Önálló laboratórium beszámoló

Távközlési és Médiainformatikai Tanszék

készítette: **Hujbert Patrik**
patrik.hujbert@gmail.com
neptun-kód: **D83AE5**
ágazat: **mérnökinformatikus szak**
konzulens: **Unyi Dániel**
unyi.daniel@tmit.bme.hu
konzulens: **Dr. Gyires-Tóth Bálint**
toth.b@tmit.bme.hu

Téma címe: GNN Explainability - egy neuronháló mely részei fontosak?

Feladat:
Jelenleg sokan tartják a neurális hálózatok magyarázhatóságát (explainibility) a legfontosabb deep learning kutatási iránynak. Feladatom is ezen témakörbe tartozik: a neurális hálózat mely részeit használja a predikcióhoz. Célom egy már létező módszer ötletes felhasználása a hálózatok magyarázhatóságának vizsgálatára. Az említett módszer a MARGIN nevet viseli, ami egy egyszerű, black-box megközelítés a neurális hálózatok értelmezhetőségére. Projektem első lépése ezen megoldás egy specifikus esetének implementálása: egy konvolúciós neurális hálózattal (CNN) végzett osztályozási feladat esetén a bemeneti kép mely részei járulnak hozzá a legnagyobb mértékben az osztályozási döntéshez. Ezt követően az analízis módosítható, oly módon, hogy az osztályozási feladat során gráf bemenetek legyenek vizsgálhatók, ez esetben az osztályozást természetesen egy gráf neurális hálózat végzi. A módosítás bár aprónak tűnik, megvalósítása korántsem magától értetődő, így aztán feladatom a projekt ezen első két szakaszára fókuszált. Jelentőségét mutatja, hogy az immár módosított eljárással neurális hálózatok is vizsgálhatók, kihasználva a tényt, hogy maguk a neurális hálózatok is gráfok. A hálózat bizonyos részeinek fontossága az alapján adható meg, hanem hogy egyes részek maszkolása mekkora különbséget jelent a predikcióban. Ez projektem egy későbbi szakaszát képezi.

Tanév: 2022/23. tanév, II. félév

1. A laboratóriumi munka környezetének ismertetése, a munka előzményei és kiindulási állapota

1.1. Bevezető

A deep learning, vagyis mélytanulás, robbanásszerű fejlődése számos alkalmazási területen nagy hatást gyakorol. Azonban a mélytanulás alkalmazása megannyi kihívással jár, különösen a neurális hálózatok magyarázhatósága (explainability) és értelmezhetősége (interpretability) tekintetében. Ennek fő oka, hogy a neurális hálózatok fekete dobozként működnek, ami következményként azt vonja maga után, hogy meglehetősen nehéz megmondani, hogyan hoznak döntéseket. Ennek a kérdéskörnek a feltérképezése és mélyebb megértése a mélytanulás kutatók körében jelenleg egy fontos feladat, saját munkám is ezen törekvések erősítését igyekszik szolgálni.

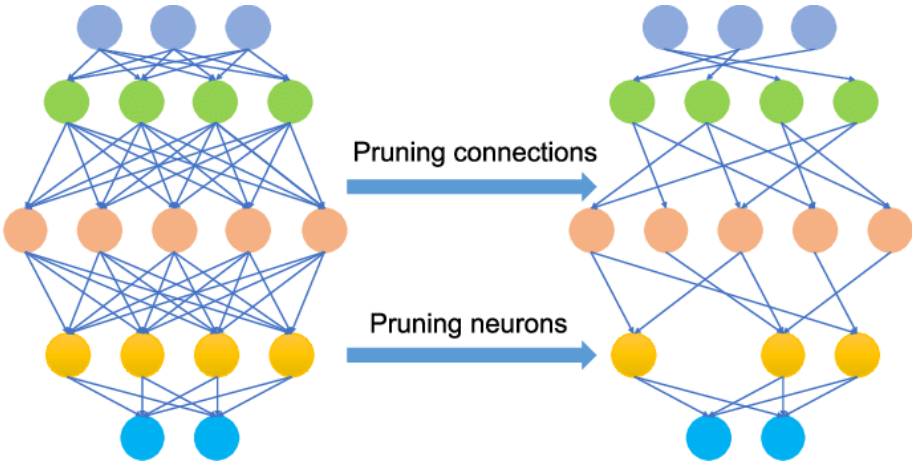
1.2. Elméleti összefoglaló

A bevezetőben említett mély neurális hálózatok magyarázhatóságának hiányossága problémát jelenthet olyan kritikus alkalmazásokban, mint például az orvosi diagnosztika vagy az autonóm járművek. Az értelmezhetőség növelése bizalmat épít a rendszer iránt, segítve a felhasználókat és a szakembereket abban, hogy megbízzanak a hálózatok döntéseiben. Az értelmezhetőség továbbá hozzájárul az AI rendszerek felelősségteljes felhasználásához. Ha egy rendszer nem képes megmagyarázni vagy értelmezni döntéseit, akkor nehéz megérteni, hogy miért hibázik vagy milyen tényezők befolyásolják a kimenetet.

