

Szegedi Tudományegyetem
Informatikai Intézet

OTDK-dolgozat

Balogh András

2022

Szegedi Tudományegyetem

Informatikai Intézet

**Robusztus modellvarrás: neurális
reprezentációk funkcionális hasonlósága a
pontosságon túl**

OTDK-dolgozat

Készítette:

Balogh András

programtervező informatikus MSc
szakos hallgató

Témavezető:

Dr. Jelasity Márk

egyetemi tanár

Szeged
2022

Tartalomjegyzék

Absztrakt	5
1. Bevezetés	6
2. Irodalmi áttekintés	9
2.1. A robusztusság-pontosság kompromisszum	9
2.2. Reprezentációs hasonlósági metrikák	10
2.3. Funkcionális hasonlóság	12
2.3.1. Reprezentáció-megfeleltetés	12
2.3.2. Modellvarrás	13
3. Kísérleti háttér	15
3.1. Jelölések	15
3.2. Modellvarrás	15
3.3. Tanítás, kiértékelés	16
3.4. Felhasznált modellek	18
4. Osztályozási pontosság modellvarrással	20
4.1. Reprezentációk funkcionális hasonlósága	20
4.2. A robusztusság instabilitása	23
5. Robusztus modellvarrás	26
5.1. A robusztus modellvarrás tulajdonságai	27
5.2. Reprezentációk robusztus funkcionális hasonlósága	29
5.3. Robusztusság a komponens modellek ellenpéldáin	31
5.4. A robusztus modellvarrás általánosítása	33
6. Diskusszió	37
7. Összefoglalás	39

Köszönetnyilvánítás	40
Irodalomjegyzék	41

Absztrakt

Napjainkban a neurális hálózatok a gépi tanulás legelterjedtebb modelljei, azonban magas teljesítményüket árnyalják a működésükkel kapcsolatos nyitott, aktívan kutatott kérdések. Ilyen kérdéskör a hálózatok támadhatósága, amely során pl. képosztályozó hálózatok bemenetéhez emberi szemmel érzékelhetetlen perturbációk hozzáadásával a modellek kimenetei drasztikusan megváltoztathatók, ezáltal a hálózatok félrevezethetők.

Az ilyen támadás ellen védett (ún. robusztus) és nem robusztus hálózatok közötti különbségek feltárása fontos probléma, mivel ezzel felfedhető, hogy a támadások a hálózaton belül hol fejtik ki a hatásukat. Ezekből lehet következtetni a nem robusztus hálózatok gyenge pontjaira, amelyek ismerete elősegítheti robusztusabb hálózatok előállítását. A hálózatok összehasonlításának egyik meghatározó módszere a hálózatokat alkotó rétegek és az általuk kinyert reprezentációk elemzése. A dolgozat célja a robusztus és nem robusztus hálózatok reprezentációi közötti funkcionális hasonlóságok és eltérések vizsgálata.

Reprezentációk hasonlóságának vizsgálatára több módszer is ismert, a dolgozatban ezek közül a modellvarrást használom. Ennek segítségével megmutatom, hogy a robusztusság a reprezentációk funkcionális tulajdonsága, ezáltal a reprezentációk funkcionális hasonlóságát szükséges az osztályozási pontosság mellett a robusztusság szempontjából is vizsgálni. A robusztus funkcionális hasonlóság pontosságtól elkülönített vizsgálatának céljából bevezetem a robusztus modellvarrás módszerét.

A dolgozatban standard és robusztus modellvarrás használatával megmutatom, hogy robusztus és nem robusztus reprezentációk az osztályozási pontosság szempontjából funkcionálisan hasonlóak, azonban a robusztusság szempontjából a funkcionális eltérések a bemenethez közeli rétegekben keresendők. Az eredményeim továbbá mutatják, hogy robusztus reprezentációk a pontosság megtartása mellett a robusztusságot megőrző és elvesztő módon egyaránt varrhatók. Ezen felül megállapítom, hogy robusztus és nem robusztus hálózatok az osztályozási feladatot azonosan, a reprezentációik klaszterezésével oldják meg.

1. fejezet

Bevezetés

Napjainkban széles körben elterjedt a gépi tanulási modellek alkalmazása különböző feladatok megoldására [1, 2]. Ezen modellek közül a mesterséges neurális hálózatoknak köszönhető a közelmúlt legnagyobb áttörései a számítógépes látás [3, 4] és a természetes nyelvfeldolgozás [5] területein, valamint további, tudományterületeket átívelő (pl. orvosi) alkalmazásokban [6]. A neurális hálózatok magas teljesítményét azonban árnyalják a működésükkel és tanításukkal kapcsolatos nyitott, aktívan kutatott kérdések [7, 8, 9, 10].

Az egyik legfontosabb ilyen kérdéskör, különösen modern, biztonságkritikus alkalmazásokban a modellek sebezhetősége és támadhatósága [11, 12, 13, 14]. A közelmúltban megmutatták, hogy neurális hálózatokhoz generálhatók olyan ellenséges bemenetek, amelyekre a háló magas konfidenciával ad hibás eredményt [15, 16, 17]. Ez a jelenség a képfeldolgozási feladatoknál a legszembetűnőbb, ahol a bemenet minimális, ember számára érzékelhetetlen módosításával a modell kimenete drasztikusan megváltoztatható [18, 19].

Több közelmúltbeli tudományos eredmény is megmutatta, hogy lehetséges a támadás elleni védekezést megvalósító, ún. robusztus hálózatok előállítása, amennyiben a támadás során a bemenetet módosító perturbáció korlátozott egy megfelelően megválasztott, alacsony ϵ küszöbvel valamely ℓ_p normában [20, 21, 22, 23, 24]. A robusztus hálózatokról azonban közismert, hogy a teljesítményük nem éri el a nem robusztus (innenről: normál) versenytársaik szintjét, ezáltal a mai napig fennáll egy kompromisszum (trade-off) a hálózatok teljesítménye (pontossága) és támadhatósága (robusztussága) között [25, 26, 27]. A kompromisszum okának feltárása [28, 29, 30], illetve a jelenség áthidalása [31], valamint a neurális hálózatok támadása [32, 33, 34] és a támadások elleni védekezési módszerek [35, 36, 37] egyaránt aktívan kutatott területek.

A robusztusság problémája mellett egy másik, hasonló fontosságú, mai napig megoldatlan feladat a neurális hálózatok működésének értelmezése és magyarázása [38, 39]. Ebben segíthet a modellt alkotó önálló rétegek, illetve egymást követő rétegek által al-

kotott részháló elemzése [40, 41, 42]. Szintén nyitott kérdés, hogy a mély neuronháló rejtett rétegei milyen funkciókat tanulnak meg, azaz hogyan reprezentálják a bemenetet, illetve ezt a reprezentációt a modell további rétegei hogyan használják fel a feladat megoldására [43, 44].

A reprezentációk elemzésére és hálózatok közötti összehasonlítására a kutatók az utóbbi években több reprezentációs hasonlósági metrikát is kidolgoztak [45, 46, 47]. Ezek közös tulajdonsága, hogy nem alkalmasak a reprezentációk funkcionális tulajdonságainak összehasonlítására [48], valamint értelmezésük nehézkes és nem egyértelmű [49, 50]. Ennek ellenére, egyszerűségüknek és előnyös matematikai tulajdonságaiknak köszönhetően a reprezentációs hasonlósági metrikákat igen gyakran alkalmazzák neurális hálózatok reprezentációinak vizsgálatára [51, 52, 53, 54].

Ezzel szemben a reprezentációk funkcionális hasonlóságát vizsgáló módszerek betekintést nyújtanak a reprezentációk felhasználásába is [55]. Ilyen módszerek a reprezentáció-megfeleltetés [56] és a modellvarrás [57], amelyekkel a hálózatok egyes részeinek kompatibilitása vizsgálható egy adott feladat megoldására.

A dolgozat célja egy új megközelítésben vizsgálni a robusztus és normál hálózatok által megtanult reprezentációk viszonyát a következő központi kérdés megválaszolásával:

Funkcionálisan mennyiben térnek el a normál és robusztus hálózatok reprezentációi?

A kérdés megválaszolására a modellvarrás módszerével vizsgálom normál és robusztus osztályozó modellek reprezentációinak kompatibilitását az osztályozási pontosságon felül az osztályozási robusztusság, mint a reprezentációk funkcionális tulajdonságának elemzésével. E két funkcionális tulajdonság vizsgálatára az osztályozási pontosságot előtérbe helyező, "standard" modellvarrás mellett bevezetem a robusztus modellvarrást, amellyel a reprezentációk kompatibilitása elsődlegesen az osztályozási robusztusság szempontjából vizsgálható. A dolgozatban továbbá tanulmányozom a reprezentációk robusztusságának és pontosságának függetlenségét, valamint az egyes tulajdonságok robusztusság-pontosság kompromisszumból adódó instabilitását, amennyiben a másik tulajdonság megtartása, illetve felerősítése a cél.

A dolgozat főbb eredményei a következők:

- Normál és robusztus hálózatok reprezentációi az osztályozási pontosság szempontjából funkcionálisan kompatibilisek.
- A normál konvolúciós hálózatok első rétegeinek reprezentációi önmagukban nem robusztusak, azonban az osztályozási robusztusság szempontjából funkcionálisan kompatibilisek a robusztus hálózatok reprezentációival. Ugyanez a normál hálózatok mélyebb rétegeire nem igaz.

- A reprezentációk robusztussága instabil, azaz az osztályozási pontosság előtérbe helyezésével robusztus hálózatok reprezentációi elveszítik a robusztusságukat. Ez az instabilitás a hálózatok különböző rétegeiben eltérő mértékű.
- Az osztályozási robusztusság megőrzéséhez, illetve a robusztusságot maximalizáló modell eléréséhez a teljes modellt robusztusan kell tanítani.

A dolgozatban ezen felül a normál és robusztus hálózatok, mint különböző eloszlású adatokon tanított modellek [58] összehasonlításának általánosításaként különböző adatbázisokon tanított osztályozó hálózatok reprezentációinak funkcionális hasonlóságát is vizsgálom. Ennek segítségével megállapítom, hogy különböző eloszlásokon tanított osztályozó hálózatok mélyebb rétegei, a reprezentációk klasztereződéséből [59] adódóan funkcionálisan kompatibilisek.

Az eredményeim bemutatására általam létrehozott, illetve state-of-the-art robusztus konvolúciós hálózatokat [60], valamint képfeldolgozási feladatra kialakított transformer hálózatokat használtam [61]. A dolgozatban az MNIST [62], Fashion MNIST [63], SVHN [64] és CIFAR10 [65] képosztályozási feladatokat megoldó hálózatokat vizsgáltam. A dolgozathoz kapcsolódó PyTorch [66] implementáció és a használt modellek publikusan elérhetők: `Erasiel/robust_similarity_and_matching`

2. fejezet

Irodalmi áttekintés

Neurális hálózatok reprezentációinak elemzésére több szakirodalmi eredmény is irányul, több különböző módszer használatával. Egy k kimeneti neuronnal rendelkező réteg által kinyert reprezentáció egy n elemű példahalmazon a réteg aktivációit tartalmazó $n \times k$ méretű mátrix, ennek megfelelően a reprezentációk vizsgálata gyakran egyszerűsíthető mátrixok vizsgálatára. Ezen módszerek a reprezentációk funkcionális elemzésére kevésbé alkalmasak, így a funkcionális tulajdonságok vizsgálatához más módszerekhez kell fordulni. A reprezentációk elemzésének módszereivel több kutatás is irányult a robusztus és normál hálózatok közötti különbség- és a robusztusság-pontosság kompromisszum vizsgálatára.

2.1. A robusztusság-pontosság kompromisszum

Az utóbbi években több tudományos publikáció jelent meg a képosztályozó neurális hálózatoknál tapasztalt robusztusság-pontosság kompromisszum okának feltárásával kapcsolatban. Tsipras és munkatársai [26] megmutatták, hogy a jelenség a robusztus modellek tanítási folyamatának következménye. Ezen folyamat során a tanítópéldákból az aktuális hálózatnak megfelelően ellenséges példákat állítunk elő, majd az így kapott, ellenségesen perturbált példákat használjuk tanításhoz (vö: adat-augmentáció) [20, 22]. A szerzők úgy vélekednek, hogy egy osztályozási feladaton akkor érhető el magas pontosság nem robusztus hálózatok által, ha a tanító adathalmaz tartalmaz kis számú, a modell által kihasználható, jól szeparáló és egyben jól általánosító jellemzőt. Mivel ezen jellemzőket a támadás módosíthatja, a modell osztályozási pontosságának növekedésével a *robusztus pontossága* (az ellenséges példákon vett osztályozási pontossága) csökken. Ezzel szemben, érvelésük szerint a robusztus hálózatok olyan reprezentációkat tanulnak meg, amelyek a nagyobb számú, kevésbé szeparáló, ugyanakkor kevésbé támadható jellemző használatának prioritizálásával jönnek létre. Következésképpen a robusztus neurális hálózatok

osztályozási pontossága alacsonyabb, ugyanakkor a standard tanítással kapott modellek nem lesznek robusztusak.

Tsipras és munkatársai [26] felvetették a kevésbé támadható (stabil, robusztus) jellemzők és a támadható (instabil, nem robusztus) jellemzők létezését. Erre a felvetésre építve Ilyas és munkatársai [29] megmutatták, hogy egy robusztusan tanított hálózat alapján előállítható a tanító adatbázis egy olyan módosítása, amelyben a példák csak a robusztus, stabil jellemzőket tartalmazzák. A szerzők az így kapott adathalmazon standard (nem robusztus) tanítással tudtak olyan modellt készíteni, amely magas osztályozási robusztusságot ér el az eredeti adatbázison. A publikációban ezen felül megmutatták, hogy a robusztus képosztályozó hálók "reprezentációs", azaz utolsó előtti rétegeinek kimenetei invertálhatók, vagyis belőlük egy optimalizálási módszerrel, előzetes információk használata nélkül az eredeti bemenethez emberi szemmel hasonló tartalmú képet kapunk, míg ugyanez a módszer nem robusztus hálók esetén emberileg értelmezhetetlen kimenetet ad.

A robusztus és normál hálózatok információtartalmával, illetve az utolsó előtti réteg kimenetének invertálhatóságával kapcsolatban Terzi és munkatársai [30] megállapították, hogy a robusztus hálózat ezen rétegeinek reprezentációja a modell súlyainak köszönhetően a teljes bemenetről tartalmaz információt, azonban az utolsó (osztályozó) réteg nem használja ki a reprezentáció egészét. Achille és munkatársai [67] korábban megmutatták, hogy a nem robusztus hálózatok súlyai is tartalmazhatnak információt a teljes bemenetről egészen addig, amíg az utolsó réteg csak az osztályozáshoz szükséges információhoz fér hozzá. Terzi és munkatársai kiemelik továbbá, hogy a robusztus hálózatok reprezentációi hatékonyabban újrafelhasználhatók (transfer learning), ezáltal egy funkcionális különbséget is bemutatnak normál és robusztus hálózatok reprezentációi között.

A dolgozatban a reprezentációk elemzését kiterjesztem a hálózat összes rétegeire, nem csupán az utolsó előtti réteget vizsgálom. Megmutatom, hogy a hálózatok robusztussága nagyban függ az első, "előfeldolgozó" rétegektől, illetve a robusztus és normál hálózatok reprezentációi közötti főbb különbségek ezekben a rétegekben keresendők, ugyanakkor a hálózat teljes robusztusságához a mélyebb, "szemantikai" rétegek sem elhanyagolhatók. Ezen felül megmutatom, hogy a stabil jellemzőket kihasználó robusztus reprezentációk egy része instabil, azaz a reprezentációk robusztussága részben instabil.

