Zhu, M. Zhao, S. Zhang, X.-W. Chang, and D. Precup. Revisiting heterophily for graph neural networks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 1362–1375. Curran Associates, Inc., 2022.

[64] Y. Ma, X. Liu, N. Shah, and J. Tang. Is homophily a necessity for graph neural networks? *arXiv preprint arXiv:2106.06134*, 2021.

[65] A. Moldovan, A. Caţaron, and R. Andonie. Learning in feedforward neural networks accelerated by transfer entropy. *Entropy*, 22(1):102, 2020.

[66] A. Moldovan, A. Caţaron, and R. Andonie. Learning in convolutional neural networks accelerated by transfer entropy. *Entropy*, 23(9), 2021.

[67] A. Moldovan, A. Caţaron, and R. Andonie. Information plane analysis visualization in deep learning via transfer entropy. In *2023 27th International Conference Information Visualisation (IV)*, pages 278–285, 2023.

[68] A. Moldovan, A. Caţaron, and R. Andonie. Transfer entropy in graph convolutional neural networks. In *2024 28th International Conference Information Visualisation (IV)*, pages 278–285, 2024.

[69] F. Monti, M. Bronstein, and X. Bresson. Geometric matrix completion with recurrent multi-graph neural networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[70] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Ng. Reading digits in natural images with unsupervised feature learning. *NIPS*, 01 2011.

[71] O. Obst, J. Boedecker, and M. Asada. Improving recurrent neural network performance using transfer entropy. In *Proceedings of the 17th International Conference on Neural Information Processing: Models and Applications - Volume Part II*, ICONIP'10, pages 193–200, Berlin, Heidelberg, 2010. Springer-Verlag.

[72] K. Oono and T. Suzuki. Graph neural networks exponentially lose expressive power for node classification. *ICLR2020*, 8, 2020.

[73] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford infolab, 1999.

[74] G. Panagopoulos and F. D. Malliaros. Influence learning and maximization. In *International Conference on Web Engineering*, pages 547–550. Springer, 2021.

[75] O. Platonov, D. Kuznedelev, M. Diskin, A. Babenko, and L. Prokhorenkova. A critical look at the evaluation of GNNs under heterophily: are we really making progress?, 2023.

[76] Y. Rong, W. Huang, T. Xu, and J. Huang. Dropedge: Towards deep graph convolutional networks on node classification. *arXiv preprint arXiv:1907.10903*, 2019.

[77] T. K. Rusch, M. M. Bronstein, and S. Mishra. A survey on oversmoothing in graph neural networks, 2023.

[78] K. Rusek, J. Suárez-Varela, P. Almasan, P. Barlet-Ros, and A. Cabellos-Aparicio. Routenet: Leveraging graph neural networks for network modeling and optimization in sdn. *IEEE Journal on Selected Areas in Communications*, 38(10):2260–2270, 2020.

[79] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020, dec 2019.

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

[81] T. Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85:461–464, Jul 2000.

[82] W. Shadish, T. Cook, and D. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin, 2001.

[83] O. Shamir, S. Sabato, and N. Tishby. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29-30):2696–2711, 2010.

[84] C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[85] C. E. Shannon. Coding theorems for a discrete source with a fidelity criterion (international convention record), vol. 7. *New York, NY, USA: Institute of Radio Engineers*, 1959.

[86] M. Shimono and J. M. Beggs. Functional clusters, hubs, and communities in the cortical microconnectome. *Cerebral cortex (New York, N.Y. : 1991)*, 25(10):3743–3757, Oct 2015. 25336598[pmid].

[87] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE signal processing magazine*, 30(3):83–98, 2013.

[88] R. Shwartz Ziv and Y. LeCun. To compress or not to compress—self-supervised learning and information theory: A review. *Entropy*, 26(3), 2024.

[89] R. Shwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. *CoRR*, abs/1703.00810, 2017.

[90] N. Slonim and N. Tishby. Agglomerative information bottleneck. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

[91] L. Song, Y. Zhang, Z. Wang, and D. Gildea. A graph-to-sequence model for amr-to-text generation. *arXiv preprint arXiv:1805.02473*, 2018.

[92] A. Sperduti and A. Starita. Supervised neural networks for the classification of structures. *IEEE transactions on neural networks*, 8(3):714–735, 1997.