Az előbb felsorolt aspektusok mind a felhasználói oldalról közelítik meg a problémát, mindazonáltal az értelmezhetőség jelentősége épp oly nagy a mélytanulási tudományos kutatások, vizsgálatok során is. A tudósoknak meg kell érteniük a hálózatok működését ahhoz, hogy hatékonyabban fejleszthessenek új architektúrákat és algoritmusokat. Az értelmezhetőség lehetővé teszi a kutatók számára, hogy feltárják a hálózatok működésének rejtett mechanizmusait, és mélyebb betekintést nyerjenek a gépi tanulás folyamatába.

A mélytanulás területén a tendencia azt mutatja, hogy jobb eredményeket nem is feltétlenül forradalmi új ötletekkel, architektúrákkal, eljárásokkal lehet elérni. A tapasztalat az, hogy ennél jelenleg sokkal fontosabb a számítási kapacitás növelése. Míg a matematikai háttér és a kitalált módszerek már egy jóideje rendelkezésünkre állnak, az eredményekben mégis jelentős javulás látható. Ökölszabályként tehát elmondható, hogy a modellek és a tanító adathalmaz növelése szinte mindig jobban teljesítő modellt eredményez. A modellek méretének tendenciózus növekedése azonban több negatív következményt is von maga után, éppen ezért igencsak aktív és felkapott kutatási iránnyá vált a mélytanulás közösségben a hálózatok méretének visszaszorítása. Erre egy megoldás a pruning.

A neurális hálózatok metszése, vagyis a pruning, egy olyan technika a gépi tanulásban és neurális hálózatok optimalizálásában, amelynek során a hálózatból eltávolításra kerülnek a kevésbé fontos vagy felesleges paraméterek, súlyok vagy neuronok, ezáltal csökkentve a hálózat méretét, komplexitását.



1. ábra. Egy többrétegű neurális hálózat metszése [1]

A metszés fontos több okból:

- Kompaktabb és hatékonyabb hálózatok: A metszés lehetővé teszi a hálózat méretének csökkentését, amely jelentős előnyöket kínál. Kompaktabb hálózatok kevesebb erőforrást igényelnek a tároláshoz és az értékeléshez, ami kisebb memóriahasználatot és gyorsabb futási időt eredményezhet.
- Értelmezhetőség és magyarázhatóság: A metszés általában egyszerűsíti a hálózat struktúráját, és kisebb számú fontos elemet hagy meg. Ezáltal javul az értelmezhetőség és a magyarázhatóság, mivel könnyebben megérthetővé válnak a hálózat működésének kulcsfontosságú elemei.
- Túltanulás csökkentése: A nagyobb méretű hálózatok hajlamosak a túltanulásra, vagyis túl sokat tanulnak az adathalmazból, és nehezebben általánosíthatnak új adatokra. A metszés révén csökkenthetjük a túltanulást, mivel megszabadulunk a felesleges paraméterektől, amelyek hozzájárulhatnak a túltanuláshoz.
- Részleges újratanulás: A metszés utáni finomhangolás révén a hálózat újra tanulhat olyan változásokra, amelyek következtében eltávolították a paramétereket vagy neuronokat. Ez lehetővé teszi a hálózat számára, hogy alkalmazkodjon a metszés után keletkező új adatokhoz vagy feladatokhoz.

Összességében tehát elmondható, hogy a neurális hálózatok metszése hatékony módszer a hálózatok optimalizálására, méretük csökkentésére és a túltanulás kezelésére. Ezzel javítható a hálózatok hatékonysága, értelmezhetősége és alkalmazhatósága különböző gépi tanulási feladatokban.

Projektem fő feladat egy ilyen metszési eljárás kidolgozása.

1.3. A munka állapota, készültségi foka a félév elején

Mint azt már korábban említettem a Magyarázható MI egy módfelett izgalmas és jelentős téma így aztán a tanszéken is fontos ennek kutatása. Projektem jelen félév munkájával kezdődött, így nincs ezt megelőző előzménye. A tanszéken sincs olyan korábbi munka ami szerves részét képezné a kutatásomnak, tehát onnan sem használtam fel korábbi segítséget. A kutatás alapját konzulensem, Unyi Dániel, ötlete jelenti, így őt illeti az érdem.

2. Az elvégzett munka és az eredmények ismertetése

Projektem során a félév alatt több fő állomáson vezetett az utam, ebben a fejezetben azt szeretném bemutatni, hogy pontosan milyen alfeladatokat végeztem el. Első körben a neurális hálózatok magyarázhatóságával és metszésével ismerkedtem meg, mint általánosabb témakör. Ezt követően a projektem kiindulásául szolgáló cikkről és keretrendszerről tanultam, amit aztán implementáltam is saját magam a cikkben leírtak alapján, egy konkrét esettanulmányra. Az esettanulmányban a feladat egy képek osztályozását ellátó konvolúciós neurális hálózat értelmezhetőségének vizsgálata. A félév utolsó szakaszában pedig ennek módosításával foglalatoskodtam, vagyis hogyan lehet a keretrendszert felhasználni gráfok osztályozásának az értelmezhetőségére.