2.2. Reprezentációs hasonlósági metrikák

Neurális hálózatok egyes rétegeinek reprezentációi az adott réteg aktivációi (kimenete) meghatározott bemeneteken. A reprezentációs hasonlósági metrikák két modell egy-egy rétege által kinyert reprezentációk összehasonlítását teszik lehetővé. Több reprezentációs hasonlósági metrikát is kidolgoztak az utóbbi években, ezek közül a Centered Kernel

Alignment (CKA) [47] és a Canonical Correlation Analysis (CCA) két variánsának [45, 46] használata terjedt el leginkább.

A reprezentációs hasonlósági metrikákat Cianfrani és munkatársai [53] alkalmazták először a robusztus és normál neurális hálózatok összehasonlítására, illetve a robusztus tanítás reprezentációkra gyakorolt hatásának elemzésére. A CKA metrika használatával megmutatták, hogy a normál hálózatokban megfigyelhetők "specializálódó" rétegcsoportok, azaz olyan egymást követő rétegek, amelyek egymáshoz viszonyított reprezentációs hasonlósága magas, míg robusztus modelleknél ez a jelenség nem látszik. Ezen felül megállapítják, hogy normál és robusztus modellek első rétegeinek reprezentációi tiszta és ellenséges példákön hasonló reprezentációkat mutatnak, tehát a korai rétegek reprezentációira a támadás kisebb hatást gyakorol. Végül megmutatják, hogy robusztus és normál hálózatok által megtanult reprezentációk nagyban eltérnek, mivel a reprezentációs hasonlóságuk alacsony.

A dolgozatban ezzel szemben megmutatom, hogy a normál és robusztus reprezentációk az alacsony hasonlóságuk ellenére funkcionálisan kompatibilisek az osztályozás pontossága szempontjából. A normál hálózatok korai rétegeinek reprezentációi ezen felül az osztályozási robusztusság szempontjából is funkcionálisan kompatibilisek a robusztus reprezentációkkal, azonban ezen reprezentációk önmagukban nem robusztusak. Továbbá megmutatom Terzi és munkatársai [30] és Cianfrani és munkatársai [53] eredményeinek kiegészítéseként, hogy a robusztus és normál hálózatok rétegeinek információtartalmi- és reprezentációs különbségei ellenére a reprezentációk hasonló geometriai tulajdonságokkal rendelkeznek.

A dolgozat eredményei részben ellentmondanak a reprezentációs hasonlósági metrikák használatával feltárt összefüggésekkel. Ennek oka a reprezentációs hasonlóság és a dolgozatban vizsgált funkcionális hasonlóság közötti különbség. A funkcionális vizsgálat betekintést nyújt a reprezentációk egy adott feladat megoldásához történő felhasználásába, míg a reprezentációs hasonlóság erről csupán közvetetten tartalmaz információt. Ding és munkatársai [48] megmutatták, hogy a reprezentációs hasonlósági metrikák nem alkalmasak a reprezentációk funkcionális tulajdonságainak megbízható elemzésére, mivel a metrikák túl érzékenyek a hálózat funkcióját nem befolyásoló, kisebb eltérésre, illetve zajra, vagy nem elég érzékenyek a modell funkcióját nagyban befolyásoló módosításokra. A CKA gyenge érzékenységet Mirzadeh és munkatársai [49] is bemutatták egy folyamatos tanulási (continual learning [68]) feladaton, amely során a modell rétegei közötti reprezentációs hasonlóság konstans, miközben a tanítási folyamattal a modell teljesítménye a korábbi feladatokon romlik.

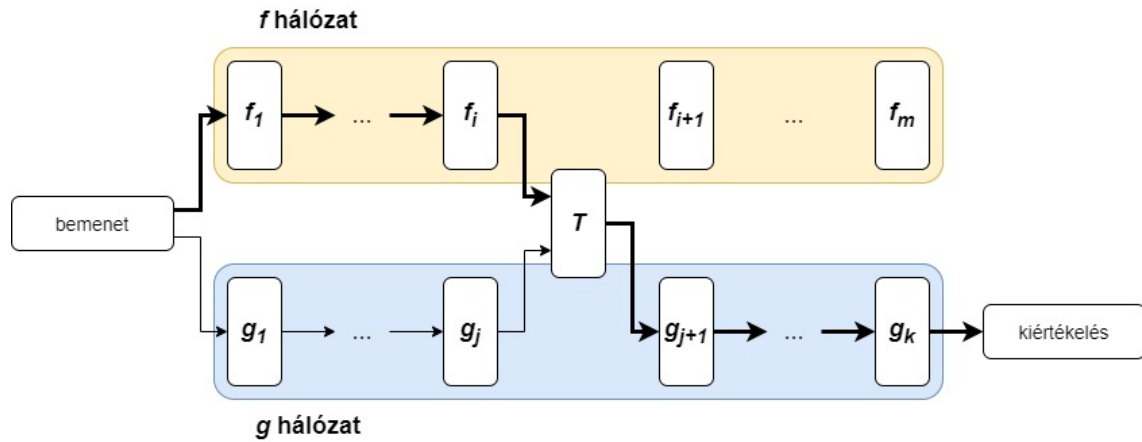
2.3. Funkcionális hasonlóság

A reprezentációk funkcionális tulajdonságainak vizsgálata betekintést nyújt a reprezentációk felhasználásába egy konkrét feladat megoldására. A funkcionális hasonlóság vizsgálatának alapkérdése a következő: Meg tudja oldani egy neurális hálózat a feladatát egy másik neurális hálózat reprezentációjával? A kérdés megválaszolására két, hasonló eszközt javasoltak: a reprezentáció-megfeleltetést és a modellvarrást. Ezen módszerek célja egy f neurális hálózat egy rögzített rétege által kinyert reprezentáció transzformálása úgy, hogy az egy g neurális hálózat egy rögzített rétege által kinyert reprezentációnak feleljen meg, ezáltal g további része meg tudja oldani a feladatát a transzformált reprezentációval, minimális teljesítmény-csökkenés mellett [56, 57]. A hasonlóság vizsgálatához a transzformációnak kellően egyszerűnek kell lennie [55], ugyanis elegendően komplex függvénnyel tetszőleges megfeleltetés megvalósítható (lásd: univerzális approximáció [69]). A szakirodalomban a reprezentáció-megfeleltetés és a modellvarrás egyaránt lineáris transzformációt használ a reprezentációk illesztésére.

2.3.1. Reprezentáció-megfeleltetés

A reprezentáció-megfeleltetés módszere egy rögzített példahalmazon határozza meg az f és g hálók adott rétegeinek reprezentációit és csak ezeket használja az illesztés megvalósításához (lásd: 2.2 ábra). A megfeleltetést megvalósító transzformációt Li és munkatársai [56] korreláción- és kölcsönös információ alapultó módszerekkel, míg Wang és munkatársai [70] a megadott rétegek aktivációs tereit illesztő permutációval határozzák meg. Módszereik meglepően sikertelenek, ugyanis a megfeleltetett reprezentáció használatával kiértékelt osztályozó hálózatok pontossága szignifikánsan csökken az eredeti teljesítményükhöz képest. Ennek következményeként kijelenthető, hogy reprezentációk permutációjával a megfeleltetés nem valósítható meg szignifikáns teljesítmény-csökkenés nélkül.

Csiszárík és munkatársai [55] megmutatták, hogy az f hálózat kijelölt rétege által kinyert reprezentáció pszeudoinverzének [71] használatával optimális illesztést kapunk. Módszerükkel a megfeleltetett reprezentációt használó g hálózat teljesítménye továbbra sem éri el az eredeti hálózatét, azonban a teljesítmény-csökkenés a korábbi módszerekhez képest sokkal enyhébb. Módszerüket Jones és munkatársai [72] használták robusztus és normál osztályozó hálózatok utolsó előtti rétegeinek illesztésére és megállapították, hogy a reprezentációk megfeleltetésével ki lehet javítani az eredeti hálózatok predikciói közötti eltéréseket. Ez a javíthatóság a hálózatok robusztusságával növekszik, ami szintén arra utal, hogy a robusztus hálózatok utolsó előtti rétegének reprezentációja gazdagabb, mint normál hálózatoké.



2.1. ábra. A reprezentáció-megfeleltetés sematikus ábrája. T a megfeleltetést végző transzformáció, amely az illesztést az f és g hálók megadott rétegeinek reprezentációja alapján végzi. Vastaggal jelezzük a bemenet feldolgozásának útját. Az ábra forrása: [55].

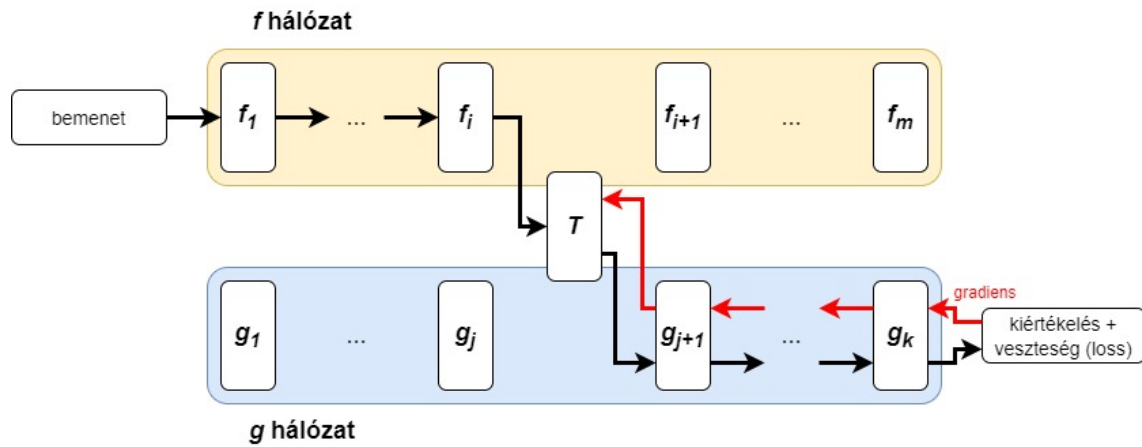
A dolgozatban a reprezentáció-megfeleltetés helyett a modellvarrást alkalmazom a reprezentációk funkcionális tulajdonságainak vizsgálatára (lásd: 2.3.2 fejezet). Csiszárík és munkatársai [55] megmutatták és az előzetes kísérleteim is igazolták, hogy modellvarrással tovább csökkenthető a megfeleltetéssel tapasztalt teljesítmény-csökkenés. Ezen felül a megfeleltetéssel a reprezentációk robusztussága csak közvetetten vizsgálható, az optimális illesztés nem kereshető robusztusan, csupán az egyes hálókhöz generált ellenséges példák reprezentációi alapján. A dolgozatban azonban megmutatom, hogy a reprezentációk robusztusságának vizsgálatához nem elegendő a két hálózat ellenséges példáinak használata, ennek megfelelően az ellenséges példákon történő reprezentáció-megfeleltetést sem vizsgálom.

2.3.2. Modellvarrás

A modellvarrás módszere annyiban különbözik a reprezentáció-megfeleltetéstől, hogy a g hálózat választott rétegének reprezentációját nem használjuk, azonban g további rétegeit igen, ezáltal a transzformáció egy megadott veszteségfüggvény segítségével tanulható (lásd: 2.2 ábra). A modellvarrást Lenc és Vedaldi [57] vezették be képosztályozó neurális hálózatok reprezentációinak vizsgálatára. A szerzők megállapították, hogy különböző architektúrájú, különböző eloszlásokon (adatbázisokon) tanított modellek korai rétegei funkcionálisan kompatibilisek, míg a mélyebb rétegek feladatspecifikusak, ezáltal kevésbé kompatibilisek. A dolgozatban megmutatjuk, hogy a későbbi rétegek reprezentációinak geometriai tulajdonságaiból (a reprezentációk klasztereződéséből) adódóan a mélyebb rétegek is funkcionálisan kompatibilisek a modellvarrás alapfeladatának megfe-

elő módosításával.

Modellvarrás használatával Csiszárík és munkatársai [55], valamint Bansal és munkatársai [50] egyaránt megállapították, hogy egyező architektúrájú, azonos eloszláson, de különböző inicializációval és hiperparaméterekkel tanított hálózatok megegyező rétegei funkcionálisan kompatibilisek. Bansal és munkatársai ezen felül megmutatják, hogy azonos osztályozási feladaton magas pontosságot elérő modellek funkcionálisan hasonló reprezentációkat tanulnak. Ehhez hasonlóan Mcneely-White és munkatársai [73] megmutatták, hogy Inception [74] és ResNet [75] hálózatok utolsó előtti rétegei funkcionálisan kompatibilisek.



2.2. ábra. A modellvarrás sematikus ábrája. T a varrást végző transzformáció, amely a pirossal jelölt gradienssel tanulható. Vastaggal jelezzük a bemenet feldolgozásának útját. Az ábra forrása: [55].

A dolgozatban a modellvarrást robusztus és normál hálózatok reprezentációinak funkcionális összehasonlítására használom. A robusztusságot a reprezentációk funkcionális tulajdonságaként vizsgálom, ennek megfelelően a robusztusság mélyebb megértése céljából bevezetem a robusztus modellvarrást. Ezen felül megmutatom, hogy magas teljesítményű osztályozó modellek mélyebb rétegeinek funkcionális kompatibilitása a reprezentációk klasztereződéséből adódóan triviális.

3. fejezet

Kísérleti háttér

3.1. Jelölések

A felügyelt tanulás feladata egy $f : X \rightarrow Y$ függvény megtanulása az $X \times Y$ feletti $\mathcal{D}_{X \times Y}$ eloszlásból származó $D = \{(x_l, y_l)\}_{l=1}^n$ (ahol $(x_l, y_l) \sim \mathcal{D}_{X \times Y}$) tanító adatbázis alapján úgy, hogy $\forall (x, y) \in D : f(x) = y$. A dolgozatban képosztályozó neurális hálózatokat vizsgálunk, azaz Y egy véges halmaz, amely elemei az egyes osztályok címkéi, X képek halmaza, illetve f egy neurális hálózat. Az f osztályozó által modellezett *a posteriori* eloszlást jelölje $\mathcal{D}_{Y|X}[f]$.

Egy $m \in \mathbb{N}$ rétegű f neurális hálózat a rétegeit alkotó függvények kombinációja: $f = f_m \circ \dots \circ f_1$, ahol $f_i : \mathcal{A}_{f,i-1} \rightarrow \mathcal{A}_{f,i}$ ($\mathcal{A}_{f,i}$ az i -edik réteg aktivációs tere, $\mathcal{A}_{f,0} = \mathcal{X}$). A reprezentációk elemzéséhez a hálózatot egy rögzített, l -edik rétege mentén kell vizsgálni, ehhez bevezetjük a következő jelöléseket: $f_{\leq l} = f_l \circ \dots \circ f_1$ és $f_{> l} = f_m \circ \dots \circ f_{l+1}$, értelemszerűen $f = f_{> l} \circ f_{\leq l}$. Az l -edik réteg által egy $x \in \mathcal{X}$ bemeneten kinyert reprezentáció $f_{\leq l}(x) \in \mathcal{A}_l$.

