# Képszegmentálási modellek interpretálhatósági metrikái

Szenkovits Annamária

BBTE, Matematika és Informatika Kar, Kolozsvár

16. Matematika és informatika alkalmazásokkal konferencia
2025. november 29.

# Áttekintés

# Interpretálhatóság mérése?

- ideális esetben: valós személyektől jövő visszajelzés
  - költséges
- kiértékelés minőségének **számszerűsítése**, automatizált mérése
  - metrikák definiálása
  - nincs konszenzus
  - dinamikusan változó terület: gyakran jelennek meg új modellek, új metrikák

# Szakirodalmi áttekintés - néhány metrika típus [4]

- 361 feldolgozott cikk a 2014 és 2020 közötti időszakból: magyarázható modellek
- **12 metrikát** javasolnak a modellek által biztosított magyarázatok **minőségi kiértékelésére**

- 3 kategória:
  - **tartalom**: magyarázat hűségét és teljességét vizsgálja a magyarázott fekete doboz modellhez képest
  - **megjelenítés**: a magyarázat formátumával és elrendezésével foglalkozik
  - **felhasználó**: a magyarázat felhasználóra gyakorolt hatását és a felhasználói igények figyelembevételét vizsgálja

# Szakirodalmi áttekintés - néhány metrika típus [4]

| Metrika típusa | Metrika leírása |
|---|---|
| **helyesség** | a magyarázat mennyire tükrözi hűen a magyarázott fekete doboz modell működését |
| **teljesség** | a magyarázat mekkora mértékben fedi le a magyarázott fekete doboz viselkedését |
| **konzisztencia** | mennyire következetes, determinisztikus a magyarázat |

Table: A magyarázat **tartalmára** vonatkozó metrikák

# Szakirodalmi áttekintés - néhány metrika típus [4]

| Metrika típusa | Metrika leírása |
|---|---|
| **helyesség** | a magyarázat mennyire tükrözi hűen a magyarázott fekete doboz modell működését |
| **teljesség** | a magyarázat mekkora mértékben fedi le a magyarázott fekete doboz viselkedését |
| **konzisztencia** | mennyire következetes, determinisztikus a magyarázat |

Table: A magyarázat **tartalmára** vonatkozó metrikák

# Szakirodalmi áttekintés - néhány metrika típus [4]

| Metrika típusa | Metrika leírása |
|---|---|
| **tömörség** | a magyarázat mérete |
| **kompozíció** | a magyarázat megjelenítési formátuma |

Table: A magyarázat **megjelenítésére** vonatkozó metrikák

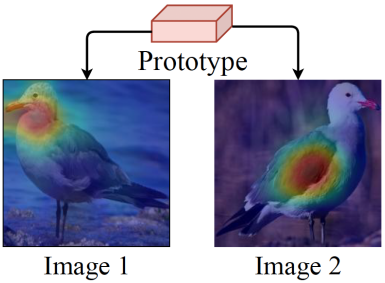
# Szakirodalmi áttekintés - néhány metrika típus [4]

| Metrika típusa | Metrika leírása |
|---|---|
| **kontextus** | mennyire releváns a magyarázat a felhasználó és az igényei szempontjából |
| **koherencia** | mennyire egyezik a magyarázat a korábbi ismeretekkel |

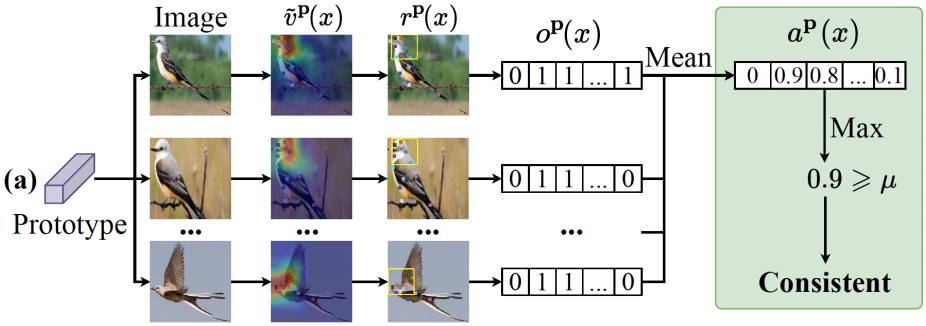
Table: A magyarázatnak a **felhasználóra** gyakorolt hatását mérő metrikák

# Prototípusok konzisztenciája képosztályozó modellek esetében

- kiindulópont: [2] – prototípusalapú képosztályozó modellekre (pl. ProtoPNet [1]) vezettek be **interpretálhatósági metrikákat**
- prototípusok **konzisztenciájának** mérése: egy adott prototípus különböző képek esetében mennyire következetesen aktiválódik az objektumok ugyanazon részeiben (pl. madár csőre)?

