Prototípus-alapú felismerő rendszerek

– eredmények és kihívások –

Csató Lehel
Matematika és Informatika Kar
Babeş–Bolyai Tudományegyetem, Kolozsvár

2025 október 24

DataMin:

Bajcsi Adél,
Bodó Zalán
Csató Lehel,
Lieb Hanna,
Máté Ditmár,
Mester Attila,
Portik Ábel,
Sándor Csanád,
Szenkovits Annamária,
...
...

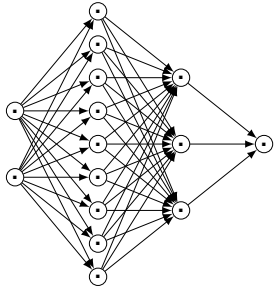
Digitális Székelyföld Konferencia
Székelyudvarhely

Tartalom

Neurális architektúrák magyarázhatósága „Explainability” I

42 – Douglas Adams

Válasz az élet értelmére is.
Azonban nem tudjuk, hogy a válasz miért pont ennyi.



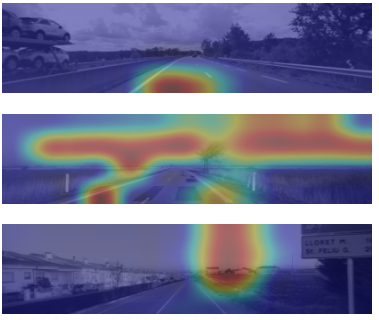
A neurális hálók nem visszakövethetőek
(?bonyolultak?).

Szeretnénk, ha a neurális háló a döntését meg tudnánk indokolni a döntését.

- Képek esetében – „indoklás” – a lokáció meghatározása.
- Például: „sáros az út, mert a képen vannak arra utaló jelek”.

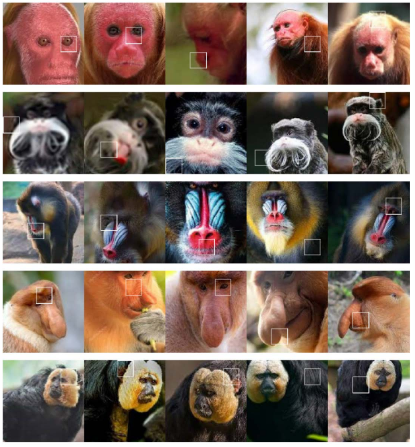
Neurális architektúrák magyarázhatósága „Explainability” II

1. példa Magyarázatok
intenzitás-térképek – heat-map-ek – segítségével:



Neurális architektúrák magyarázhatósága „Explainability” III

2. példa Szeretnénk, ha a *következtetésekat* (magyarázatokat) is „tanulnánk”.



Tartalom

Magukat magyarázó modellek „Self-Explainable”

„Ön”-magyarázó modellek – self-explainable models

- Képek feldolgozásánál használtak (osztályozásra vagy szegmentálásra);
- Architekturális módosításokat tartalmaz, mely segítségével a döntés „támogatásaként” létrejön a magyarázat is;
- Általában a döntési mechanizmust teszik „láthatóvá”.
- Cél, hogy az osztályozás vagy a szegmentálás *tanulásával* együtt tanuljuk meg az architekturális elemeket.

PrototypeDL architektúra [Li et al., 2018]

- A bemeneti adatokat – kódoló rendszerhez hasonlóan – egyszerűsítjük;

$$x \to f(x) = \Phi_x$$

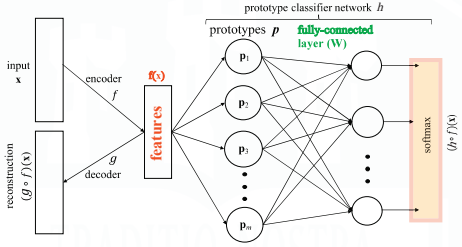
- A *feature*-vektorok tartalmazzák a visszaállításhoz szükséges információt;

$$\Phi_x \to g(x) = \hat{x}$$

- Cél, hogy a visszaállítás minél jobb legyen

$$\min \|x - \hat{x}\|$$

- A *feature*-vektorokat nevezhetjük a képek reprezentánsának.