2.1. The Lottery Ticket Hyphotesis

A kutatásom kezdete során fontosnak tartottam feltérképezni a neurális hálózatok metszésére szolgáló jelenleg state-of-the-art módszereket és eljárásokat, melyek nem képezik feltétlenül szerves részét a saját megoldásomnak, azonban ismeretükkel nagyobb rálátást nyerhettem a teljes témakörre, annak mikéntjére, jelentőségére.

A megismert módszerek közül a The Lottery Ticket Hypothesis című cikk [2] tekinthető a legfontosabbnak, ezért én is ezzel foglalkoztam a legtöbbet, és az ebből levont konzekvenciákat szeretném a következőekben részletezni. A cikk szerzői, Jonathan Frankle és Michael Carbin, új megközelítést javasolnak a mély neurális hálózatok hatékonyabb és értelmezhetőbb kialakítására. Egy olyan módszert dolgoztak ki, amivel a neurális hálózatok méretét töredékére tudják csökkenteni, oly módon, hogy az nem veszít a pontosságából. A modell paramétereinek 90 százalékos csökkentésével kisebb tárhely és kevesebb számítási kapacitás szükséges a modell használata során. A cikkben azonban a metszett hálózatok egy általánosabb problémájára is megoldást találnak, miszerint általánosságban elmondható, hogy a metszett hálózatok nehezen taníthatók. Márpedig a számítási igény a tanítás során is roppant mód fontos.

A hipotézis állítása szerint a véletlenül inicializált, többrétegű, előrecsatolt hálózatok tartalmaznak olyan alhálózatokat, "nyereményjegyeket", amiket önmagukban tanítva épp ugyan olyan pontosság érhető el a teszt adatokon mint a teljes betanított hálózat esetén.

A nyertes alhálózatok az inicializációjuk miatt különlegesek. A hipotézis szerint ezek azok a paraméterek a hálózatban, amik a helyes inicializációjuk miatt a tanítást effektívvé teszik.

A The Lottery Ticket Hypothesis lényege, hogy ha képesek vagyunk azonosítani ezeket a kulcsfontosságú "nyereményjegyeket", akkor lehetséges a hálózatot kisebb méretűre metszeni a felesleges paraméterek eltávolításával. Ezáltal csökken a hálózat mérete, de a teljesítmény és a pontosság megmarad.

A cikkben a szerzők bemutatják a metszett hálózatok hatékonyságát és a reprodukálható eredményeket, számos különböző gépi tanulási feladaton és adathalmazon. Emellett részletesen tárgyalják a metszési módszereket és az optimalizációs technikákat a "nyereményjegyek" azonosítására és a hálózatok hatékony újratanítására.

A cikk tanulsága tehát abban nyilvánult meg számomra, hogy a neurális hálózatok metszésének két lényegesen eltérő szempontja van:

- Inference modell optimalizálása: A már betanított nagy méretű és számításigényű modell méretének csökkentése, hogy a modell használhatósága javuljon.
- Tanítás optimalizálása: Már a tanítás kezdetén egy olyan metszett alhálózat találása, amely könnyebben és gyorsabban tanítható mint az eredeti hálózat, azonban ugyan olyan eredményt ér el a tanítás után. Ennek természetesen következménye az is, hogy az inference modell is már egy metszett hálózat.

Saját projektem tehát ebből a szemszögből vizsgálva egy kisebb problémát fed le, mint a cikkben leírt eljárás, ugyanis az én módszerem egy már betanított hálózatot vizsgál. Csak erre a betanított hálózatra tudja megmondani, hogy mely részei fontosak, melyek nem. Vagyis a tanítás optimalizálására nem terjed ki a megoldásom.

2.2. MARGIN: Uncovering Deep Neural Networks using Graph Signal Analysis

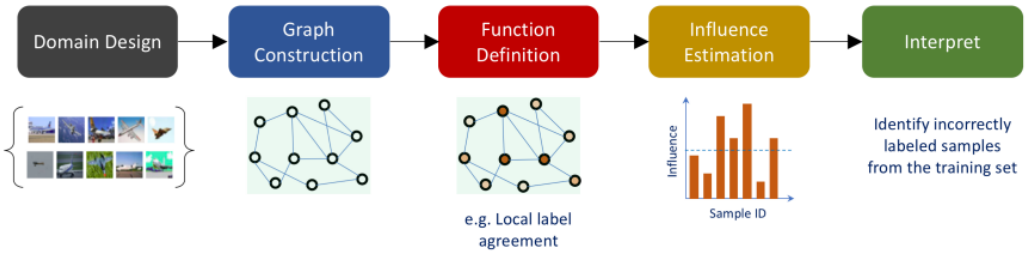
Megoldásom a "MARGIN: Uncovering Deep Neural Networks using Graph Signal Analysis" című cikkben [3] bemutatott Model Analysis and Reasoning using Graph-based Interpretability (MARGIN) eljárást

használja fel, így aztán a cikk beható tanulmányozása fontos részét képezte a kutatásomnak. Beszámolóm ezen szekciójában ezt a cikket szeretném bemutatni.

