Aufgaben

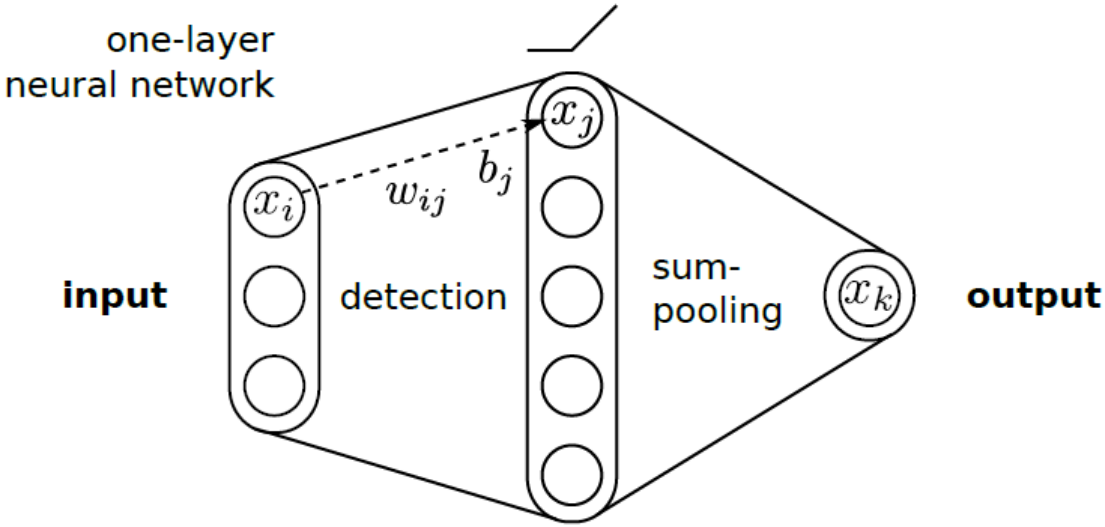
1. Arbeit an einem CNN für den Pascal VOC 2012 Datensatz fortsetzen
2. Implementierung des Ansatz der Deep Taylor Decomposition für DNN (insbesondere CNN)
3. Vergleich LRP $\leftrightarrow$ Deep Taylor Decomposition

Aufgabe

- Einführung CNN Part und Wdh. Deep Taylor

Deep Taylor Decomposition - Rückblick

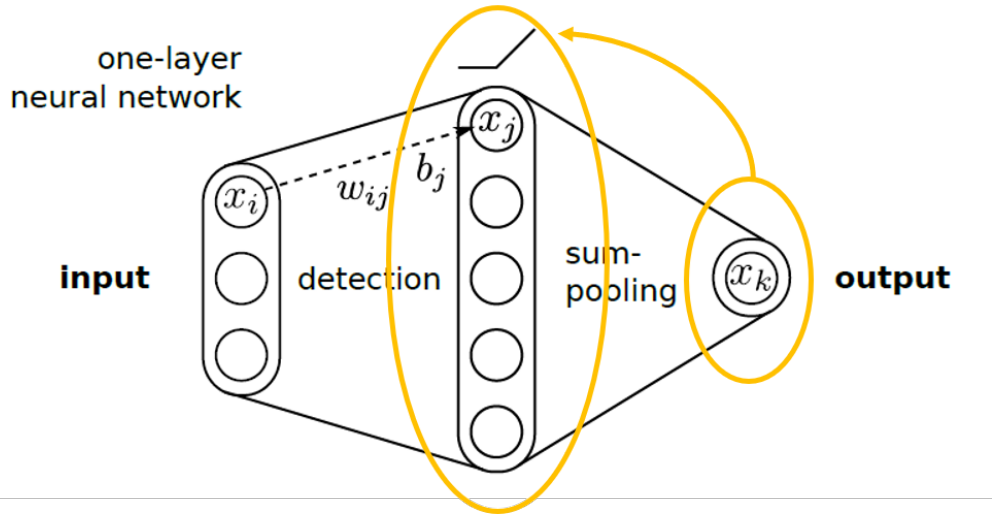
- Einfaches Netzwerk mit einem Hidden Layer, ReLU Aktivierung und Sum-Pooling als Output.
- Zusätzliche Voraussetzung: $b_j \leq 0$.