3.2. Modellvarrás

A funkcionális hasonlóság vizsgálatának alapkérdése a következő: meg tudja oldani egy g hálózat a feladatát egy f hálózat reprezentációjával? Formálisan, adott két neurális hálózat, $f = f_m \circ \dots \circ f_1$ és $g = g_k \circ \dots \circ g_1$ ($m, k \in \mathbb{N}$) és ezek egy-egy rögzített rétege, f_i és g_j ($1 \leq i < m$ és $1 \leq j < k$). A feladat egy olyan $T : \mathcal{A}_{f,i} \rightarrow \mathcal{A}_{g,j}$ transzformáció megadása, amivel $\mathcal{D}_{Y|X}[g_{>j} \circ T \circ f_{\leq i}] \approx \mathcal{D}_{Y|X}[g]$.

Modellvarrással a T transzformáció tanulható (lásd: 2.2 ábra), azaz a bemenet az f és g modellek, ezek egy-egy rétege (f_i és g_j), egy $\mathcal{D}_{X \times Y}$ eloszlás feletti $D = \{(x_l, y_l)\}_{l=1}^n$ tanító adathalmaz, valamint egy $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ veszteségfüggvény. A tanítás során

csupán a T transzformáció θ paramétereit tanuljuk, azaz a következő optimalizálási problémát oldjuk meg:

$$\arg \min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}_{X \times Y}} [\mathcal{L}(g_{>j} \circ T_{\theta} \circ f_{\leq i}(x), y)] \quad (3.1)$$

ahol a cél a veszteségfüggvény várható értékének minimalizálása a tanításhoz használt eloszlás felett. A modellvarrás alapfeladatában $\mathcal{D}_{X \times Y}$ a g hálózat feladatához tartozó eloszlás, azonban a dolgozat későbbi részében kiterjesztem a vizsgálatot az f hálózat feladatához tartozó eloszlásra. Ezzel a módosítással a feladat az f által megtanult *a posteriori* eloszlás modellezése modellvarrással, azaz olyan T transzformáció megtanulása, amelyre $\mathcal{D}_{Y|X}[g_{>j} \circ T \circ f_{\leq i}] \approx \mathcal{D}_{Y|X}[f]$.

Az f és g hálózatok paramétereit a tanítás során nem módosulnak, ezért a 3.1 képletből ezeket elhagytam. Továbbá, amennyiben egyértelmű, a modellvarrást végző transzformáció θ paramétereit is elhagyom és jelben T -ként hivatkozok rá. A dolgozatban a modellvarrással létrehozott $g_{>j} \circ T \circ f_{\leq i}$ modellre a varrás eredményeként kapott, röviden varrt modellként, míg a varrás alapját képező f és g modellekre a varrás komponens modelljeként hivatkozok. Ezen belül f a varrás első-, g a varrás hátsó modellje (vagy feje), avagy az f modell van elöl, g hátul a varrásban.

A modellek adott rétegeit összekötő T transzformáció a modellvarrás esetében lineáris. Ez konvolúciós rétegek esetén 1×1 -es konvolúciós kernelekből álló transzformáció, egyéb rétegek esetén a szokásos lineáris transzformáció, mindkét esetben aktivációs függvény és további komponensek (pl. batch normalizáció) nélkül.

A dolgozatban reziduális hálózatok varrását a reziduális blokkok végén hajtom végre, azaz nincs olyan eset, amikor a varrást a reziduális kapcsolat "átugorja". A dolgozatban a modellvarrást egyező architektúrájú hálózatok között hajtom végre. Kornblith és munkatársai [47] a CKA reprezentációs hasonlósági metrikával megmutatták, hogy egyező architektúrájú neurális hálózat rétegeinek reprezentációs hasonlósága az egyező rétegek között a legnagyobb, ennek megfelelően a dolgozatban a hálózatok azonos rétegeinek funkcionális kompatibilitását vizsgálom modellvarrással.

3.3. Tanítás, kiértékelés

A dolgozatban általam előállított normál- és robusztus neurális hálózatokat, valamint state-of-the-art robusztus modelleket használok, utóbbiakhoz a RobustBench [60] publikus ranglistán keresztül fértem hozzá. A dolgozatban VGG-11 [76], ResNet-18, ResNet-50 [75], PreActResNet-18 [77], WideResNet-28-10 [78] és Pooling Transformer (PViT) [61] architektúrájú hálózatokat vizsgáltam.

A normál modellek tanítását az sztochasztikus gradiens (SGD) optimalizálóval végez-

tem, 0.9-es Nesterov-momentummal, 10^{-4} ℓ_2 -regularizációs (weight decay) együtthatóval. A kezdeti tanulási ráta VGG-11 modellek tanításánál 10^{-3} , a többi modellnél 10^{-1} , ami a tanítás harmadánál és kétharmadánál a tizedére csökken. Az MNIST adatbázison tanított modelleket 100, míg a Fashion MNIST, CIFAR10 és SVHN adatbázisokon tanított modelleket 200 epochig tanítottam. A tanított hálózatok egyes rétegeinek paramétereit véletlenszerűen inicializáltuk a $[-\sqrt{1/p}, \sqrt{1/p}]$ intervallumból egyenletes eloszlással, ahol p az adott réteg bemenő paramétereinek száma. A modellek és a modellvarrás tanításánál adat-augmentációként vízszintes tükrözést és a képek egy részének véletlenszerű kivágását és az eredeti képméretre történő felnagyítását alkalmazom. A képek intenzitásértékeit a $[0, 1]$ intervallumba skáláztam. A hiperparaméterek értékei megegyeznek Csiszárík és munkatársai [55] publikációjában.

Az általam előállított robusztus modellek tanítási paramétereit hasonlóan állítottam be, azzal a különbséggel, hogy az Adam optimalizálót [79] használtam. Ennek egy oka, hogy Bansal és munkatársai [50] megmutatták, hogy SGD optimalizálóval tanított hálózatok rétegeinek reprezentációi modellvarrással funkcionálisan kompatibilisek, azonban SGD és Adam optimalizálókkal tanított hálózatokról hasonló kijelentésük nincs.

Robusztus tanításhoz a Madry-féle tanítási protokollt [22] és a TRADES [31] veszteségfüggvényen alapuló módszert használtam. A Madry-protokoll a θ paraméterekkel megadott f hálózat tanítása során a következő min-max problémát oldja meg:

$$\arg \min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}_{X \times Y}} [\max_{\delta \in \mathcal{S}} \mathcal{L}(f_{\theta}(x + \delta), y)] \quad (3.2)$$

A képletben $\mathcal{S} = \{\delta : \|\delta\|_p \leq \epsilon\}$ jelöli a megengedett ellenséges perturbációkat (ahol $\|\cdot\|_p$ az ℓ_p norma), azaz a bemenet módosításának lehetséges értékeit. A dolgozatban az ℓ_{∞} normán alapuló fenyegetési modellt (threat model) használjuk, amelyben a modell bemenetéhez dimenzióként abszolút értékben legfeljebb ϵ mértékű perturbáció adható. ϵ értéke a szakirodalomnak és a RobustBench [60] ranglistának megfelelően a CIFAR10 és SVHN adatbázisokon 8/255, az MNIST és Fashion MNIST adatbázisokon 0.3.

A Madry-protokoll alkalmazásával egy támadó jelenlétében minden x -hez előállítható olyan megengedett $\delta \in \mathcal{S}$ perturbáció, amely maximalizálja a $\mathcal{L}(f_{\theta}(x + \delta), y)$ veszteséget. Az algoritmus a Projected Gradient Descent (PGD) iteratív, gradiens-alapú támadással generálja az ellenséges példákat. A dolgozatban a Madry-protokollhoz használt PGD iterációszáma 7, lépésköze CIFAR10 és SVHN adatbázisokon 2/255, MNIST és Fashion MNIST adatbázisokon 0.1.

A robusztus tanítás megvalósítható ellenséges példák generálása nélkül is, a TRADES [31] módszer a veszteségfüggvényt módosítja, amellyel a teljes optimalizálási feladat a következő:

$$\arg \min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}_{X \times Y}} [\mathcal{L}(f_{\theta}(x), y) + \beta \max_{\delta \in \mathcal{S}} \mathcal{L}(f_{\theta}(x), f_{\theta}(x + \delta))] \quad (3.3)$$

A TRADES veszteségfüggvény első része az osztályozási feladatokból ismeretes, a második része felelős a robusztusságért, ez szabályozható a β súllyal. Magasabb β megadásával robusztusabb, de kevésbé pontos modelleket kapunk, míg $\beta = 0$ esetén az algoritmus a standard (nem robusztus) tanítás. Robusztus hálózatok tanításához a TRADES módszert $\beta = 10$ paraméterezéssel használtam.

Osztályozó hálózatok robusztusságának kiértékeléséhez az osztályozási feladathoz tartozó adatbázis tesztalmazának egyes elemeit támadom, és a kapott ellenséges példák, azok eredeti címkéivel értékelem ki a hálózat pontosságát. A dolgozatban az ellenséges példák előállítását az AutoAttack [34] támadás "rand" verziójának az Adversarial Robustness Toolbox (ART) [80] által adaptált implementációját használtam. Az AutoAttack standard verziója négy különböző támadással generálja az ellenséges példákat, míg a "rand" verzió csupán az Auto Projected Gradient Descent (aPGD) [34] támadás két különböző veszteségfüggvénnyel megvalósított változatát használja erre a célra. A "rand" verziójú támadás a standard verzióval jóval kevésbé költséges, ezen tovább javít az ART gyorsított implementációja, emellett a támadás erőssége csupán minimálisan marad el a standard AutoAttack erősségétől (lásd: 3.2 táblázat). A támadások célpont nélküliek, azaz a támadás célja olyan ellenséges példa előállítása, amelyen a modell predikciója hibás. A dolgozatban az aPGD-ket 20 iterációval, újraindítás nélkül futtattam, a maximális megengedett ellenséges perturbáció mértéke a CIFAR10 és SVHN adatbázisokon tanított hálózatok esetén $\epsilon = 8/255$, míg MNIST és Fashion MNIST adatbázisokon tanított hálózatok esetén $\epsilon = 0.3$. A kezdeti optimalizálási lépésköz, szintén a szakirodalmi paraméterezéseknek megfelelően [22] az előbbi adatbázisok esetén $2/255$, míg az utóbbiak esetén 0.1 .

Modellvarrás esetén a varró réteg tanítási paraméterei a korábbiakhoz hasonlóak, azzal a különbséggel, hogy a kezdeti tanulási ráta 10^{-3} , illetve a varrást csak 30 epochig tanítottam. Amíg ezt külön nem említem, a felhasznált modellek-, és a vizsgált modellvarrások tanítási paramétereit nem optimalizáltam. A modellvarrás megvalósításához felhasználtam Csiszárík és munkatársai [55] implementációját.

3.4. Felhasznált modellek

A dolgozatban vizsgált hálózatok adatait a 3.1 táblázat tartalmazza. A dolgozatban felhasználtuk Addepalli és munkatársai [81], Sehwag és munkatársai [82], Rade és munkatársai [83], Engstrom és munkatársai [84] és Goyal és munkatársai [85] CIFAR10 adatbázison tanított robusztus hálóit. Ezen modellek használata lehetővé teszi a korábban

ismertetett támadás erejének felmérését. A dolgozatban alkalmazott támadás előnye az AutoAttack [34] standard megvalósításával szemben az időigénye, ugyanakkor valamelyest gyengébb az ellenséges példák előállításában. A dolgozatban használt támadás és az AutoAttack teljesítményének összehasonlítását a 3.2 táblázat tartalmazza.

Architektúra	Adatbázis	Pontosság	Robusztusság	Algoritmus
VGG-11	MNIST	99,61%	0%	
VGG-11	Fashion MNIST	94,12%	0%	
VGG-11	SVHN	95,35%	0%	
VGG-11	CIFAR10	90,23%	0%	
ResNet-18	MNIST	99,97%	0%	
ResNet-18	Fashion MNIST	94,77%	0%	
ResNet-18	SVHN	95,36%	0%	
ResNet-18	CIFAR10	90,67%	0%	
ResNet-18	CIFAR10	93,66%	0%	
PreActResNet-18	CIFAR10	93,82%	0%	
ResNet-50	CIFAR10	94,47%	0%	
WideResNet-28-10	CIFAR10	92,91%	0%	
PViT	CIFAR10	82,33%	0%	
ResNet-18	MNIST	99,39%	94,10%	Madry
ResNet-18	Fashion MNIST	90,65%	45,40%	Madry
ResNet-18	SVHN	85,32%	52,15%	TRADES
ResNet-18	CIFAR10	79,97%	47,50%	TRADES
ResNet-18	CIFAR10	80,24%	52,07%	[81]
ResNet-18	CIFAR10	84,59%	56,66%	[82]
PreActResNet-18	CIFAR10	89,02%	58,34%	[83]
ResNet-50	CIFAR10	87,03%	51,53%	[84]
WideResNet-28-10	CIFAR10	87,50%	64,31%	[85]
PViT	CIFAR10	67,28%	28,44%	Madry

3.1. táblázat. A dolgozatban használt modellek. Robusztus hálózatok esetén az utolsó oszlop a tanításhoz használt algoritmust, illetve state-of-the-art modellek esetén a publikációt mutatja.

Modell	Publikáció	AutoAttack ("rand")	AutoAttack (standard)
ResNet-18	[81]	52,07%	51,06%
ResNet-18	[82]	56,66%	55,54%
PreActResNet-18	[83]	58,34%	57,67%
ResNet-50	[84]	51,53%	49,25%
WideResNet-28-10	[85]	67,28%	63,44%

3.2. táblázat. A dolgozatban vizsgált state-of-the-art modellek robusztussága az általam használt, "rand" verziójú- és a RobustBench [60] ranglista által használt standard Auto-Attack használatával.