[93] L. Spillmann, B. Dresp-Langley, and C.-H. Tseng. Beyond the classical receptive field: the effect of contextual stimuli. *Journal of Vision*, 15(9):7–7, 2015.

[94] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method, 1999.

[95] N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5, 2015.

[96] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.

[97] R. Vicente, M. Wibral, M. Lindner, and G. Pipa. Transfer entropy—a model-free measure of effective connectivity for the neurosciences. *Journal of Computational Neuroscience*, 30(1):45–67, Feb 2011.

[98] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt. Graph kernels. *The Journal of Machine Learning Research*, 11:1201–1242, 2010.

[99] H. Wang, J. Wang, J. Wang, M. Zhao, W. Zhang, F. Zhang, X. Xie, and M. Guo. Graphgan: Graph representation learning with generative adversarial nets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[100] J. Wang, Y. Guo, L. Yang, and Y. Wang. Understanding heterophily for graph neural networks. *arXiv preprint arXiv:2401.09125*, 2024.

[101] T. Wang, D. Jin, R. Wang, D. He, and Y. Huang. Powerful graph convolutional networks with adaptive propagation mechanism for homophily and heterophily. In *AAAI Conference on Artificial Intelligence*, volume 36, pages 4210–4218, Jun. 2022.

[102] J. Weston, F. Ratle, and R. Collobert. Deep learning via semi-supervised embedding. In *Proceedings of the 25th international conference on Machine learning*, pages 1168–1175, 2008.

[103] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[104] C. Xie, J. Zhou, S. Gong, J. Wan, J. Qian, S. Yu, Q. Xuan, and X. Yang. Pathmlp: Smooth path towards high-order homophily, 2023.

[105] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017.

[106] K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[107] Y. Yan, Y. Chen, H. Chen, M. Xu, M. Das, H. Yang, and H. Tong. From trainable negative depth to edge heterophily in graphs. *Advances in Neural Information Processing Systems*, 36, 2024.

[108] Y. Yan, M. Hashemi, K. Swersky, Y. Yang, and D. Koutra. Two sides of the same coin: Heterophily and oversmoothing in graph convolutional neural networks. In *IEEE International Conference on Data Mining (ICDM)*, pages 1287–1292, Los Alamitos, CA, USA, dec 2022. IEEE Computer Society.

[109] H. Yang, K. Ma, and J. Cheng. Rethinking graph regularization for graph neural networks. In *Proceedings of the aaai conference on artificial intelligence*, volume 35, pages 4573–4581, 2021.

[110] M. Yoon. Introduction to graph neural networks, 2022.

[111] J. You, R. Ying, X. Ren, W. L. Hamilton, and J. Leskovec. Graphrnn: Generating realistic graphs with deep auto-regressive models. In *International Conference on Machine Learning*, 2018.

[112] W. Yu, C. Zheng, W. Cheng, C. C. Aggarwal, D. Song, B. Zong, H. Chen, and W. Wang. Learning deep network representations with adversarially regularized autoencoders. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2663–2671, 2018.

[113] S. Yun, M. Jeong, R. Kim, J. Kang, and H. J. Kim. Graph transformer networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[114] C. Zhang, Q. Li, and D. Song. Aspect-based sentiment classification with aspect-specific graph convolutional networks. *arXiv preprint arXiv:1909.03477*, 2019.

[115] M. Zhang and Y. Chen. Link prediction based on graph neural networks. *Advances in neural information processing systems*, 31, 2018.

[116] K. Zhao, Q. Kang, Y. Song, R. She, S. Wang, and W. P. Tay. Graph neural convection-diffusion with heterophily, 2023.

[117] L. Zhao and L. Akoglu. Pairnorm: Tackling oversmoothing in gnns. *arXiv preprint arXiv:1909.12223*, 2019.

[118] E. Zheleva and L. Getoor. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th*

International Conference on World Wide Web, WWW '09, page 531–540, New York, NY, USA, 2009. Association for Computing Machinery.

[119] T. Zhou, L. Lü, and Y.-C. Zhang. Predicting missing links via local information. *The European Physical Journal B*, 71:623–630, 2009.

[120] J. Zhu, Y. Yan, L. Zhao, M. Heimann, L. Akoglu, and D. Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7793–7804. Curran Associates, Inc., 2020.

[121] X. Zhu, Z. Ghahramani, and J. D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003.