- Für das Outputneuron x_k gilt: $x_k = max(0, \sum_j x_j)$

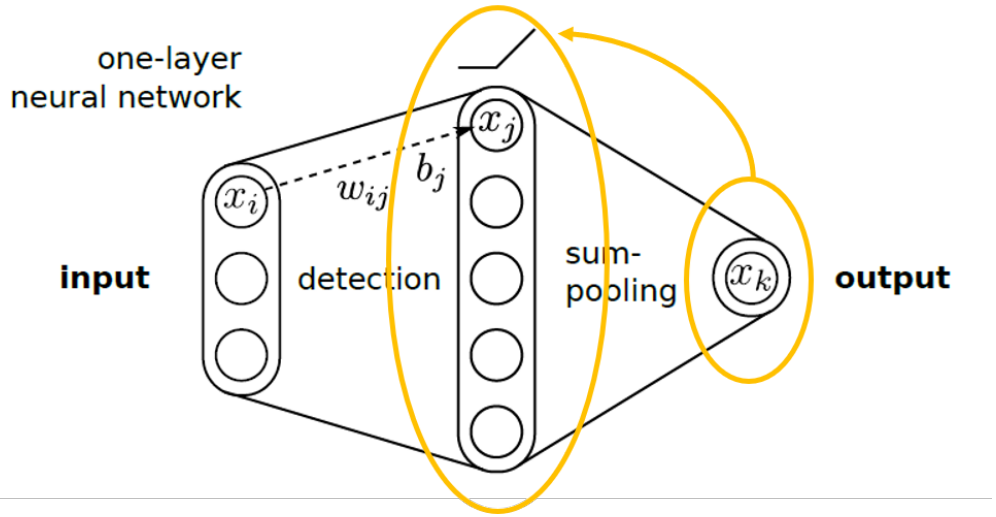
Deep Taylor Decomposition - Rückblick

- Suche eine Nullstelle für die Taylorentwicklung von $R_k(\boldsymbol{x}) = \sum_j x_j$.



Deep Taylor Decomposition - Rückblick

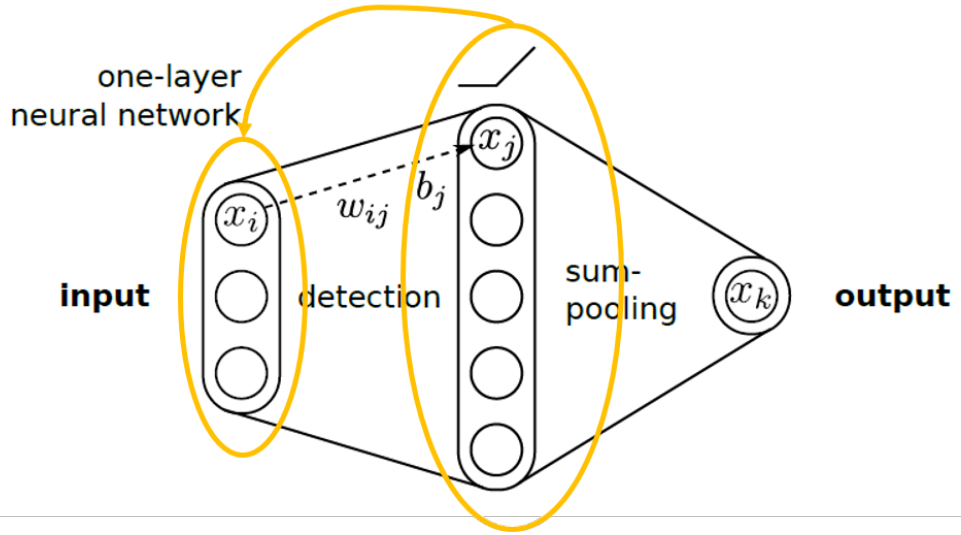
- Suche eine Nullstelle für die Taylorentwicklung von $R_k(\boldsymbol{x}) = \sum_j x_j$.



- Wg. ReLU Aktivierung im vorherigen Layer und $\sum_j x_j \overset{!}{=} 0$ ist $\tilde{\boldsymbol{x}} = \mathbf{0}$ die einzige Nullstelle von R_k.
- Wegen $R_j = \frac{\partial R_k}{\partial x_j}(x_j - \tilde{x}_j) = 1 \cdot (x_j - 0)$ gilt also
- $R_j = x_j = \max\left(0, \sum_i x_i w_{ij} + b_j\right)$

Deep Taylor Decomposition - Generische Regel

- Es gilt $R_j = x_j = \max(0, \sum_i x_i w_{ij} + b_j)$



- Unterscheide nun 2 Fälle:
  1. $R_j = 0$: Nicht aktivierte Neuronen sollen keine Relevanz zurückverteilen. Insbesondere gilt hier $\tilde{\boldsymbol{x}} = \boldsymbol{x}$.
  2. $R_j > 0$: Hierfür wird ein Richtungsvektor $\boldsymbol{v}^{(j)}$ definiert. $\tilde{\boldsymbol{x}}$ soll von der Form $\boldsymbol{x} + t \cdot \boldsymbol{v}^{(j)}$, mit $t \in \mathbb{R}$ sein.