4. fejezet

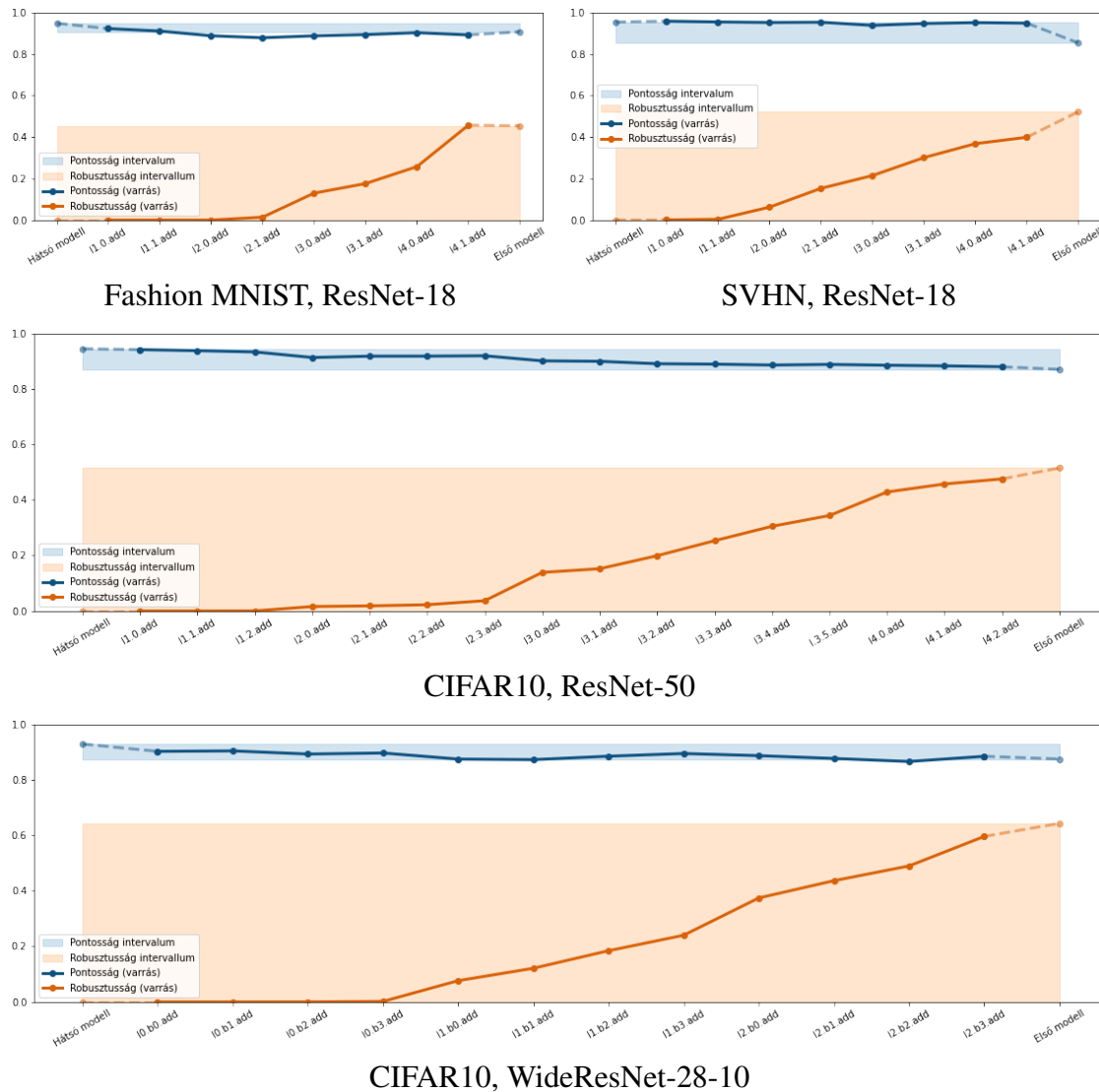
Osztályozási pontosság modellvarrással

A dolgozat alapkérdése, hogy a robusztus és nem robusztus (normál) hálózatok reprezentációi közötti mekkora a funkcionális eltérés. Ennek megfelelően elsőként a normál és robusztus reprezentációk funkcionális kompatibilitását vizsgálom az eredeti osztályozási feladaton. Ezt a feladatot mindkét típusú háló megoldja, azonban eltérő hatékonysággal (lásd: robusztusság-pontosság kompromisszum), illetve vitatott, hogy a modellek azonos módon oldják-e meg az eredeti feladatot, vagy robusztus hálók teljesen más reprezentációt tanulnak meg [26, 30]. Az ellenséges példákon történő robusztus tanítás során nem használjuk az eredeti (ellenséges perturbáció nélküli) bemeneteket, míg a módosított veszteségfüggvénnyel történő tanítással az osztályozási pontosság mellett magasabb szerepet élvez a robusztusság elérése, így jogos lehet a feltevés, miszerint az alapfeladatot a robusztus hálózatok máshogy oldják meg.

Ennek vizsgálatára különböző robusztus és normál hálózatok egyes rétegeit varrom össze az eredeti feladaton. Amennyiben a varrt modell osztályozási pontossága szignifikánsan alacsonyabb, mint a komponens modelleké, a varrt rétegek reprezentációi funkcionálisan nem hasonlók, ellenkező esetben kompatibilisek.

4.1. Reprezentációk funkcionális hasonlósága

A dolgozat első eredménye, hogy azonos architektúrájú robusztus és normál hálózatok összes rétege által kinyert reprezentációk funkcionálisan kompatibilisek az eredeti feladaton elért osztályozási pontosság szempontjából (lásd: 4.1 és 4.2 ábra). Ezen felül a varrt modell osztályozási pontossága a legtöbb esetben közel lineárisan interpolálja a komponens modellek pontosságát a varrás helyének függvényében, azaz az első réteg varrásánál a hátsó modell pontosságát közelíti, míg az utolsó előtti réteg varrásánál az elsőét. Ez a megállapítás független a modellek sorrendjétől, azaz robusztus első- és normál hátsó modellel és fordított esetben is fennáll.



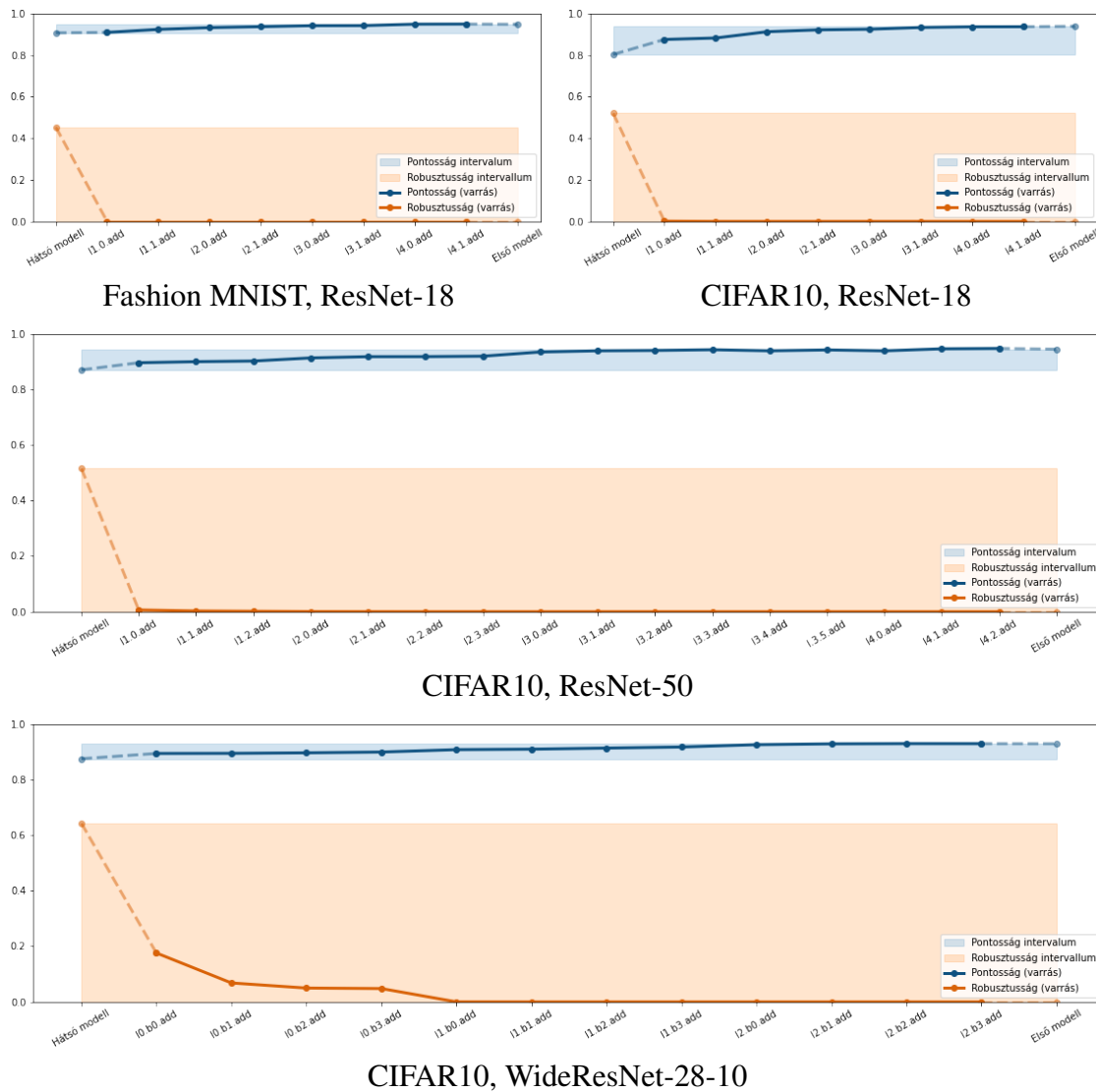
4.1. ábra. Modellvarrás a jelölt adatbázisokon tanított robusztus első- és normál hátsó modellel. A grafikonok egyes pontjai a varrt rétegeket mutatják. A kék vonal a varrt modell pontosságát, a narancssárga a varrt modell robusztusságát jelöli. A kék háttér a komponens modellek pontossága-, míg a narancssárga háttér azok robusztussága közötti intervallumot mutatja. Az ábra tartalmazza az első- és hátsó modellek pontosságát és robusztusságát (a modellvarrástól függetlenül), amelyeket a folytonosság érzékeltetése céljából szaggatott vonal köt a modellvarrás eredményeihez.

Normál és robusztus hálózatok varrásával nem csupán a varrt modell pontosságát lehet vizsgálni, hanem annak robusztusságát is. Ugyan tanítás során csak a varró réteg paraméterei módosulnak, a teljes varrt hálózathoz generálhatók ellenpéldák gradiens-alapú támadásokkal. Mivel a robusztus hálózatoknak az eredeti feladaton felül a robusztus osztályozás is feladata, a varrt modell robusztusságának vizsgálata lehetőséget nyújt a reprezentációk robusztusságának elemzésére. A feldolgozott szakirodalom alapján ez az első olyan tanulmány, amelyben a reprezentációk funkcionális hasonlóságának vizsgálata a

varrt modell robusztusságának elemzésével együttesen történik.

Amennyiben a varrás első modellje robusztus, a varrt modell osztályozási robusztussága a pontossághoz hasonlóan, a varrás helyének függvényében interpolálódik (lásd: 4.1 ábra). Ez jól mutatja, hogy a robusztusság a reprezentációk tulajdonsága is, hiszen modellvarrással az eredeti osztályozási feladat megoldása a cél, azonban a varrt modell a transzformált reprezentációkkal, nem robusztus fejjel is részben meg tudja oldani a robusztus osztályozási feladatot.

Amennyiben a varrás első modellje nem robusztus, a varrt modell robusztussága már az első rétegek varrásával teljesen eltűnik (lásd: 4.2 ábra). Ez meglepő, hiszen a legelső rétegnél végrehajtott varrás esetén a varrt modell az első rétegen és a varrást végző transzformáción kívül csupán robusztus rétegekből áll. Ezen felül általános feltevés, hogy



4.2. ábra. Modellvarrás a jelölt adatbázisokon tanított normál első- és robusztus hátsó modellel. A jelölések megegyeznek a 4.1 ábra jelöléseivel.

képosztályozó hálózatok első rétegei általános jellemzőkinyerést, többnyire "alacsony-szintű" képfeldolgozó műveleteket (pl. zajszűrés, éldetektálás) végeznek [57], azonban a robusztus osztályozási feladat megoldásához ezen rétegek reprezentációi önmagukban nem elegendők. Cianfrani és munkatársai [53] megmutatták a reprezentációs hasonlóság vizsgálatával, hogy normál modellek első rétegei robusztusak, azonban a funkcionális hasonlóság vizsgálatával ez már nem igaz, az első rétegek reprezentációi akkor sem robusztusak, ha a hálózat további része robusztus. Ezt az állítást a dolgozat későbbi részében a robusztusság célzott vizsgálatával pontosítom.

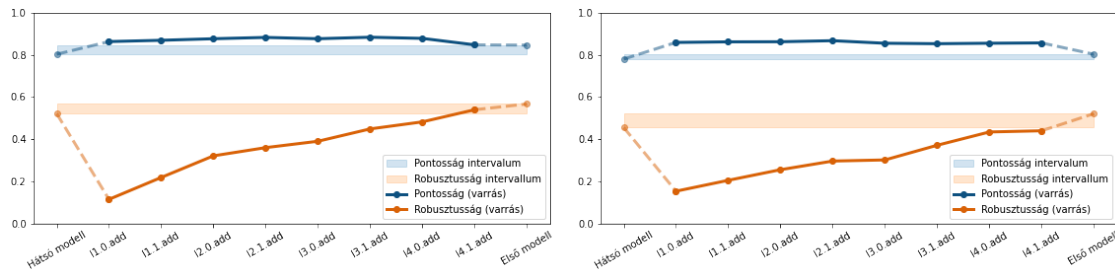
A kísérletek eredményei mutatják, hogy a robusztusság szempontjából a hálózat eleje nagyon fontos. Ehhez kapcsolódóan korábban Schott és munkatársai [86] megmutatták, hogy a gyakorlatilag bináris MNIST adatbázison a robusztus osztályozási feladat megoldható a hálózat elején egy egyszerű küszöbölés alkalmazásával, azonban ez a kijelentés nem általánosítható további, nehezebb osztályozási feladatokra.

4.2. A robusztusság instabilitása

A robusztusság a robusztus reprezentációk funkcionális tulajdonsága, azonban a modell-varrás figyelmen kívül hagyja a robusztusságot, hiszen a cél csupán az eredeti osztályozási feladaton vett funkcionális hasonlóság vizsgálata. Reprezentációk egy A funkcionális tulajdonsága legyen instabil egy másik, B funkcionális tulajdonsággal szemben, ha B előtérbe helyezésével (pl. célzott vizsgálatával, felerősítésével) az A funkció redukálódik. A dolgozatban vizsgált két funkcionális tulajdonságról, a robusztusságról és a pontosságról köztudott, hogy jelenlegi ismereteink szerint kompromisszumban állnak, azaz robusztusabb hálózatok nem érik el a normál hálózatok osztályozási pontosságát. A dolgozat ezen részében ezt a jelenséget a reprezentációk szintjén vizsgálom, azaz az elemzés célja a robusztusság instabilitásának vizsgálata a pontossággal szemben.

A robusztusság pontossággal szemben vett instabilitását különböző robusztus hálózatok varrásával vizsgálom. Ebben az esetben a varrás az eredeti osztályozási feladat megoldását helyezi előtérbe, a robusztus osztályozási feladat megoldása nem célja. Ennek megfelelően az instabilitás akkor áll fenn, ha a robusztusság a varrással szignifikánsan csökken, ellenkező esetben a reprezentációk robusztussága, az eredeti osztályozási feladat célzott vizsgálata ellenére stabil az osztályozási pontossággal szemben.

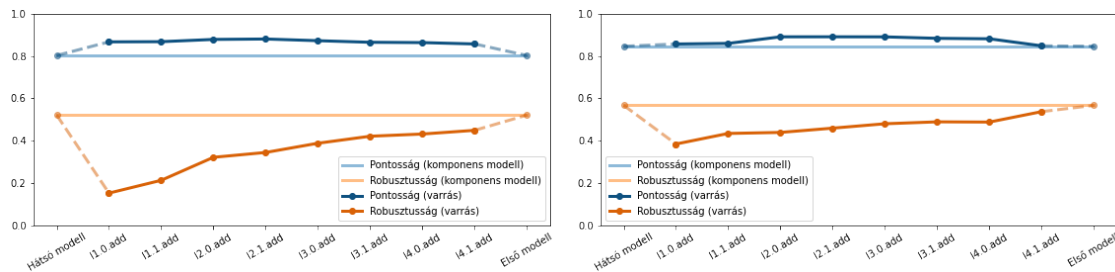
A dolgozat második eredménye, hogy reprezentációk robusztussága az osztályozási pontossággal szemben instabil. Ezen instabilitás mértéke rétegenként eltérő, a korai rétegekben jóval jelentősebb, mint a mélyebb rétegekben (lásd: 4.3 ábra). Ez tovább erősíti a feltételezést, amely szerint a korai rétegek a robusztusság szempontjából sokkal fontosabbak, hiszen a robusztusság ezekben a rétegekben jóval instabilabb, ugyanakkor a



4.3. ábra. Különböző robusztus, ResNet-18 architektúrájú hálózatok varrása a CIFAR10 adatbázison. A robusztusság pontossággal szemben vett instabilitásából adódóan a varrt modell robusztussága, főleg a korai rétegek varrásánál szignifikánsan csökken, a pontossága minimálisan nő.