A szerzők kiindulási alapját az képezte, hogy bár az értelmezhetőség egy meghatározó irányzat a mélytanulás területén, nincs egy jól elfogadott definíciója. Így aztán megközelítéstől függően számos dolgot jelenthet, éppen ezért több különböző alfeladat is a témakör alá tartozik. Néhány ilyen feladat lehet például anomáliák meghatározása a tanító adathalmazban, modellek debuggolása, vagy épp a számomra releváns feladat: meghatározni, hogy egy konkrét bemenet esetén a bemenet mely részei voltak a legfontosabbak a predikció előállításához. Ugyan ezekre már mind léteztek a cikk előtt is megoldások, de ezek egytől egyig probléma specifikusak voltak. A szerzők felvetése szerint viszont kezelhetők egységesen, hiszen végső soron mindegyik feladat mögött ugyan az a kérdés húzódik meg: A relatív változások milyen mértékben befolyásolják a modell predikcióját, legyen az a változás lokális vagy globális. Erre a feladatra nyújt megoldást a MARGIN keretrendszer.

Az eljárás során az értelmezési feladatot egy hipotézisvizsgálatként kell kezelni, aminek a viszonylatában a metódus megad egy befolyásolási mértéket arra, hogy a feladat bemenetének melyik része támasztja alá a leginkább a hipotézist. Ez pontosítva annyit jelent, hogy minden bemenetből egy olyan gráfot lehet előállítani, amelynek a csúcsai az értelmezendő adat, az éleit pedig a hipotézist leíró függvény adja meg. A gráfon ezután egy gráf jel analízist (graph signal analysis) hajt végre az eljárás, ami végül a becsült befolyásolásokat szolgáltatja. A MARGIN eljárás lépései tehát a következők:

1. A bemeneti tartomány (domain) meghatározása
2. A gráf csúcsainak meghatározása
3. A hipotézis függvény definiálása
4. Gráf jel analízis
5. Magyarázatok készítése az analízis kimenet alapján



2. ábra. A MARGIN lépései [3]

2.3. CNN Explainability: Predikció értelmezhetősége képek osztályozására

A MARGIN keretrendszer megismerése után következő feladatom a keretrendszer egy konkrét esettanulmányának az implemetálása volt: predikciók értelmezhetősége képek osztályozása esetén. Az esettanulmány ugyanis egy kiváló kiindulási alap, hogy a projekt egy későbbi fázisában eljussak egy olyan módszerhez, amivel nem csak képek osztályozási feladata esetén tudom alkalmazni a MARGIN-t, hanem konkrét neurális hálózatok lehetnek a MARGIN bemenetei. Ebben az alfejezetben az implementáció menetét szeretném részletezni.

Az implementáció technikai információi: A Google Colab fejlesztőkörnyezetében dolgoztam, ahol python nyelven írtam a kódot, felhasználva a mélytanulásos projektek során leggyakrabban használt könyvtárakat, mint például a pytorch, torchvision, scikit-learn és matplotlib.

Mivel az értelmezendő predikció képek osztályozása, így a MARGIN implementálása előtt két dologra is szükségem volt:

1. Képek osztályozására alkalmas adatbázis

2. A kiválasztott adatbázis képeinek osztályozására alkalmas konvolúciós neurális hálózat.

Az adathalmaz és a modell megválasztása során fontos szempont volt számomra, hogy már egy biztosan jól működő párost válasszak, hiszen ebben az esetben a legszemléletesebb a MARGIN működése, és implementálása során nagyobb fókuszt helyezhettem a konkrét implementálásra. Továbbá egy nagy modellt szerettem volna választani, ami nagy számú osztályba tudja sorolni a képeket, mivel természetesen bonyolultabb modellek esetén még érdekesebb az értelmezhetőség kérdése. A választásom ezen okokból kifolyólag a méltán híres ImageNet [4] adathalmazra és AlexNet mélytanulásos modellre esett. Pontosabban az ImageNet adathalmaznak csak egy részhalmazát használtam, ez a Stanford Dogs Dataset [5]. Az adathalmaz az ImageNet képeiből és annotációból lett felépítve és 120 különböző kutyafajtáról tartalmaz 20580 képet.