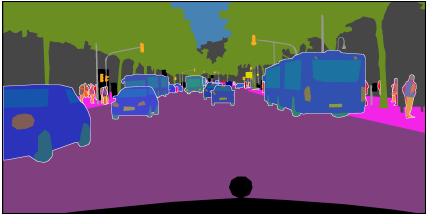


# Prototípusok konzisztenciája képosztályozó modellek esetében - ProtoPNet

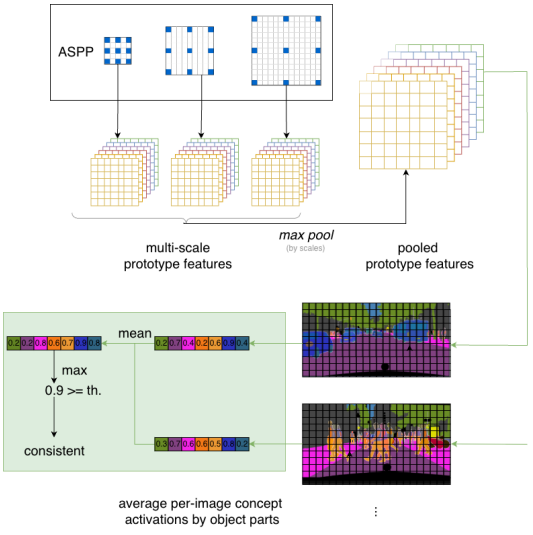


# Koncepciók konzisztenciája képszegmentáló modell esetében

- a bemutatott metrika adaptálása szegmentálásra
- felhasznált adathalmaz: **Cityscapes-Panoptic-Parts** [3]
- a Cityscapes adathalmaz **kibővített változata**: szemantikus osztályokon belül be vannak jelölve a különböző objektumok részei is, pl:
  - **személyek** törzse, feje, karja, lába
  - **autók** kereke, szélvédője, rendszámtáblája



# Koncepciók konzisztenciája képszegmentáló modell esetében



# Előzetes eredmények

| **küszöbérték** | **konzisztens koncepciók száma** |
|---|---|
| 0.6 | 107/256 |
| 0.7 | 82/256 |
| 0.8 | 7/256 |

Table: A szegmentáló modell konzisztens koncepcióinak száma különböző küszöbértékek esetében, a Cityscapes adathalmazon, 0.64-es tanítási mIoU érték esetében

# További lépések

- további metrikák definiálása
- metrikák lemérése a ProtoSeg modellre is, összehasonlítás
- egyéb adathalmazok (pl. Pascal VOC)

# References I


Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su.
This looks like that: deep learning for interpretable image recognition.
*Advances in Neural Information Processing Systems*, 32, 2019.

Qihan Huang, Mengqi Xue, Wenqi Huang, Haofei Zhang, Jie Song, Yongcheng Jing, and Mingli Song.
Evaluation and improvement of interpretability for self-explainable part-prototype networks.
In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2011–2020, 2023.

# References II


Panagiotis Meletis, Xiaoxiao Wen, Chenyang Lu, Daan de Geus, and Gijs Dubbelman.
Cityscapes-Panoptic-Parts and PASCAL-Panoptic-Parts datasets for Scene Understanding.
arXiv preprint arXiv:2004.07944, 2020.


Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert.
From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai.
*ACM Computing Surveys*, 55(13s):1–42, July 2023.

# Köszönöm a figyelmet!

A kutatást a **“Romanian Hub for Artificial Intelligence – HRIA”**, Smart Growth, Digitization and Financial Instruments Program, 2021–2027, MySMIS no. 351416 projekt, illetve a **Magyar Tudományos Akadémia 62/22/2025/HTMT azonosítójú Domus** ösztöndíja támogatta.