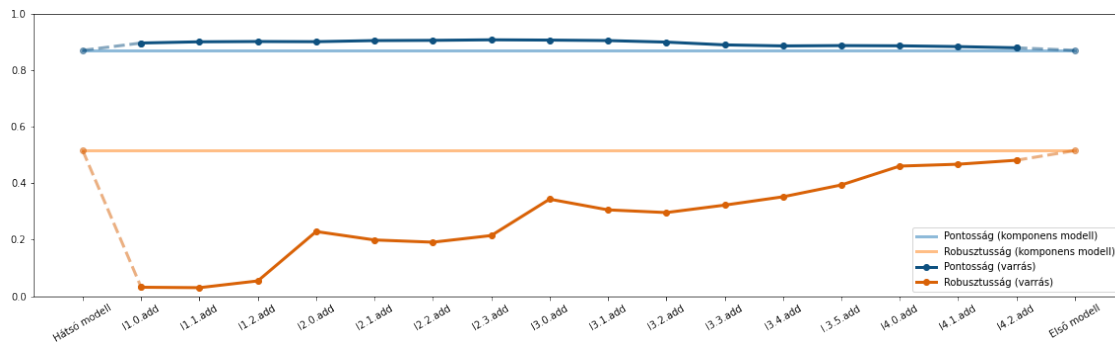
későbbi rétegekben fennálló instabilitás mutatja, hogy a hálózat egészének robusztusnak kell lennie a maximális robusztusság eléréséhez.

Különböző robusztus hálózatok varrásával kapcsolatban, a robusztusság instabilitását látva felmerülhet a kétely, miszerint a robusztusság megfeleltetése azért nem megvalósítható, mivel a reprezentációk "máshogy robusztusak", a hálózatok máshogy oldják meg a robusztus osztályozási feladatot. Ennek megcáfolására, illetve annak bemutatására, hogy a robusztusság instabilitását valóban a modellvarrás hozza felszínre, a varrást



ResNet-18

ResNet-18



ResNet-50

4.4. ábra. Azonos robusztus hálózatok varrása a CIFAR10 adatbázison. A modellvarrást az identikus leképezéssel inicializáltam. A kék folytonos vonal jelzi az eredeti modell pontosságát, a narancssárga vonal a robusztusságát.

azonos robusztus hálózatok között végzem el. A reprezentációk robusztusságának instabilitása ebben az esetben is ugyanúgy fennáll, ráadásul akkor is, amikor a modellvarrás súlyait a robusztusság szempontjából optimális megfeleltetéssel, az identikus leképezéssel inicializálom (lásd: 4.4 ábra). Neurális hálózatokról ismert, hogy a modell súlyainak kezdőértéke nagyban befolyásolhatja a tanítás sikerességét, azonban ez a kísérlet mutatja, hogy nem a varrás paramétereinek inicializálása okozza a robusztusság instabilitását.

Ilyas és munkatársai [29] érvelése alapján a robusztus hálózatok a bemenet jellemző-készletének robusztus, kevésbé támadható részét használják az osztályozáshoz. A funkcionális hasonlóság vizsgálatával azonban látható, hogy az osztályozási feladat megoldása szempontjából normál és robusztus hálózatok kompatibilis reprezentációkat nyernek ki. Ezen felül a robusztusság instabilitása mutatja, hogy a robusztus hálózatok által kinyert reprezentációknak is van nem robusztus része, amelynek előtérbe helyezésével a pontosság minimálisan növelhető, de a robusztusság szignifikánsan csökkenthető.

Robusztus és normál hálózatok varrásával, illetve a robusztusság instabilitásának ki-mutatásával látszik, hogy a standard modellvarrás nem alkalmas a robusztusság, mint reprezentációk funkcionális tulajdonságának vizsgálatára. Ennek ellenére az eredmények mutatják, hogy az osztályozási és a robusztus osztályozási feladat megoldásában a rep-rezentációk funkcionálisan leginkább a korai rétegekben térnek el, mivel az instabilitás ezekben a legjelentősebb, illetve robusztus hálózatokban legelső rétegek normál rétegek-vel történő helyettesítésével, az osztályozási pontosságra fókuszálva a hálózat robusztus-sága teljesen eltűnik.

5. fejezet

Robusztus modellvarrás

A szakirodalomban nem robusztus reprezentációk funkcionális kompatibilitását csupán a hálózatok eredeti feladatán mutatták meg. A robusztus hálózatoknak azonban a robusztus osztályozási feladatot is meg kell tudniuk oldani, ráadásul az ellenséges példákon történő robusztus tanítási módszerekkel létrehozott hálózatok esetén a robusztus osztályozási feladat felfogható a hálózat alapfeladataként.

A 4. fejezetben megmutattam, hogy robusztus és normál reprezentációk funkcionálisan kompatibilisek az eredeti (nem robusztus) osztályozási feladaton, valamint, hogy a robusztusság a reprezentációk funkcionális tulajdonsága. Ebben a fejezetben bevezetek egy új módszert, a robusztus modellvarrást, amellyel a normál és robusztus reprezentációk funkcionális kompatibilitását a robusztusság szempontjából vizsgálom.

A robusztus modellvarrás módszere a standard modellvarráshoz hasonló (lásd: 3.2. fejezet), azonban a varrást végző transzformáció tanítása robusztusan történik. Ehhez a korábban említett Madry-protokollt [22], a TRADES veszteségfüggvényt [31] és egy részben ellenpéldákra alapuló módszert használok. Ez utóbbi módszer a Madry-protokollhoz hasonlóan ellenséges példákat generál a tanítás során, azonban ha egy x példát nem lehet sikeresen támadni, azaz nem lehet olyan $\delta \in \mathcal{S}$ megengedett perturbációt találni, hogy a modell az $x + \delta$ példát helytelenül osztályozza, a támadás az eredeti példát adja vissza, hozzáadott perturbáció nélkül. Robusztus tanításhoz ebben az esetben a Madry-protokoll által használt PGD-nél erősebb támadásra van szükség, ezért ehhez az AutoAttack "rand" verziójának olyan módosítását használok, amelyben az AutoAttack támadásainak iterációs számát 10-re csökkentettem (a kiértékeléshez 20 iterációt használok). Az algoritmus tehát a támadható példákhoz erős ellenpéldát generál, a nem támadható példák esetén az eredeti példákat használja, így a TRADES módszerhez hasonlóan tartalmaz "standard" és "robusztus" komponenst. Fontos megjegyezni, hogy az ellenséges példák a robusztusság kiértékeléséhez hasonlóan az ellenpéldákra alapuló robusztus modellvarrás során is a teljes varrt modellhez generáltak, míg a tanítás során csak a varró réteg paraméterei mó-

dosulnak. Kísérleteimben a TRADES veszteségfüggvénnyel megvalósított robusztus modellvarrás esetén a robusztusságot befolyásoló β paramétert 1-re állítottam, amennyiben a varrás legalább egy komponens modellje nem robusztus és 10-re, ha a mindkét komponens hálózat robusztus, ezzel is kiegyensúlyozva a robusztus és normál reprezentációk robusztusságával kapcsolatos elvárásokat. Minden további hiperparaméter megegyezik a korábban ismertetett paraméterekkel (lásd: 3.3. fejezet).

Normál és robusztus reprezentációk funkcionális kompatibilitása az eredeti osztályozási feladaton vizsgált pontosságon kívül elemezhető a robusztusság szemszögéből is. Ehhez a különböző robusztus modellvarrási módszereket alkalmazom normál és robusztus hálózatok egyes rétegeinek varrására. Elsődlegesen a varrt modell robusztussága vizsgálendő, ugyanakkor a robusztusan tanított modellek az eredeti feladatot is megoldják, így a reprezentációk robusztus funkcionális kompatibilitására a varrt modell pontosságából is következtethetünk.

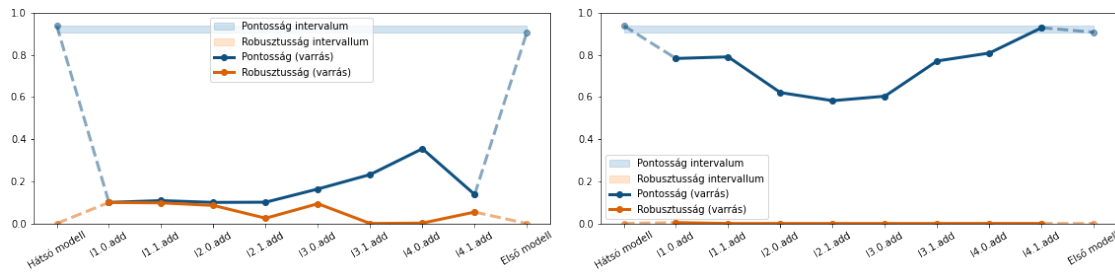
5.1. A robusztus modellvarrás tulajdonságai

A modellvarrás tulajdonságait Bansal és munkatársai [50], Csiszárík és munkatársai [55] és Jones és munkatársai [72] is vizsgálták és megállapították, hogy egy hálózat pontossága nem növelhető szignifikánsan modellvarrással. Természetesen minimális növekedés elvárható, hiszen a varró réteg megnöveli a hálózat paramétereinek számát.

A robusztus modellvarrással szemben hasonló elvárások támaszthatók, azzal a különbséggel, hogy a módszer célja az osztályozási robusztusság elemzése, így a robusztus modellvarrás célja a robusztusság megőrzése, esetleg minimális növelése. Természetesen ez csak akkor történhet meg, ha a reprezentációk robusztusságát a hátsó modell ki tudja használni.

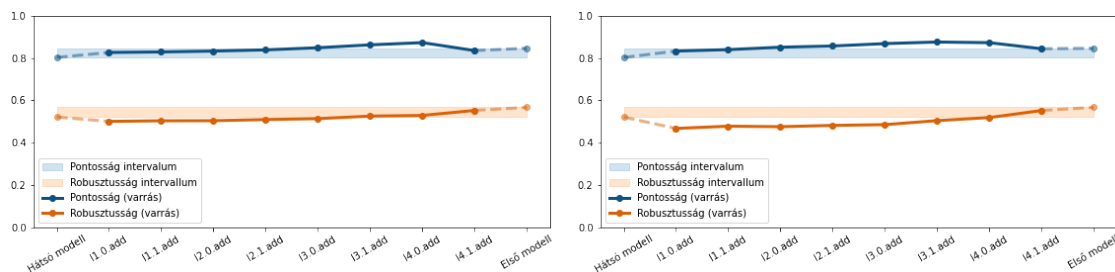
Ennek megfelelően robusztus modellvarrással nem robusztus modellek reprezentációi funkcionálisan teljesen inkompatibilisek. A normál hálózatokon végrehajtott robusztus modellvarrásnak két kimenete lehetséges: a varrt modell robusztussága 0% és az eredeti modellekhez képest csökken az osztályozási pontosság (a robusztusság-pontosság kompromisszumból adódóan), vagy a varrt modell pontossága és robusztussága egyaránt $1/C$, ahol C a lehetséges osztálycímkek száma (lásd: 5.1 ábra). Ez utóbbi esetben a modellvarrás egy robusztus optimumpontot talál, a reprezentáció "kinullázását". Ezzel minden bemenet, beleértve az ellenséges példákat, ugyanazt a címkét kapja, illetve a gradiens-alapú támadások nem fognak működni (a 0 súlyok mentén a varrásnál eltűnik a gradiens).

A robusztus modellvarrással kiküszöbölhető a robusztusság standard modellvarrásnál tapasztalt instabilitása, azaz robusztus reprezentációk osztályozási pontossága a robusztussággal szemben nem instabil. Ez elvárható, hiszen a robusztus hálózatok elsődleges



5.1. ábra. Robusztus modellvarrás különböző normál, CIFAR10 adatbázison tanított, ResNet-18 hálózatok között. A bal ábra a Madry-protokollal, a jobb ábra a részben ellen-séges példákon történő tanítás eredményeit mutatja.

feladata a robusztus osztályozási feladat megoldása, azonban ezzel az eredeti osztályozási feladaton is magas pontosságot érnek el. Különböző robusztus hálózatok tetszőleges réte-geinek robusztus modellvarrásával a varrt modell robusztussága megközelíti a komponens hálózatok robusztusságát, valamint azonos robusztus hálózat varrása esetén a robusztus modellvarrás véletlen inicializálással szintén az eredeti hálóval közel azonos robusztussá-gú varrt modellt eredményez. Ezzel a reprezentációk már funkcionálisan kompatibilisek a robusztusság szempontjából, emellett a robusztus modellvarrás az osztályozási pontossá-got is megőrzi.



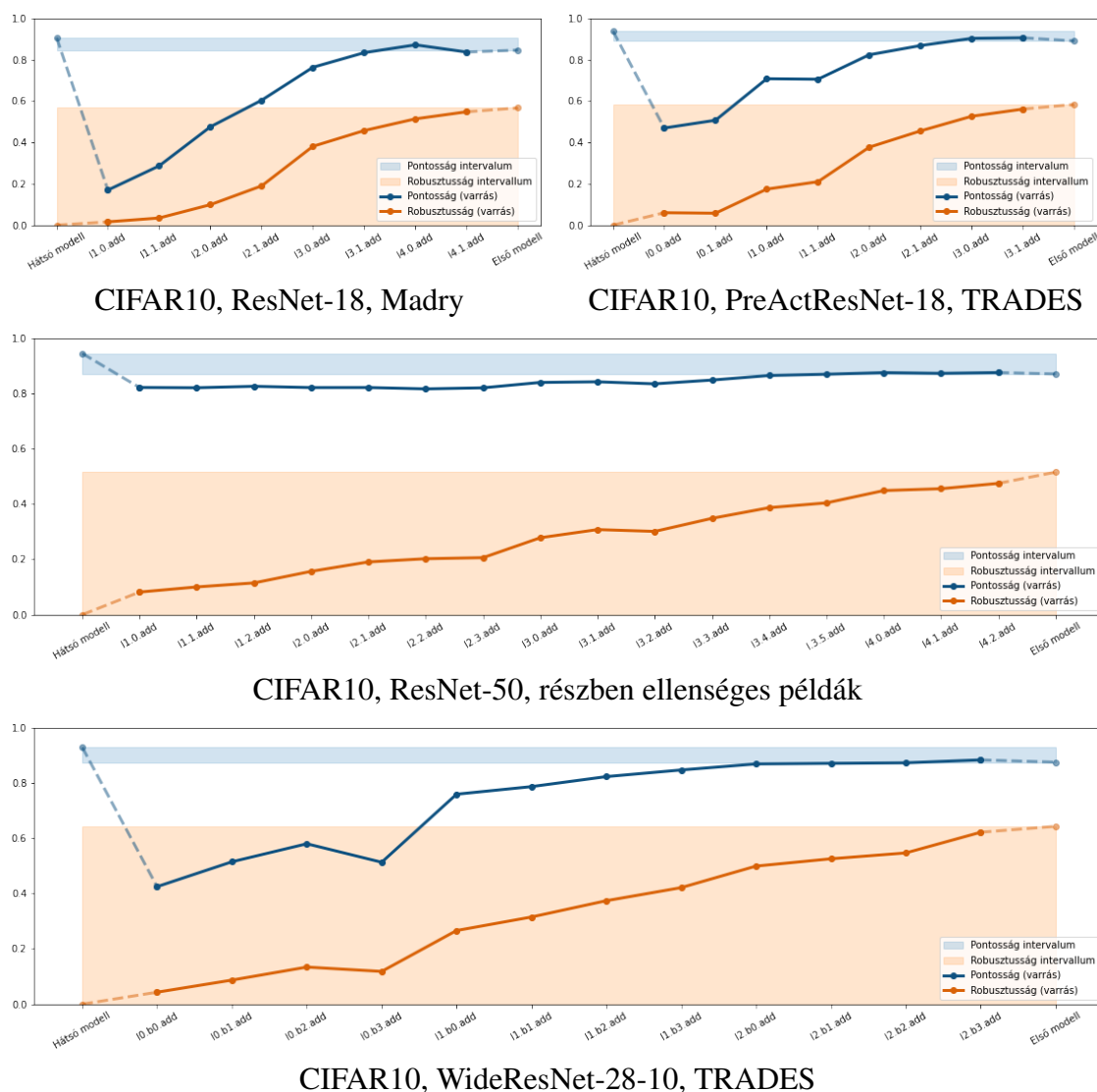
5.2. ábra. Robusztus modellvarrás különböző robusztus, CIFAR10 adatbázison tanított, ResNet-18 hálózatok között. A bal ábra a Madry-protokollal, a jobb ábra a TRADES veszteségfüggvényen történő tanítás eredményeit mutatja.