Az adathalmaz és a predikciót elvégző hálózat megválasztása után a következő lépés a MARGIN eljárás általános lépéseinek specifikálása volt a jelenlegi konkrét esetre. Ezt szeretném először összefoglalni, majd az egyes lépéseket kifejteni. A lépések összefoglalva:

1. A bemeneti tartomány: Az adathalmaz egy bizonyos képe
2. A gráf csúcsai: A képből előállított szuperpixelek
3. A hipotézis függvény: Az egyes szuperpixelek relatív fontossága
4. Gráf jel analízis
5. Magyarázat: A képből előállított Dense Saliency Map és a gráf jel analízis kimenetének Hadamard-szorzata

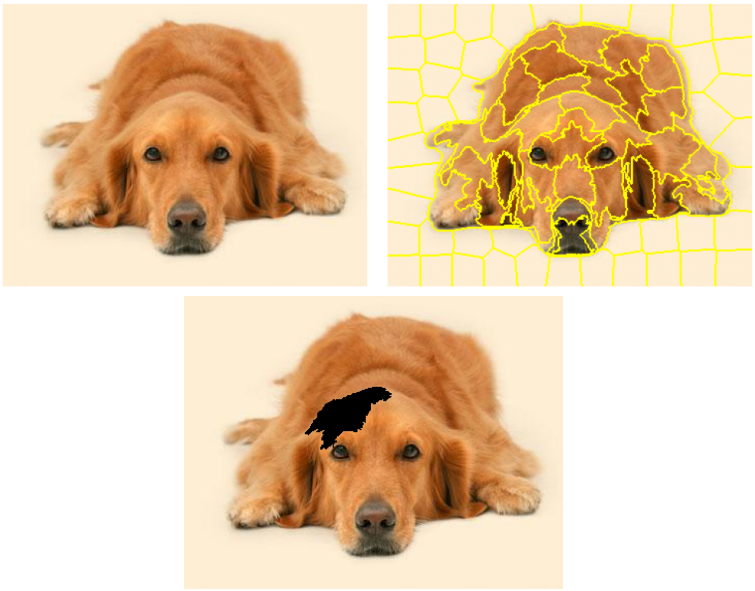
A bemeneti tartomány meghatározása természetesen egy magától értetődő lépés volt, szimplán kiválasztottam egy képet a Stanford Dogs adathalmazból. A szomszédossági gráf csúcsainak meghatározásához a képet szuperpixelekre bontottam fel. Ehhez a SLIC algoritmust [6] használtam segítségként. Annak érdekében, hogy minnél több és különfajta szuperpixel álljon a rendelkezésemre, az algoritmust többször is futtatam, oly módon, hogy különböző számú szuperpixelre bontsa fel a képet. Ezek a pixelszámok: $[50, 100, 150, 200, 250, 300]$. A hipotézis függvény értékeinek kiszámolásához a neurális hálózat predikcióira volt szükségem az egyes szuperpixelek által kimaszkolt képen. Ahhoz, hogy a predikciókat hatékonyan tudjam kiszámolni, egy saját Pytorch Dataset-et hoztam létre, amely tartalmazza az eredeti képet és a SLIC algoritmus által meghatározott szuperpixeleket. Az adathalmaz egyes elemei pedig az egyes szuperpixelek által kimaszkolt eredeti kép.

```
class MaskedImage(Dataset):
    def __init__(self, og_img, img_label, transform=None,
                 segment_list = [50, 100, 150, 200, 250, 300]):
        self.og_img = og_img
        self.img_label = img_label
        self.transform = transform
        self.super_pixels = []
        img_array = np.asarray(self.og_img)
        for n_segments in segment_list:
            segments = slic(img_array, n_segments=n_segments)
            segment_values = np.unique(segments)
            for segment_value in segment_values:
                segment_mask = segments == segment_value
                self.super_pixels.append(segment_mask)

    def __len__(self):
        return len(self.super_pixels)

    def __getitem__(self, idx):
        mask = self.super_pixels[idx]
        img_array = np.asarray(self.og_img)
        img_array[mask] = 0
```

```
masked_img = Image.fromarray(img_array)
if self.transform:
    masked_img = self.transform(masked_img)
return masked_img
```



3. ábra. Balról jobbra: eredeti kép, szuperpixelekre felbontott kép, egy szuperpixellel maszkolt kép

A maszkolt képeken történő predikció után már minden fontos adat rendelkezésre állt számomra ahhoz, hogy kiszámítsam a szomszédossági mátrixát a bemeneti gráfnak, és a hipotézis függvény értékeit. A szomszédosségi mátrix értékeinek meghatározásához először is szükségem volt az egyes szuperpixelek saliency értékére. Ez tulajdonképpen annak a mértéke, hogy az adott szuperpixel milyen mértékben járul hozzá a helyes kimeneti predikció értékéhez. Mivel a módszer teljes mértékben fekete doboz megoldás, ezért a saliency értékek kiszámításához a

$$|p_i(x_{img}) - p_i(img)| \tag{1}$$

képletet használtam, ahol x_{img} jelöli az x. szuperpixellel maszkolt képet, az img jelöli az eredeti képet, a $p_i()$ medig a konvolúciós hálózat i. osztályhoz tartozó softmax értékét jelöli, ha a paramétere a hálózat bemenete. Az élek értékei innen már könnyen meghatározhatók, az

