

Visuell tolkning av (dype) konvolusjonelle nevralnett

Layer-wise relevance propagation

Innhold

- Bakgrunn: Konvolusjon med lærte filtre som egenskapsuttrekning
- “Layer-wise relevance propagation”: Lagvis visualisering av viktigste piksler for en gitt nevronaktivering

Bakgrunn konvolusjon

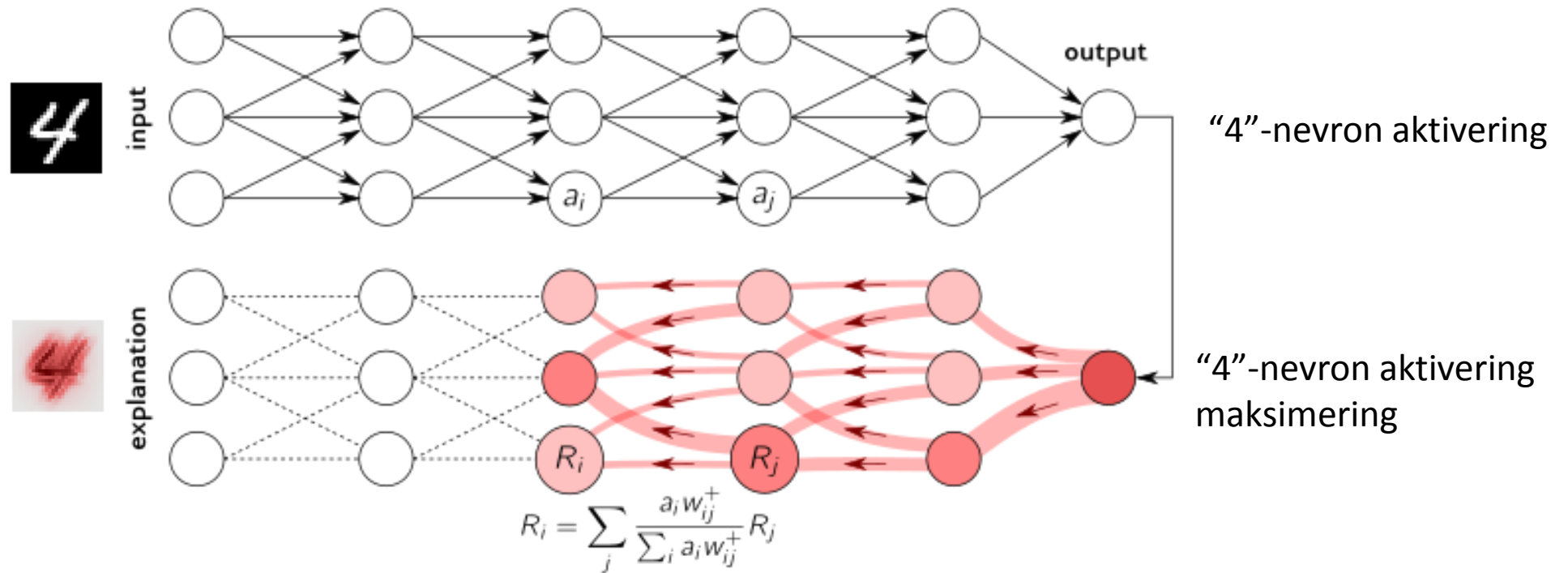
Delvis overlappende (strided) konvolusjon med gaussisk kule som filter: Nedskalert + utjevnet (smoothed) bilde

- Konvolusjonelle nevralnett: Filter kan være en hvilken som helst lært egenskap/struktur i bildet, ikke bare en gaussisk kule.
- Tilsvare et ettlags nett som har lært utjevning.
- Flerlags nett: mange lærte filtre/strukturer.

Layer-wise relevance propagation

En form for sensitivitetsanalyse

Heat-map over viktigste piksler for aktivering av et bestemt nevron



Tilbakepropageringformel

RLP med $\alpha\beta$ -formel

- α : positiv relevans/gradient (+) (rød)
- β : negativ relevans/gradient (-) (blå)

$$R_i = \sum_j \left(\alpha \frac{x_i w_{ij}^+}{\sum_i x_i w_{ij}^+} - \beta \frac{x_i w_{ij}^-}{\sum_i x_i w_{ij}^-} \right) R_j$$



Demo: <https://lrpserver.hhi.fraunhofer.de/image-classification>

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLOS ONE, 10(7), e0130140. <https://doi.org/10.1371/journal.pone.0130140>

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. Digital Signal Processing, 73, 1–15. <https://doi.org/10.1016/j.dsp.2017.10.011>