Ezekkel az önellenőrzésnek is tekinthető vizsgálatokkal kijelenthető, hogy a robusztus modellvarrás alkalmas a reprezentációk funkcionális kompatibilitását a robusztusság szempontjából vizsgálni. A dolgozat egy fontos üzenete, hogy a reprezentációk funk-ciónális hasonlósága nem csupán az osztályozási pontosság szempontjából vizsgálható, illetve a több feladatot (pl. osztályozás és robusztus osztályozás) megoldó reprezentációk funkcionális hasonlósága a megoldott feladatokon külön vizsgálandó. Ez a szemléletmód a neurális hálózatok vizsgálatában módszertani újítás.

5.2. Reprezentációk robusztus funkcionális hasonlósága

A dolgozat harmadik eredménye, hogy azonos architektúrájú robusztus és normál hálózatok reprezentációi a robusztusság szempontjából, tetszőleges robusztus modellvarrási módszerrel vizsgálva *részben* kompatibilisek.

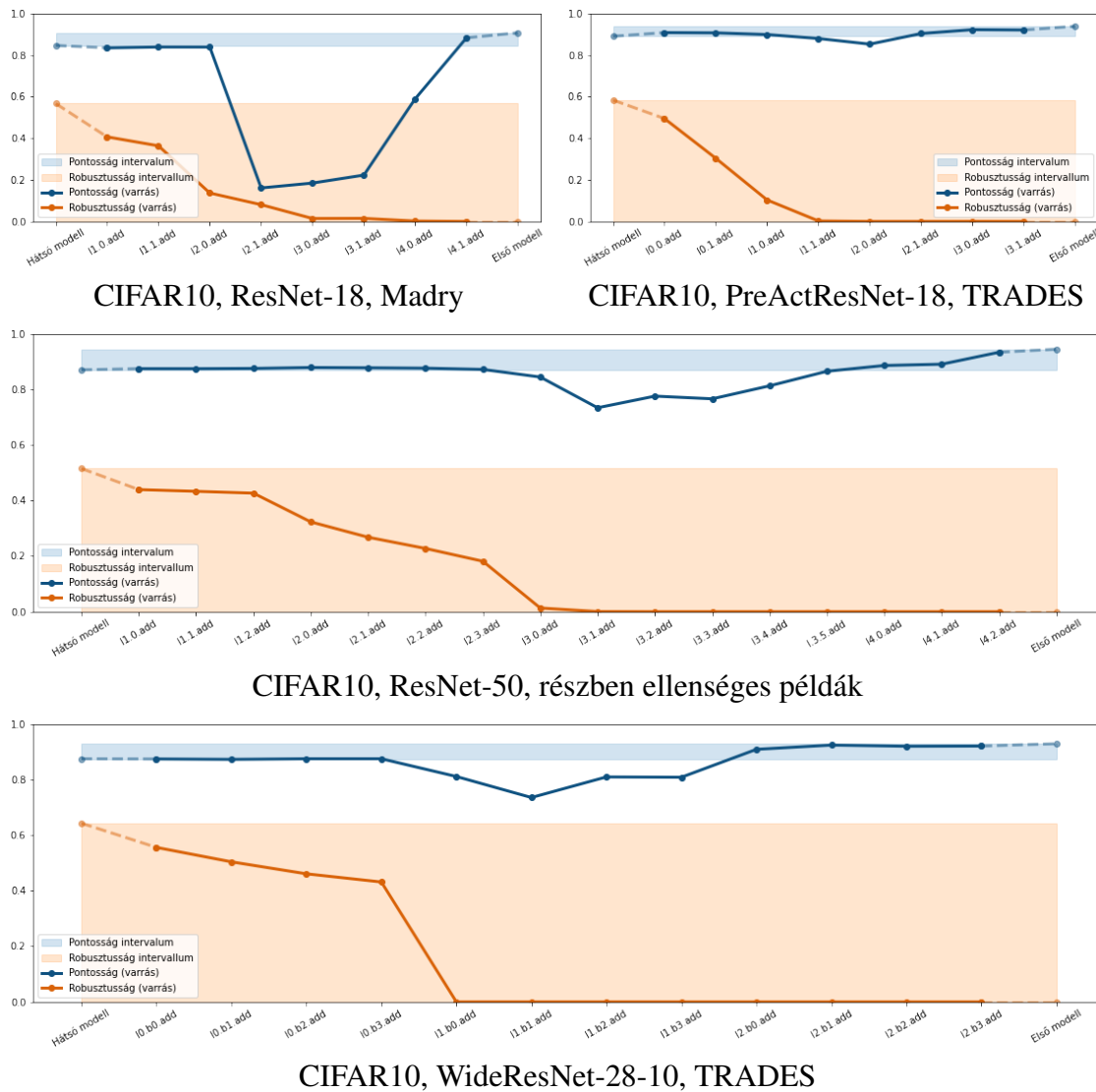
Amennyiben a varrás első modellje robusztus, a robusztus modellvarrással eredményeként kapott modell robusztussága minimálisan nő a standard modellvarrással elérhető robusztussághoz képest, ugyanakkor az osztályozási pontosság szignifikánsan csökken annak függvényében, hogy a normál hátsó modell mekkora részben szerepel a varrt modellben (lásd: 5.3 ábra). A robusztusság növekedése érthető (elvégre a modellvarrás már a robusztusságot helyezi előtérbe), az osztályozási pontosság korai rétegek varrásánál meg-



5.3. ábra. Különböző robusztus modellvarrási módszerek robusztus első- és normál hátsó modellekkel. A jelölések megegyeznek a 4.1 ábra jelöléseivel.

figyelhető csökkenését a dolgozat későbbi részében magyarázom a robusztus modellvarrás ezen esetének (robusztus első- és normál hátsó modell) általánosításával (lásd: 5.4. fejezet).

Amennyiben a varrás hátsó modellje robusztus, a robusztus modellvarrás eredményei változatosak (lásd: 5.4 ábra). A korai normál rétegek reprezentációi tetszőleges robusztus modellvarrási módszer alkalmazásával funkcionálisan kompatibilisek a robusztus rétegekkel, azaz a varrt modell robusztussága megközelíti a robusztus hátsó modell robusztusságát az osztályozási pontosság csökkenése nélkül. Korábban megmutattam, hogy normál hálózatok első rétegei önmagukban nem robusztusak, azonban a robusztus modellvarrás eredményei mutatják, hogy lineáris transzformációval ezen rétegek reprezentációi robusztussá tehetők. Ez tovább erősíti a feltevést, mely szerint a robusztus osztályozás megva-



5.4. ábra. Különböző robusztus modellvarrási módszerek normál első- és robusztus hátsó modellekkel. A jelölések megegyeznek a 4.1 ábra jelöléseivel.

lósításához a hálózatok első rétegeinek robusztussága fontos, valamint igazolja, hogy az első rétegek a robusztusság szempontjából is általános jellemzőket nyernek ki. Ez az eredmény azt is mutatja, hogy a standard modellvarrással tapasztalt eltűnő robusztusság (lásd: 4.2 ábra) nem abból adódik, hogy a robusztus és normál hálózatok különböző jellemzőkészletet használnak, elvégre a korai normál reprezentációk robusztussá tehetők.

A mélyebb rétegeknél, robusztus hátsó modellel végrehajtott robusztus modellvarrási módszerek eredményei a varrt modell osztályozási pontosságában jelentősen eltérők, azonban az osztályozási robusztusság ezen rétegek varrásánál konzisztensen 0%. Ez azt mutatja, hogy a robusztusság szempontjából a normál hálózatok mélyebb rétegeinek reprezentációi nem kompatibilisek a robusztus hálózatok reprezentációival. Ez ismét hangsúlyozza a korai rétegek fontosságát, hiszen ezeket a mélyebb rétegekben végrehajtott modellvarrás nem módosítja, ezáltal a robusztusságuk hiánya továbbterjed a mélyebb rétegekbe.

A varrt modell osztályozási pontossága többnyire megközelíti a robusztus komponens modell pontosságát, azonban néhány esetben drasztikus csökkenések tapasztalhatók. Mivel a robusztus modellvarrásnak nem direkt célja az osztályozási pontosság megőrzése, így a hálózatok mélyebb rétegeiben, ahol a robusztusság szempontjából inkompatibilisek a reprezentációk, a varrt modell osztályozási pontossága alapján csak a robusztus osztályozási feladat megoldásának nehézségére lehet következtetni. Mivel erre a normál reprezentációk nem alkalmasak, így a robusztus modellvarrás a robusztus osztályozási feladat (többnyire sikertelen) megoldása érdekében a pontosság beáldozására kényszerül.

5.3. Robusztusság a komponens modellek ellenpéldáin

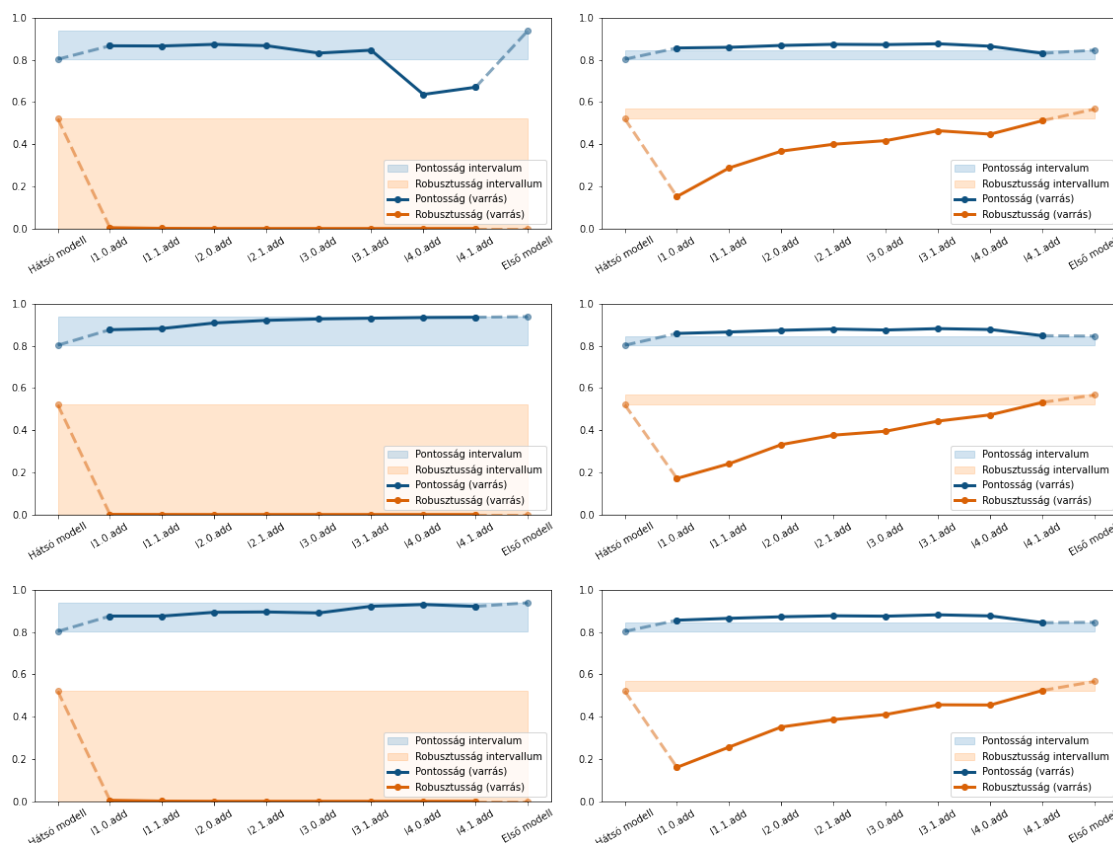
A robusztus modellvarrással elemezhető hálózatok reprezentációinak robusztussága, azonban ehhez a varró réteget robusztusan kell tanítani. Az ellenséges példákon történő robusztus modellvarrásnál felmerülhet a kétely, miszerint a varrt modell ellenséges példái csupán az egyik komponens hálózat ellenséges példái, ha az adott hálózat a varrásban "többségben van", azaz a varrt modell ezen hálózat jóval több rétegét tartalmazza. Közismert, hogy az ellenséges példák hálózatok között "transzferálódnak", azaz egy hálózat-hoz generált ellenséges példák másik hálózatokat is félrevezetnek. Ezek alapján felmerül a kérdés, hogy a reprezentációk robusztussága vizsgálható-e robusztus modellvarrás nélkül, csupán a varrás komponens modelljeinek ellenséges példáinak használatával.

Ennek vizsgálatára a modellvarrás tanítását a komponens hálózatok ellenséges példáin végzem. Az eredeti modellek ellenséges példáit a Madry-protokoll által használt PGD támadással állítom elő. A transzferált ellenséges példákon történő tanítást három különböző módon valósítom meg: csak az első-, illetve csak a hátsó modellhez generált

ellenpéldák használatával, valamint a két hálózathoz generált ellenséges példák együttes használatával. Utóbbi esetben egy epoch során egy bemenet kétszer kerül felhasználásra (az első- és hátsó modellhez generált ellenséges példa előállításához), a tanulás ideje ennek megfelelően duplázódik.

A dolgozat negyedik eredménye, hogy a reprezentációk robusztusságának vizsgálatához a varró réteget is robusztusan kell tanítani, a komponens hálózatok ellenséges példáin történő tanítás nem alkalmas a reprezentációk robusztusságának elemzésére (lásd: 5.5 ábra). A robusztus modellvarrás és az eredeti modellek ellenséges példáin történő modellvarrás eredményei drasztikusan eltérnek, ez utóbbi módszer eredményei a standard modellvarrással kapott eredményekhez hasonló képet mutatnak (lásd: 4.2 és 4.3 ábrák).