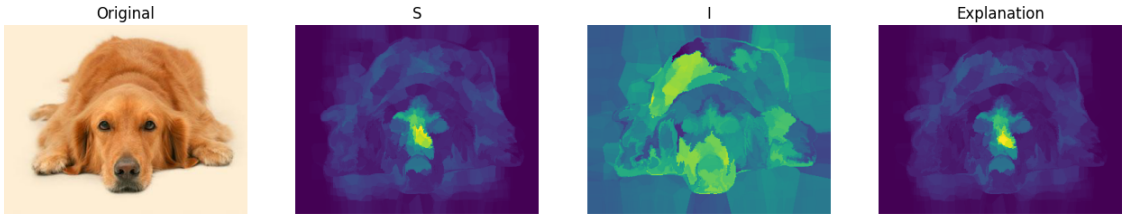
$$|s_x - s_y| \tag{2}$$

képlettel, ahol x és y az élhez tartozó két csúcs és s_j az j. szuperpixellel maszkolt kép saliency értéke. A hipotézis függvény értékeit az egyes szuperpixelek relatív méretével adtam meg, vagyis elosztottam a szuperpixel méretét a legnagyobb szuperpixel méretével. A MARGIN eljárás tehát azon csúcsokat keresi a gráfban az analízis során, amelyeknek a relatív mérete a legkisebb. Intuitívan a MARGIN eljárás biztosítja számomra, hogy a végső kimenet a lehető legritkább (sparse) legyen. Ez egy meghatározó aspektusa a metódusnak, ugyanis az értelmezés során nem csak az a fontos, hogy a legnagyobb saliency értékkel rendelkező részeit találja meg a bemeneti képnek, hanem az is, hogy ez minnél ritkább legyen. Ez a későbbiekben még fontosabb szerepet fog játszani, hiszen a neurális hálózatok metszése esetén pontosan az előbbiekben leírt jellemzőkkel rendelkező alhálózat a keresendő.

A gráf jel analízishez így minden bemenet rendelkezésre áll már, az algoritmus kódját a cikkből emeltem át.

A metódus utolsó lépése a tényleges kimenet előállítása. A kimenet a gráf jel analízis kimenetének és a saliency értékek (1 képlet eredményei) alapján előállított Dense Saliency Map Hadamard-szorzata:

$$S_{final} = S \odot I \tag{3}$$



4. ábra. Magyarázat, hogy a kép mely része a legfontosabb (és egyben a legritkább) ahhoz, hogy a hálózat predikciója megegyezzen az eredeti kép cimkéjével

Maszkolás	42.3%	45%	88.64%
Top 5 predikció			
1.	Golden retriever 79.5%	Labrador retriever 21.19%	Pencil sharpener 72.38%
2.	Labrador retriever 4.21%	Toyshop 20.24%	Conch 7.2%
3.	Tibetan mastiff 3.14%	Dogsled 9.98%	Jellyfish 6.9%
4.	Irish terrier 2.24%	Golden retriever 8.83%	Lampshade 2.87%
5.	Chow 1.9%	Irish terrier 5.03%	Torch 1.61%

1. táblázat. A különböző mértékben maszkolt képekre adott predikció az AlexNet modellel

ahol S a saliency map, I a MARGIN kimenete, $\odot$ pedig a Hadamard-szorzat.

Az implementáció elkészülte után teszteltem az eljárást. A fő kérdés, amire igazán kíváncsi voltam, hogy az eredményül kapott magyarázat alapján vajon az eredeti kép hány százalékát lehet maszkolni ahhoz, hogy még mindig jó predikciót adjon a modell. Ez a kérdés természetesen azért fontos, mert a későbbi módosítások során a neurális hálókat épp így szeretném majd metszeni, mint ahogy most a képet maszkolni. Tehát a most elért eredmények iránymutatóak lehetnek, hogy a jövőben mennyire lehet sikeres a projektem. A tesztelés alatt egy bemeneti képet próbáltam ki, amit az értelmezés alapján különböző mértékben maszkoltam (lásd 5 ábra). Mint azt az 1 táblázat is mutatja, a kép 42%-ig teljesen maszkolható az eljárás eredménye alapján, a predikció az eredeti 88%-hoz képest csak kis mértékben csökken, és a hálózat még mindig nagy magabiztossággal találja el a kép osztályát (golden retriever). 45%-os maszkolás esetén a helyes predikció még bent van a top 5 eredményben, ami az ImageNet adathalmaz esetén még egy jó eredménynek mondható, sőt a többi predikció is majdnem mind kutya, tehát itt még egész értelmezhető eredményt ad a modell. Azonban azt is figyelembe kell venni, hogy ez pont az a határ, amikor már nem a kép kutyán kívüli részei kerülnek maszkolásra, hanem maga a kutya is. 45% felett már teljesen rossz predikciókat ad a hálózat. Konklúzióként elmondható, hogy az implementált módszer működőképes, így továbbléptem projektem következő alfeladatára.

