ProtoInit: prototípus-inicializálás önmagyarázó hálókra

Lieb Hanna-Georgina, Sándor Csanád, Bodó Zalán

Babeș–Bolyai Tudományegyetem, Matematika-Informatika kar, Kolozsvár

hanna.lieb@ubbcluj.ro, csanad.sandor@ubbcluj.ro, zalan.bodo@ubbcluj.ro

A prototípus-alapú hálók elősegítik a modellek magyarázhatóságát azáltal, hogy az osztályozási kimeneteket vizuális prototípusokhoz kapcsolják. Mivel a prototípusok általában véletlenszerűen vannak inicializálva, előfordulhat, hogy értelmes osztályspecifikus jellemzők helyett irreleváns részekre fókuszálnak. Jelen munkánkban egy olyan új prototípus inicializálási technikát vezetünk be, amely egy előtanított jelemzőkivonási gerinchálót használ az egyedi, osztályspecifikus jellemzők azonosítására. Rangsorolván és kiválasztván a legkiugróbb jellemzőket képenként, a prototípusok inicializása vagy közvetlenül ezekből valósul meg, vagy a klaszterezett jellemzővektorok centroidjaiból. Ez az inicializáció a tanulási folyamatot informatívabb vizuális bizonyítékot nyújtó prototípusok felé tereli.

A kísérleti eredmények azt mutatják, hogy a javasolt módszer javíthatja a modell hatékonyságát és a prototípus-alapú magyarázatok megbízhatóságát, lehetővé téve a prototípusok számának jelentős csökkentését egy magas osztályozási pontosság megtartása mellett. Különösen jelentős javulást sikerült elérni azon gerinchálóknál, ahol a véletlenszerűen inicializált prototípusok tanítás után is a háttérre fókuszáltak. Ezekben az esetekben a mi módszerünkkel inicilizált prototípusok olyan aktivációs térképeket eredményeztek, amelyek jobban összhangban vannak az osztállyal kapcsolatos objektumokkal, ezáltal csökkentvén a háttérfókusz problémát. Összességében, az eredmények arra utalnak, hogy az irányított prototípus inicializálás javíthatja a prototípus-alapú önmagyarázó hálók megbízhatóságát és hatékonyságát.

Hivatkozások

[1] Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: deep learning for interpretable image recognition. Advances in Neural Information Processing Systems, 32.

[2] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1 (5), 206–215.

[3] Borji, A., & Itti, L. (2012, June). Exploiting local and global patch rarities for saliency detection. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (pp. 478-485). IEEE.

[4] Kim, S. S., Watkins, E. A., Russakovsky, O., Fong, R., & Monroy-Hernández, A. (2023, April). " Help me help the AI": Understanding how explainability can support human-AI interaction. In proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1-17).

[5] Nauta, M., Schlötterer, J., Van Keulen, M., & Seifert, C. (2023, June). PIP-net: Patch-based intuitive prototypes for interpretable image classification. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2744-2753). IEEE.