Ez az eredmény megmutatja továbbá, hogy a reprezentációk funkcionális hasonlósága a robusztusság szempontjából nem vizsgálható reprezentáció-megfeleltetéssel. A reprezentáció-megfeleltetés az f_i és g_j rétegek által, egy előre rögzített példahalmazon



5.5. ábra. Modellvarrás a komponens hálózatok ellenpéldáin: normál első- és robusztus hátsó modellel (bal oszlop), illetve robusztus modellek között (jobb oszlop). A modellvarrás mindhárom esetben (első modell ellenpéldái: felső sor, hátsó modell ellenpéldái: középső sor, mindkét modell ellenpéldái: alsó sor) a standard modellvarrás eredményeihez áll közel (lásd: 4.2 és 4.3 ábrák), tehát a robusztusság elemzéséhez nem elegendő a komponens modellek ellenpéldáinak használata.

kinyert reprezentációk alapján állítja elő a reprezentációkat illesztő transzformációt, tanulás és további finomhangolás nélkül. Az megfeleltetést megvalósító réteg támadása nélkül a robusztus illesztés csupán az ellenséges példák reprezentációit tudná használni, azonban ez a tanulással finomhangolt modellvarrással sem alkalmas a reprezentációk robusztusságának vizsgálatára.

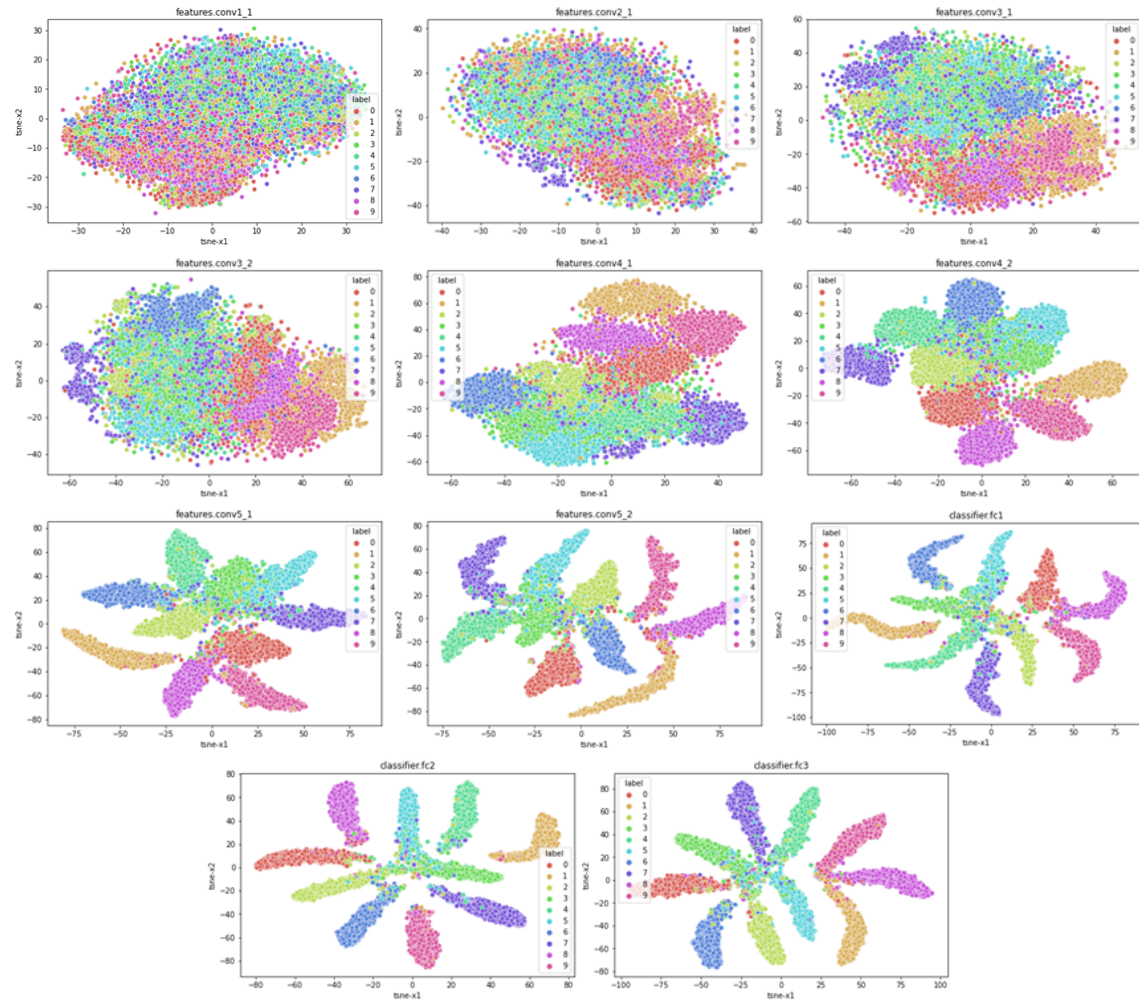
5.4. A robusztus modellvarrás általánosítása

Emlékezzünk vissza a modellvarrás alapkérdésére (3.2. fejezet): meg tudja oldani egy g hálózat a feladatát egy f hálózat reprezentációjával? A robusztus modellvarrás azonban nem ezt a kérdést válaszolja meg abban az esetben, amikor a varrás első modellje robusztus, a hátsó modellje viszont nem, mivel a normál hálózatoknak alapértelmezetten nem feladata a robusztus osztályozási feladat megoldása. Ebben az esetben a reprezentációk robusztussága minimálisan felerősíthető (lásd: 5.3 ábra), ugyanakkor már a standard modellvarrással végrehajtott vizsgálatoknál is megállapítottam, hogy a robusztusság a reprezentációk funkcionális tulajdonsága, így ennek felerősítése a robusztus modellvarrással elvárható.

A dolgozat ezen részében a modellvarrás működését vizsgálom a robusztus modellvarrás előbb ismertetett esetének általánosításával. A modellvarrás feladatát a következőképp módosítom: adott két, *különböző eloszláson tanított* neurális hálózat, $f = f_m \circ \dots \circ f_1$ és $g = g_k \circ \dots \circ g_1$ ($m, k \in \mathbb{N}$) és ezek egy-egy rögzített rétege, f_i és g_j ($1 \leq i < m$ és $1 \leq j < k$), a cél egy olyan $T : \mathcal{A}_{f,i} \rightarrow \mathcal{A}_{g,j}$ transzformáció tanulása, amivel $\mathcal{D}_{Y|X}[g_{>j} \circ T \circ f_{\leq i}] \approx \mathcal{D}_{Y|X}[f]$. Mivel az ellenséges példák normál hálózatok számára *eloszláson kívüli* (out-of-distribution) példáknek tekinthetők, illetve a robusztus osztályozás nem feladata a normál hálózatoknak, a feladat módosítása a robusztus első- és normál hátsó modellel végrehajtott robusztus modellvarrás általánosításának is tekinthető.

Goldfeld és munkatársai [59] megmutatták, hogy a neurális hálózatok az "Information Bottleneck" elmélet [87] következményeként tömörítik a reprezentációikat. Publikációjukban kisebb hálózatokra bemutatták, hogy a tömörítés osztályozó hálózatok esetén a reprezentációk klasztereződésében valósul meg, azaz a hálózatok mélyebb rétegeinek reprezentációi egy osztályba tartozó példák esetén egymáshoz közel esnek (páronkénti távolságuk alacsony), a különböző osztályokba tartozó példák reprezentáció egymástól távol esnek. A klaszterezés célja a reprezentációk lineáris szeparálhatóságának elérése, hiszen a hálózatok utolsó rétege lineáris osztályozást végez.

A reprezentáció-klasztereződés a dolgozatban használt, nagyobb modellekre is teljesül, a rétegenként fokozatosan elkülönülő klasztereket a t-SNE [88] vizualizációs módszerrel létrehozott 5.6 ábra szemlélteti. A klasztereződést a korábban bevezetett általá-

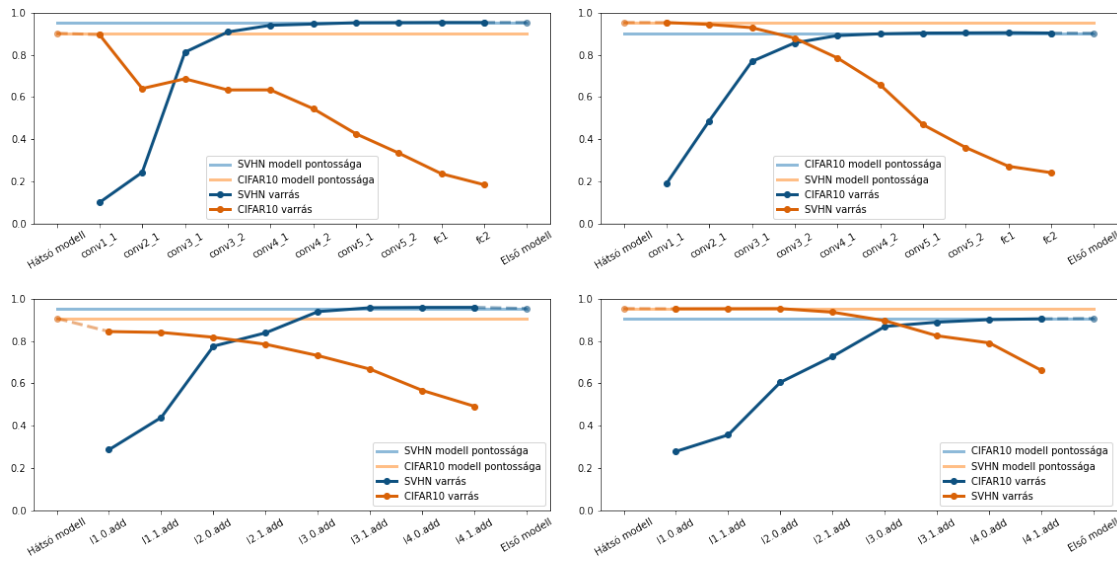


5.6. ábra. Egy CIFAR10 adatbázison tanított VGG-11 hálózat reprezentációinak osztályonkénti klaszterei rétegenként, a CIFAR10 adatbázis tesztalmazán. A rétegek sorfolytonosan követik egymást, az egyes osztályok különböző színnel szerepelnek.

nosított robusztus modellvarrás megvalósítására használok fel, amellyel a modellvarrás működését is megmagyarázom.

A dolgozat ötödik eredménye, hogy a mélyebb rétegekben kialakuló reprezentáció-klaszterek modellvarrással megfeleltethetők (lásd: 5.7 ábra). Ehhez különböző feladatokon tanított hálózatokat varrok össze úgy, hogy a varró réteget az első hálózat tanításához használt adatbázison tanítom. Amennyiben a varrt réteg által kinyert reprezentációknál már fellép a klaszterezés, a varrt modell pontossága teljesen megegyezik az első modell pontosságával, tehát a két hálózatban kialakuló reprezentáció-klaszterek megfeleltethetők. Ez az eredmény meglepő, mivel a hálózatok a középső rétegektől kezdve, teljesítményvesztés nélkül varrhatók, nem csupán az utolsó előtti rétegnél.

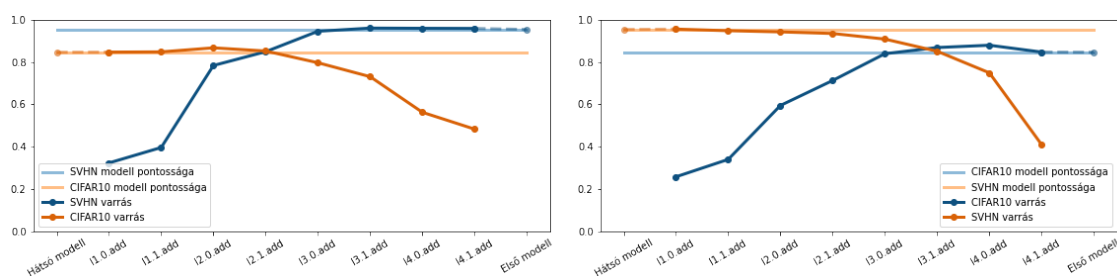
A hátsó modell tanításához használt adatbázison tanított modellvarrás esetén Lenc és Vedaldi [57] eredményei látszanak, miszerint a korai rétegek általános jellemzőket nyer-



5.7. ábra. A CIFAR10 és SVHN adatbázisokon tanított VGG-11 (felső sor) és ResNet-18 (alsó sor) hálózatok varrása. A bal oszlopban az SVHN-, a jobb oszlopban a CIFAR10 használatával tanított hálózat a varrás első modellje. A kék görbe jelöli az első hálózat feladatához tartozó adatbázison tanított modellvarrás pontosságát (általánosított robusztus modellvarrás), míg a narancssárga görbe a hátsó háló adatbázisán tanított modellvarrás pontosságát (standard modellvarrás).

nek ki, még akkor is, ha azok teljesen más adatbázison voltak tanítva, így ezen rétegek varrhatók. A mélyebb rétegek által kinyert reprezentációk feladat-specifikusságát az eloszláson kívüli példákon történő modellvarrás sikertelensége is mutatja, ugyanis a klaszterezés csak az eloszláson belüli példákon valósul meg.

További meglepő eredmény, hogy a klasztereződést kihasználó általánosított robusztus modellvarrás robusztus hálózatokon is megegyező eredménnyel alkalmazható (lásd: 5.8 ábra). Ez megmagyarázza a robusztus és normál reprezentációk közötti, az osztályozási pontosság szempontjából értelmezett teljes funkcionális kompatibilitást, hiszen robusztus és normál hálózatok egyaránt reprezentáció-klaszterezés segítségével oldják meg



5.8. ábra. Az SVHN adatbázison tanított normál hálózat és a CIFAR10 adatbázison tanított robusztus modell varrása. A bal ábrán az SVHN modell, míg a jobb ábrán a CIFAR10 hálózat a modellvarrás első modellje. A jelölések megegyeznek a 5.7 ábra jelöléseivel.

az osztályozási feladatot.

A klasztereződést kihasználó modellvarrás eredményei a neurális hálózatok tanítására, illetve a betanított hálózatok újrafelhasználására (transfer learning) is hatással vannak. Mivel a klaszterezés a probléma struktúrájától (az osztályok számától) függ, így egy új feladat megoldására előállítható olyan neurális hálózat, amely egy hasonló struktúrájú feladatot megoldó (betanított) hálózat mélyebb rétegeit finomhangolás nélkül felhasználja és csupán a bemenethez közeli rétegeket tanítja újra. Hasonló jelenséget a beszéd felismerés területén Zhang és munkatársai [89] is megfigyeltek, azonban ennek okát nem vizsgálták.

6. fejezet

Diszkusszió

A dolgozat alapvető célja robusztus és nem robusztus hálózatok reprezentációinak elemzése, valamint ezen reprezentációk funkcionális hasonlóságainak és különbségeinek feltárása. A dolgozat központi megfigyeléseként megállapítottam, hogy a reprezentációk funkcionális hasonlósága a pontosság és a robusztusság szempontjából külön vizsgálándó, ezáltal a funkcionális hasonlóságnak több dimenziója létezik attól függően, hogy a neurális hálózatok milyen feladatokat oldanak meg. A jövőben kutatási téma lehet olyan osztályozó hálózatok reprezentációinak vizsgálata, amelyek a robusztusságtól eltérő további feladatokat oldanak meg (pl. eloszláson kívüli példák detekcióját), illetve olyan, akár robusztus hálózatokat, amelyek komplexebb feladatokat oldanak meg (pl. rekonstrukció, szegmentálás). Szintén vizsgálandó a komplexebb feladatot megoldó hálózatok reprezentációinak tömörítési módja, amely eddig csak osztályozó hálózatokról ismert.