5. ábra. Maszkolt képek balról jobbra: 42.3%, 45% és 88.64% maszkolva

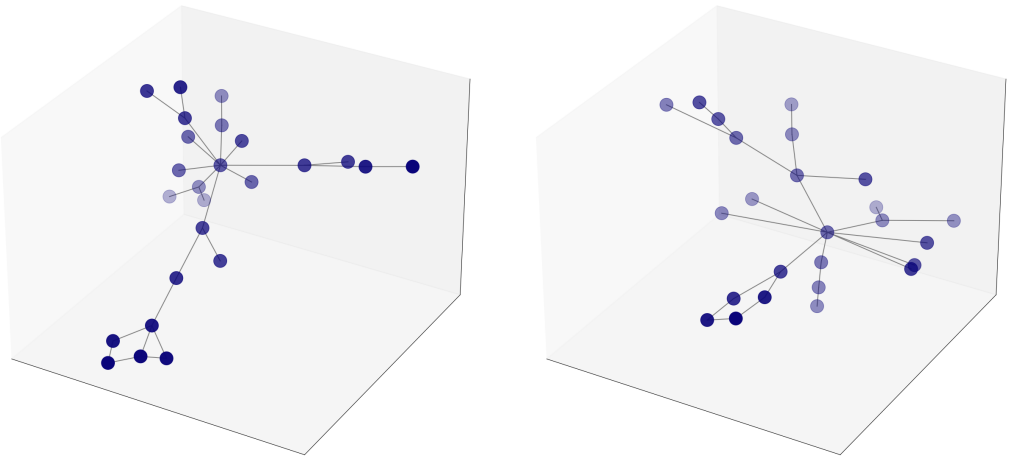
2.4. GNN Explainability: Predikció értelmezhetősége gráfok osztályozására

Ebben az alfejezetben a projekt következő állomásáról szeretnék beszélni, ami az előzőekben taglalt MARGIN eljárás módosítása. Ahhoz, hogy egy lépéssel közelebb kerüljek a végcélhoz, szükségem volt arra, hogy az implementált eljárást ne csak képi bemenetre tudjam alkalmazni, hanem gráfokra is. Így ennek az

alfeladatnak a célja az, hogy lecseréljem az előzőekben használt konvolúciós hálózatot egy gráfok osztályozására alkalmas gráf neurális hálózatra. Ehhez természetesen szükségem volt egy új adathalmazra, ami alkalmas gráfok osztályozásásra. Ennél fogva az implemetálást most is kutatómunka előzte meg. Minthogy a magyarázhatóság témaköre meglehetősen felkapott a mélytanulásos világban, több olyan cikk jelent meg az elmúlt években, amely gráf neurális hálózatok magyarázhatóságával foglalkozik. Ezek közül az egyik az Explainability in Graph Neural Networks. A Taxomonic Survey [7] című cikk, amiben a szerzők a már létező technikákat és módszereket foglalják össze. Itt taglalnak több, külön erre a specifikus feladatra létrehozott szintetikus adathalmazt is, közülük pár:

- BA-Shapes: csúcs klasszifikációs adathalmaz 4 osztállyal
- BA-Community: csúcs klasszifikációs adathalmaz 8 osztállyal
- Tree-Cycle: csúcs klasszifikációs adathalmaz 2 osztállyal
- Tree-Grids: csúcs klasszifikációs adathalmaz 2 osztállyal
- BA-2Motifs: gráf klasszifikációs adathalmaz 2 osztállyal

Annak érdekében, hogy könnyebb legyen az átállás képekről gráfokra, számomra egy gráf klasszifikációs feladat volt a leginkább kézenfekvő, szemben egy csúcs klasszifikációval, ezért a BA-2Motifs-ra esett a választásom. Az adathalmaz számszerűen 1000 gráfot tartalmaz, és mindegyik gráf egy egyszerű BA gráfból és egy ehhez csatolt különleges mintázatú gráfból áll. A két különböző mintázat, vagyis a ház (house) és a kör (cycle) határozzák meg, hogy a gráf melyik osztályba tartozik (6 ábra).



6. ábra. Ház és kör mintázatú gráfok a BA-2Motif adathalmazból

Az adathalmaz választása után egy megfelelő modell architektúrát kellett választanom a gráf klasszifikáció elvégzéséhez. Két különböző modellt is választottam, az egyiket saját magam tanítottam be az adathalmazon, a másik pedig már egy előre betanított volt. Az általam tanított modell három Graph Isomorphism Network (GIN) rétegből épül fel, egy-egy réteg sorban lineáris, batch normalizásiós, relu aktivációs, lineáris és relu aktivációs réteg szekvenciáját tartalmazza. A tanítást követően a hálózat 100%-os eredményt ért el a teszt adathalmazon. A készen kapott modell egy korábbi, szintén a gráf neurális hálózatok magyarázhatóságáról készült cikkből [8] implementáltam. A hálózat 3 gráf konvolúciós rétegből áll, mindhárom rétegre L2 normalizációt alkalmazva. A bemeneti dimenzió az adathalmaz dimenzionalitásából kifolyólag 10, míg a rejtett rétegek dimenziója egyenként 20-20. A cikk szerzői a hálózatot be is tanították a BA-2Motifs adathalmazra. A betöltött súlyparaméterekkel 98%-os pontosságú osztályozást ér el a modell. A megfelelő adathalmaz és GCN modell párossal minden készen áll ahhoz, hogy a MARGIN eljárást implementálni tudjam gráfok vizsgálatára. Ez a jövőbeli terveim részét képezi.