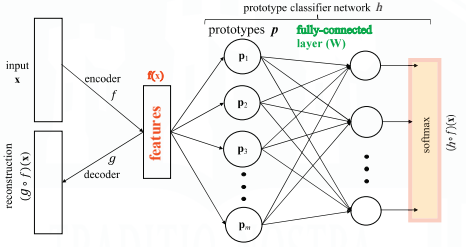


PrototypeDL II

[Li et al., 2018]

A PrototypeDL működése:

- Egy feature-vektort rendelünk a teljes képhez – eltakarja a részleteket;
- Osztályozás céljára megfelel, azonban az *interpretálhatóság* nincs jelen.

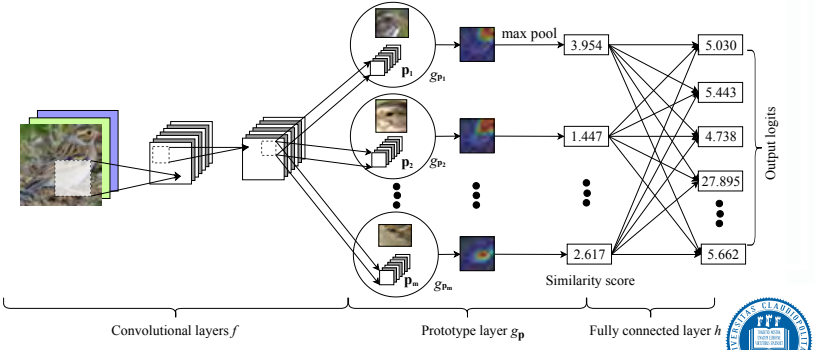


Szükséges finomítás!

ProtoPNet architektúra [Chen et al., 2019]

This Looks Like That: Deep Learning for Interpretable Image Recognition

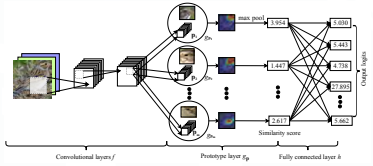
- Több reprezentánsa is van egy képnek; ezek koordinátával rendelkeznek;
- A prototípusok kiszámítása *konvolúciós hálóval* történik;
- A rendszer ezekből a reprezentáns-vektorokból készíti a prototípusokat;



ProtoPNet jellemzői [Chen et al., 2019]

A rendszer működése:

- A konvolúciós réteg kimenete a prototípus-jelölt
- A prototípus-réteg kiszámolja és továbbítja a hasonlóság mértékét (LVQ)
- A hasonlóság alapján számoljuk a kimenetet az utolsó réteg teljesen összekötött.



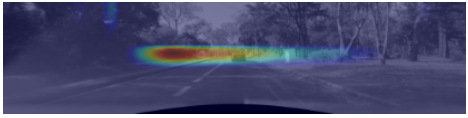
- Az *interpretálhatóság* elősegítésére a prototípusok adott osztályhoz tartoznak (minden osztály adott számú prototípussal rendelkezik);
- A prototípusokat tanították képek részeinek a megjegyzésére hasonló: perceptual grounding.

ProtoPNet eredmények I

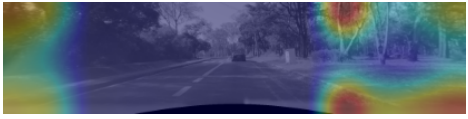
[Chen et al., 2019]

Helyes osztályozás:

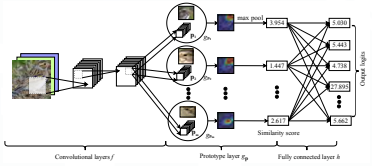
motorway



rural



urban



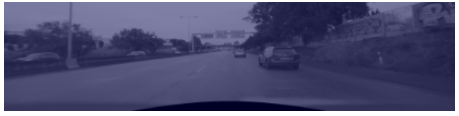
ProtoPNet eredmények II

[Chen et al., 2019]

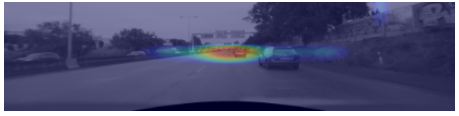
Helytelen osztályozás:

Incorrect motorway class

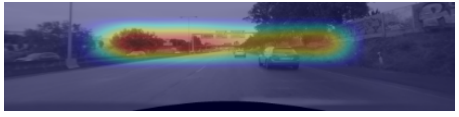
motorway

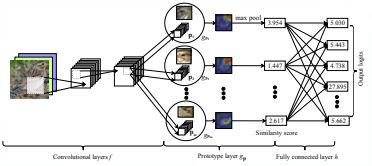


rural



urban





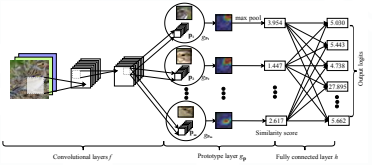
ProtoPNet eredmények II

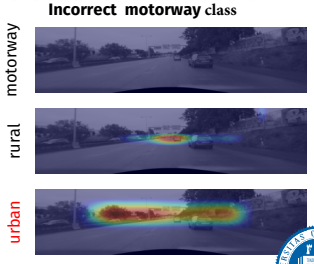
[Chen et al., 2019]

Kísérletek összeoglalása:

- A *ProtoPNet* konvolúciós része fekete dobozként működik;
- Hiba esetén nem tudtuk hogyan javítsunk ezen.
- Az osztályozó redundáns információkat kell mellőzzön – nem sikerül mindig
  (a céges informatikusoknak a panasza: az adatok rosszak).

Szükséges finomítás!

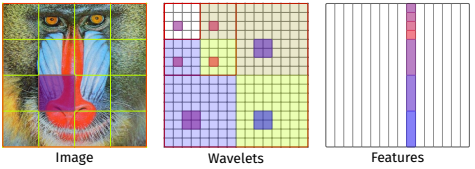


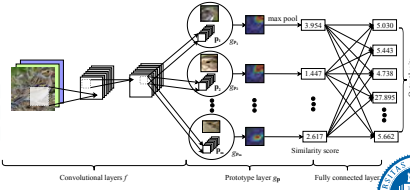


A Wave-ProtoPNet rendszer [Kaszta et al., 2023]

Jellemzők:

- A konvolúciós modul nem adaptív, hanem determinisztikus;
- Wavelet-transzformációt használunk, a *feature*-vektorok kiszámítására.
- Mivel azonos műveleteket végzünk, ezek konvolúcióként működnek.
- Mivel nincs tanítás a konvolúcióknál, a rendszer gyorsan tanul.
- Az ábrázolás értelmezhető, a reprezentáció sűríthető (természetesen információvesztéssel).





A Wave-ProtoPNet eredmények I [Kaszta et al., 2023]

Helyes osztályozás

(urban)

(rural)

(expressway)

(motorway)

A Wave-ProtoPNet eredmények II [Kaszta et al., 2023]

Helytelen osztályozás

(urban)

(rural)

(expressway)

(motorway)

Tartalom

Következtetések

A prototípusokról:

- A prototípus-architektúra egy lehetséges iránya a magyarázhatóságnak,
- Szükséges feltétel a prototípusok egyszerűsége, ami redundanciához vezethet (pl. nem rotáció-invariánsak a tárolt prototípusok).

Oktatás / módszertan:

- A programozás julia-ban történt, hasznosak voltak a funckionális és lineáris algebrai jellemzők.
- A diákok szeretnek programozók lenni – nehéz a más munkára rávezetni.
- A GPT-k hurrá-optimizmusa ellehetetleníti a kritikus gondolkodást.
- Nem kell mindig nagyobb rezolúció; az nem jelent több információt.

Köszönet...

Köszönöm a figyelmet!

Köszönetnyilvánítás:
„Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization, and Financial Instruments Program, 2021-2027, MySMIS no. 334906.

Könyvészet I

Chen, C., Li, O., Tao, C., Barnett, A. J., Su, J., and Rudin, C. (2019).
This Looks like That: Deep Learning for Interpretable Image Recognition.
In *NeurIPS*, pages 8930–8941.

Kaszta, T., Lieb, H.-G., and Csató, L. (2023).
Wavelet-based prototype networks.
In *2023 IEEE 21st Jubilee International Symposium on Intelligent Systems and Informatics (SISY)*, pages 277–282.

Li, O., Liu, H., Chen, C., and Rudin, C. (2018).
Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 3530–3537.