Kísérleteimmal megmutattam több funkcionális hasonlóságot a robusztus és normál reprezentációk között, így például az osztályozási feladat megoldásának módját, amely a reprezentációk klasztereződésén alapul. A reprezentációk közötti funkcionális különbségek a robusztusság szempontjából keresendők, ráadásul főként a korai rétegekben. A dolgozat eredményei alapján az osztályozó hálózatok rétegeit a reprezentációk funkcionális tulajdonságai alapján két csoportra lehet osztani: előfeldolgozó és szemantikai rétegekre. A korai, előfeldolgozó rétegek alacsonyszintű jellemzőkinyerést, képfeldolgozást végeznek, amely funkció hasonló robusztus és normál hálózatok között, ezért ezen rétegek reprezentációi normál hálózatok esetén is robusztussá tehetők, ha a robusztusságot a háló további része (modellvarrással) ki tudja használni. A szemantikai rétegek az osztályozási feladat megoldását végzik, ezekben a rétegekben alakulnak ki a reprezentáció-klaszterek. További kutatási téma lehet az előfeldolgozó és szemantikai rétegek szeparációjára tett javaslat és az Achille és munkatársai [67] által megmutatott, súlyokban tárolt, bemenetet reprezentáló információ degradálódásának vizsgálata és a két elmélet közötti kapcsolat feltárása.

Szintén további kutatási téma lehet a robusztusság és a klasztereződés kapcsolata. Megmutattam, hogy a robusztus hálózatok is klaszterezéssel oldják meg az osztályozási feladatot, azonban a robusztus és normál hálózatok által létrehozott klaszterek között is létezhet különbség. Érdekes lehet az ellenséges példák klaszterezésre gyakorolt hatásának vizsgálata is, ugyanis az ellenséges példák reprezentációinak a klasztereken kívül kell esniük (elvégre a hálózat félrevezetése a céljuk), viszont robusztus hálózatok esetén az ellenséges példák egy része továbbra is a klaszteren belül esik. Ismert, hogy a veszteségfüggvény módosításán alapuló módszerek (pl. TRADES [31], MART [24]) a döntési felületet úgy módosítják, hogy az a példáktól távolabb legyen, ezáltal a támadásokkal létrehozott ellenséges példák továbbra is a döntési felület megfelelő oldalán maradjanak. A robusztus tanulás reprezentáció-klaszterekre gyakorolt hatása további kutatási téma, feltételezhető, hogy a robusztus osztályozási feladat megoldásához jobban szeparált klaszterek létrejötte szükséges.

A dolgozat limitált hosszából kifolyólag nem tudtam minden kísérleti eredményemet megjeleníteni, ugyanakkor a dolgozat eredményei az összes vizsgált hálózatra és adatbázisra teljesülnek. Ez magában foglalja a vizsgált robusztus és normál képosztályozó Transformer hálózatokat is (Pooling Vision Transformer [61]), ugyanakkor a robusztus Transformer hálózatok előállítása és vizsgálata további kutatási téma. Ezek a hálózatok a kisebb képeket tartalmazó adatbázisokhoz tartozó robusztus feladaton (pl. CIFAR10-en, lásd: RobustBench [60]) kevésbé teljesítenek jól, azonban a nagyobb képeket tartalmazó ImageNet [90] adatbázison kiemelkedő robusztusságot érnek el [60]. Az ImageNet adatbázison tanított robusztus és normál hálózatok reprezentációinak összehasonlítása szintén egy jövőbeli kutatási feladat, erre hardveres limitációk miatt a dolgozat megírásáig nem volt lehetőségem.

7. fejezet

Összefoglalás

A dolgozatban a robusztus és nem robusztus neurális hálózatok reprezentációit, ezek funkcionális hasonlóságait és eltéréseit, valamint általánosan a neurális hálózatok működését vizsgáltam. Megállapítottam, hogy több funkciót megvalósító neurális hálózatok esetén a reprezentációik funkcionális hasonlóságát a különböző funkciók mentén szükséges elemezni, amely a neurális hálózatok működésének vizsgálatában módszertani újítás.

A robusztus osztályozó hálózatok ilyen, több funkciót megvalósító hálózatok, amelyek reprezentációinak funkcionális hasonlóságát az osztályozási pontosságon felül az osztályozási robusztusság szempontjából is vizsgáltam a modellvarrás eszközével. A reprezentációk robusztusságának elemzésére bevezettem a robusztus modellvarrás módszerét és megmutattam, hogy a funkcionális hasonlóság robusztusság szempontjából történő vizsgálatára ez a jelenleg ismert egyetlen helyes módszer. A feldolgozott szakirodalom alapján ez az első olyan tanulmány, amelyben reprezentációk funkcionális hasonlóságának vizsgálata nem csupán az eredeti osztályozási feladaton történik.

A standard és robusztus modellvarrás módszerével megállapítottam, hogy a robusztus és nem robusztus hálózatok reprezentációi az osztályozási feladat megoldása szempontjából teljesen kompatibilisek, azonban ez a robusztus osztályozás szempontjából csupán a bemenethez közeli (korai) rétegekre igaz. A korai rétegekről megmutattam, hogy a robusztusság eléréséhez sokkal fontosabbak, mint a mélyebb, kimenethez közelebbi rétegek, ezáltal feltárva a robusztus és nem robusztus hálózatok közötti legfőbb különbséget.

A dolgozatban vizsgáltam továbbá az osztályozó neurális hálózatok működését és megállapítottam, hogy robusztus és nem robusztus hálózatok egyaránt a reprezentációik klasztereződésével oldják meg az osztályozási feladatot. Mivel a klasztereződés, ezáltal a neurális hálózatok szemantikai rétegeinek működése a feladat struktúrájától függ, ezért egy megfelelően betanított hálózat kimenethez közeli rétegei újrafelhasználhatók egy hasonló struktúrájú új feladat megoldására, amellyel csökkenthető a neurális hálózatok tanításának költsége.

Köszönetnyilvánítás

Mindenekelőtt köszönetet mondok témavezetőmnek, Dr. Jelasity Márknak, akihez bizzalommal fordulhattam a kutatás során felmerülő összes kérdésemmel és aki nélkül ez a dolgozat nem jöhetett volna létre.

Jelen kutatás a Kulturális és Innovációs Minisztérium ÚNKP-22-2 – SZTE-343 kód-számú Új Nemzeti Kiválóság Programjának a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból finanszírozott szakmai támogatásával készült.

A Robusztus Mélytanulás projekt nevében köszönetet mondok az ELKH Cloud (lásd: Héder et al. 2022 [91]; <https://science-cloud.hu>) használatáért, ami hozzájárult a publikált eredmények eléréséhez.

A kutatás az Európai Unió támogatásával valósult meg, az RRF-2.3.1-21-2022-00004 azonosítójú, Mesterséges Intelligencia Nemzeti Laboratórium projekt keretében.

Irodalomjegyzék

- [1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [2] Oludare Isaac Abiodun, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed, and Humaira Arshad. State-of-the-art in artificial neural network applications: A survey. *Heliyon*, 4(11):e00938, 2018.
- [3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.
- [4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.
- [5] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.
- [6] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O’Donoghue, Daniel Visentin, George van den Driessche, Balaji Lakshminarayanan, Clemens Meyer, Faith Mackinder, Simon Bouton, Kareem Ayoub, Reena Chopra, Dominic King, Alan Karthikesalingam, Cían O Hughes, fRosalind Raine, Julian Hughes, Dawn A Sim, Catherine Egan, Adnan Tufail, Hugh Montgomery, Demis Hassabis, Geraint Rees, Trevor Back, Peng T Khaw, Mustafa Suleyman, Julien Cornebise, Pearse A Keane, and Olaf Ronneberger. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.*, 24(9):1342–1350, September 2018.

- [7] Geoffrey E. Hinton. How neural networks learn from experience. *Scientific American*, 267(3):144–151, 1992.
- [8] Hartmut Maennel, Ibrahim M Alabdulmohsin, Ilya O Tolstikhin, Robert Baldock, Olivier Bousquet, Sylvain Gelly, and Daniel Keysers. What do neural networks learn when trained with random labels? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19693–19704. Curran Associates, Inc., 2020.
- [9] Alsallakh Bilal, Amin Jourabloo, Mao Ye, Xiaoming Liu, and Liu Ren. Do convolutional neural networks learn class hierarchy? *IEEE Transactions on Visualization and Computer Graphics*, 24(1):152–162, 2018.
- [10] Rahul Parhi and Robert D. Nowak. What kinds of functions do deep neural networks learn? Insights from variational spline theory. *SIAM Journal on Mathematics of Data Science*, 4(2):464–489, 2022.
- [11] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-22, 2018*. The Internet Society, 2018.
- [12] Chaofei Yang, Qing Wu, Hai Li, and Yiran Chen. Generative poisoning attack method against neural networks, 2017.
- [13] Adnan Siraj Rakin, Zhezhi He, and Deliang Fan. Bit-flip attack: Crushing neural network with progressive bit search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1211–1220, 2019.
- [14] Zaixi Zhang, Jinyuan Jia, Binghui Wang, and Neil Zhenqiang Gong. Backdoor attacks to graph neural networks. In *Proceedings of the 26th ACM Symposium on Access Control Models and Technologies, SACMAT '21*, page 15–26, New York, NY, USA, 2021. Association for Computing Machinery.
- [15] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2013.
- [16] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings, 2015.
- [17] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world, 2016.

- [18] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.
- [19] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search, 2019.
- [20] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2014.
- [21] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses, 2017.
- [22] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2017.
- [23] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing, 2018.
- [24] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *ICLR*, 2020.
- [25] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy? – a comprehensive study on the robustness of 18 deep image classification models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [26] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy, 2018.
- [27] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C. Duchi, and Percy Liang. Adversarial training can hurt generalization, 2019.
- [28] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy, 2020.
- [29] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features, 2019.
- [30] Matteo Terzi, Alessandro Achille, Marco Maggipinto, and Gian Antonio Susto. Adversarial training reduces information and improves transferability. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3):2674–2682, May 2021.

- [31] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019.
- [32] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks, 2015.
- [33] Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack, 2019.
- [34] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.
- [35] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations, 2017.
- [36] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa, 2018.
- [37] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations, 2017.
- [38] Deng Pan, Xin Li, and Dongxiao Zhu. Explaining deep neural network models with adversarial gradient integration. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2876–2883. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.
- [39] Hyuk Jin Kwon, Hyung Il Koo, Jae Woong Soh, and Nam Ik Cho. Inverse-based approach to explaining and visualizing convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12, 2021.
- [40] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks, 2013.
- [41] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization, 2015.
- [42] Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120(3):233–255, may 2016.

- [43] Shunsuke Imai, Shin Kawai, and Hajime Nobuhara. Stepwise pathnet: a layer-by-layer knowledge-selection-based transfer learning algorithm. *Scientific Reports*, 10, 05 2020.
- [44] Linghua Zeng and Xinmei Tian. Layer-wise training to create efficient convolutional neural networks. In Derong Liu, Shengli Xie, Yuanqing Li, Dongbin Zhao, and El-Sayed M. El-Alfy, editors, *Neural Information Processing*, pages 631–641, Cham, 2017. Springer International Publishing.
- [45] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability, 2017.
- [46] Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation, 2018.
- [47] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited, 2019.
- [48] Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity with statistical testing, 2021.
- [49] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Dilan Gorur, Razvan Pascanu, and Hassan Ghasemzadeh. Linear mode connectivity in multitask and continual learning, 2020.
- [50] Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations, 2021.
- [51] Cinjon Resnick, Zeping Zhan, and Joan Bruna. Probing the state of the art: A critical look at visual representation evaluation, 2019.
- [52] Jessica A. F. Thompson, Yoshua Bengio, and Marc Schönwiesner. The effect of task and training on intermediate representations in convolutional neural networks revealed with modified RV similarity analysis. *CoRR*, abs/1912.02260, 2019.
- [53] Christian Cianfarani, Arjun Nitin Bhagoji, Vikash Sehwal, Ben Y. Zhao, Prateek Mittal, and Haitao Zheng. Understanding robust learning through the lens of representation similarities, 2022.
- [54] Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth, 2020.

- [55] Adrián Csiszárík, Péter Kőrösi-Szabó, Ákos K. Matszangosz, Gergely Papp, and Dániel Varga. Similarity and matching of neural network representations, 2021.
- [56] Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations?, 2015.
- [57] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence, 2014.
- [58] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L. Yuille, and Quoc V. Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [59] Ziv Goldfeld, Ewout van den Berg, Kristjan Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, and Yury Polyanskiy. Estimating information flow in deep neural networks, 2018.
- [60] Francesco Croce, Maksym Andriushchenko, Vikash Sehwal, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.
- [61] Xiaofeng Mao, Yuefeng Chen, Xiaodan Li, Gege Qi, Ranjie Duan, Rong Zhang, and Hui Xue. Easyrobust: A comprehensive and easy-to-use toolkit for robust computer vision. <https://github.com/alibaba/easyrobust>, 2022.
- [62] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]*. Available: <http://yann.lecun.com/exdb/mnist>, 2, 2010.
- [63] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.
- [64] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011.
- [65] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.
- [66] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban

- Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.
- [67] Alessandro Achille, Giovanni Paolini, and Stefano Soatto. Where is the information in a deep neural network?, 2019.
- [68] Sebastian Thrun and Tom M. Mitchell. Lifelong robot learning. *Robotics and Autonomous Systems*, 15(1):25–46, 1995. The Biology and Technology of Intelligent Autonomous Agents.
- [69] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.
- [70] Liwei Wang, Lunjia Hu, Jiayuan Gu, Yue Wu, Zhiqiang Hu, Kun He, and John Hopcroft. Towards understanding learning representations: To what extent do different neural networks learn the same representation, 2018.
- [71] R. Penrose. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3):406–413, 1955.
- [72] Haydn Jones, Jacob M. Springer, Garrett T. Kenyon, and Juston Moore. If you’ve trained one you’ve trained them all: Inter-architecture similarity increases with robustness. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.
- [73] David McNeely-White, J. Ross Beveridge, and Bruce A. Draper. Inception and resnet features are (almost) equivalent. *Cognitive Systems Research*, 59:312–318, 2020.
- [74] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.
- [75] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
- [76] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

- [77] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks, 2016.
- [78] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks, 2016.
- [79] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.
- [80] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, and Ben Edwards. Adversarial robustness toolbox v1.2.0. *CoRR*, 1807.01069, 2018.
- [81] Sravanti Addepalli, Samyak Jain, Gaurang Sriramanan, Shivangi Khare, and Venkatesh Babu Radhakrishnan. Towards achieving adversarial robustness beyond perceptual limits. In *ICML 2021 Workshop on Adversarial Machine Learning*, 2021.
- [82] Vikash Sehwal, Saeed Mahloujifar, Tinashe Handina, Sihui Dai, Chong Xiang, Mung Chiang, and Prateek Mittal. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? In *International Conference on Learning Representations*, 2022.
- [83] Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. Helper-based adversarial training: Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *ICML 2021 Workshop on Adversarial Machine Learning*, 2021.
- [84] Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019.
- [85] Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy Mann. Improving robustness using generated data, 2021.
- [86] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on mnist, 2018.
- [87] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method, 2000.
- [88] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.
- [89] Shucong Zhang, Cong-Thanh Do, Rama Doddipatla, Erfan Loweimi, Peter Bell, and Steve Renals. Train your classifier first: Cascade neural networks training from upper layers to lower layers, 2021.

- [90] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [91] Mihály Héder, Ernő Rigó, Dorottya Medgyesi, Róbert Lovas, Szabolcs Tenczer, Ferenc Török, Attila Farkas, Márk Emődi, József Kadlecsek, György Mező, Ádám Pintér, and Péter Kacsuk. The past, present and future of the ELKH cloud. *Információs Társadalom*, 22(2):128, aug 2022.