2.5. Összefoglalás

A félév során egy nagyobb volumenűre tervezett projekt első lépéseivel foglalkoztam. A projekt a neurális hálózatok értelmezhetőségének, magyarázhatóságának kérdésével foglalkozik. Arra próbál egy új módszert adni, hogyan értelmezhető egy neurális hálózat működése. Ezen kérdéskörön belül pedig a fő feladata annak meghatározása, hogy a vizsgált hálózat mely részei a legfontosabbak a predikció előállításához. Ennek meghatarározása lehetőséget nyújt a hálózat méretének csökkentésére, vagyis metszésére.

A mélytanulás területén igazán jelentős kutatási terület a neurális hálózatok értelmezhetősége és metszése. Ennek fő indoka, azon bizonyos tendecia ellensúlyozása, miszerint az igazán áttörő eredményeket a számítási kapacitásnak és a mélytanulásos modellek, adathalmazok méretének növelése okozza. Kutatások eredményei azoban azt mutatják, hogy feltételezhetőleg azért van szükség nagy modellekre, hogy nagyobb esély legyen megtalálni egy olyan kisebb alhálózatát, ami önmagában is alkalmas feladatának elvégzéséhez. Következményként a modellek mérete visszaszorítható, ami rengetek előnnyel jár. Mindezen okok miatt gondolom, hogy fontos és igazán érdekes ezzel a témával foglalkozni.

Saját megoldásom kidolgozása során egy már létező értelmezhetőségi eljárás, a MARGIN implementálásán és módosításán tevékenykedtem. Ez a bizonyos eljárás egy egységes keretrendszert biztosít értelmezhetőségi feladatok ellátására, így módosításával és ügyes alkalmazásával használható neurális hálózatok metszésére is. Összefoglalva a félév alatt elvégzett munkámat: Implementáltam a MARGIN keretrendszer felhasználásával egy értelmezhetőségi feladatot. A feladat: értelmezést adni, hogy egy konvolúciós neurális hálózattal osztályozott kép mely részei járultak hozzá a leginkább a helyes predikció előállításához. Ezt követően azon dolgoztam, hogyan tudom ezt a megoldást módosítani úgy, hogy az gráfok osztályozását vizsgálja. A projekt jövőbeli tervei közé tartozik tényleges neurális hálózatok vizsgálata a módosított modellel, kihasználva, hogy azok szintén gráfok. Ha a vizsgálat sikeres és a módszer segítségével ténylegesen lehet metszeni a vizsgált hálózatokon akkor ezt kihasználva be lehet tanítani egy újabb hálózatot arra, hogy mely részeket érdemes lemetszeni az adott hálózatokból.

A MARGIN implementálása során elért eredmények azt mutatják, hogy a vizsgált bemeneteknek tényleg elég csak egy kisebb részhalmaza a megfelelő pontosságú predikció eléréséhez. Ez azt jelenti, hogy érdemes a projekt további részével foglalkozni, mert a módszer jó potenciált tartogat, hogy alkalmas legyen neurális hálózatok metszésére.

3. Irodalom, és csatlakozó dokumentumok jegyzéke

3.1. A tanulmányozott irodalom jegyzéke

[1] Chen, Jiasi & Ran, Xukan. (2019). *Deep Learning With Edge Computing: A Review*. Proceedings of the IEEE. PP. 1-20. 10.1109/JPROC.2019.2921977.

[2] Jonathan Frankle, Michael Carbin: *The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks*, 2018, ICLR 2019; http://arxiv.org/abs/1803.03635.

[3] Rushil Anirudh, Jayaraman J. Thiagarajan, Rahul Sridhar, Peer-Timo Bremer: "MARGIN: Uncovering Deep Neural Networks using Graph Signal Analysis", 2017; http://arxiv.org/abs/1711.05407.

[4] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, Li Fei-Fei: "ImageNet Large Scale Visual Recognition Challenge", 2014; http://arxiv.org/abs/1409.0575.

[5] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao and Li Fei-Fei. Novel dataset for Fine-Grained Image Categorization. First Workshop on Fine-Grained Visual Categorization (FGVC), IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

[6] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Fua Pascal, and Sabine Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34:2274–2282, 2012.

[7] Hao Yuan, Haiyang Yu, Shurui Gui, Shuiwang Ji: "Explainability in Graph Neural Networks: A Taxonomic Survey", 2020; http://arxiv.org/abs/2012.15445.

[8] Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, Shuiwang Ji: "On Explainability of Graph Neural Networks via Subgraph Explorations", 2021; http://arxiv.org/abs/2102.05152.

3.2. A csatlakozó dokumentumok jegyzéke

A projekt forráskódja elérhető a következő github repository-ban: https://github.com/phujbert/MARGIN_DL_Explainability.git