Deep Taylor - Entwicklungspunkt

- Allgemeine Vorgehensweise:
- Durch Einsetzen von $\tilde{\boldsymbol{x}} = \boldsymbol{x} + t \cdot \boldsymbol{v}^{(j)}$ in die Ebenengleichung $\sum_i \tilde{x}_i w_{ij} + b_j$ lässt sich eine allgemeine Formel für t finden.
- Somit gilt:

$$0 = \sum_i \left(x_i + t v_i^{(j)} \right) w_{ij} + b_j$$

$$\Leftrightarrow -t = \frac{\sum_i x_i w_{ij} + b_j}{\sum_i v_i^{(j)} w_{ij}}$$

$$\Rightarrow x_i - \tilde{x}_i = -t v_i^{(j)} = \frac{\sum_i x_i w_{ij} + b_j}{\sum_i v_i^{(j)} w_{ij}} v_i^{(j)}$$

Deep Taylor - Entwicklungspunkt

- Für die Umverteilung von der $l+1$-ten Schicht in die l-te Schicht gilt

$$
\begin{aligned}
R_i^l &= \sum_j R_{i\leftarrow j}^l = \sum_j \frac{\partial R_j^{l+1}}{\partial x_i^l}(x_i - \tilde{x}_i) \\
&= \sum_{j:R_j=0} \frac{\partial R_j^{l+1}}{\partial x_i^l} \cdot 0 + \sum_{j:R_j>0} w_{ij} \frac{\sum_i x_i w_{ij} + b_j}{\sum_i v_i^{(j)} w_{ij}} v_i^{(j)} \\
&= \sum_j \frac{v_i^{(j)} w_{ij}}{\sum_i v_i^{(j)} w_{ij}} R_j^{l+1}
\end{aligned}
$$

- $\Rightarrow$ Allgemeine Formel in Abhängigkeit von $v_i^{(j)}$

Deep Taylor - Die z^+-Regel

- Grundannahme der Deep Taylor Decomposition ist die Anwendung der ReLU-Aktivierungsfunktion.
  $\Rightarrow$ für Input eines Layers gilt $\boldsymbol{x} \in \mathbb{R}_+^d$
  $\Rightarrow$ zulässige Nullstelle sollte auch aus zulässigem Bereich kommen
- Gesucht wird Nullstelle für R_j auf dem Intervall

$$[\{x_i 1_{w_{ij}<0}\}, \{x_i\}]$$

- Mindestens eine Nullstelle existiert bei $\{x_i 1_{w_{ij}<0}\}$

Deep Taylor - Die z^+-Regel

- Wähle $v_i^{(j)} = x_i - x_i \cdot 1_{w_{ij} \leq 0} = x_i \cdot 1_{w_{ij} > 0}$
- Einsetzen in die Gleichung liefert:

$$R_i = \sum_j \frac{x_i 1_{w_{ij}>0} w_{ij}}{\sum_i x_i 1_{w_{ij}>0} w_{ij}} R_j^{l+1} = \sum_j \frac{z_{ij}^+}{\sum_i z_{ij}^+} R_j^{l+1}$$

- Mit $z_{ij} = x_i \cdot w_{ij}$

Herleitung z^B-Regel

- Für Inputwerte des ersten Layers (z.B. Pixelwerte) gilt Positivität i.A. nicht
- Diese sind meistens beschränkt und können auch Werte kleiner 0 annehmen
- Formal: $x \in \mathcal{B}$ mit

$$\mathcal{B} = \{x \in \mathbb{R}^d : l_i \leq x_i \leq h_i \quad \forall i \in \{0, ..., d\}\}$$

Herleitung z^B-Regel

- Gesucht wird Nullstelle für R_j auf dem Intervall

$$[l_i 1_{w_{ij}>0} + h_i 1_{w_{ij}<0}, \{x_i\}]$$

- Auf diesem Intervall existiert mindestens eine Nullstelle, denn

$$\begin{aligned} & R_j\left(\{l_i 1_{w_{ij}>0} + h_i 1_{w_{ij}<0}\}\right) \\ &= \max\left(0, \sum_i l_i 1_{w_{ij}>0} \cdot w_{ij} + h_i 1_{w_{ij}<0} \cdot w_{ij} + b_j\right) \\ &= \max\left(0, \sum_i l_i \cdot w_{ij}^+ + h_i \cdot w_{ij}^- + b_j\right) = 0 \end{aligned}$$

Herleitung z^B-Regel

- Dieses Intervall ist also für die Suche zulässig und wir fügen die Richtung $v_i^{(j)}$, mit

$$v_i^{(j)} = x_i - l_i 1_{w_{ij}>0} + h_i 1_{w_{ij}<0},$$

- in die Grundformel

$$\sum_j \frac{v_i^{(j)} w_{ij}}{\sum_i v_i^{(j)} w_{ij}} R_j$$

- ein und erhalten unsere finale $z^{\mathcal{B}}$-Regel

$$R_i = \sum_j \frac{z_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-}{\sum_i z_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-} R_j$$

Anwendung auf den Pascal-Datensatz

- VGG16 trainiert für Multilabel-Klassifizierung für die Klassen *Mensch* und *Pferd*
- $z^{\mathcal{B}}$-Regel wurde bei Anwendung aller Regeln für das Inputlayer verwendet
- LRP-Composition:
  - LRP-0 auf Dense-Layern
  - LRP-ϵ auf mittleren sechs Conv-Layern
  - LRP-γ auf den letzten sechs Conv-Layern vor Inputlayer

Aufgabe

- Vergleich LRP DTD theoretisch
- Ggf. Ausblick Min-Max?