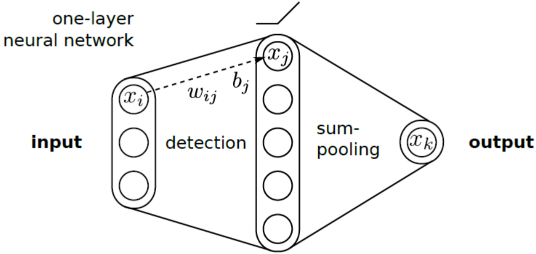
Aufgaben

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

Deep Taylor Decomposition - Rückblick

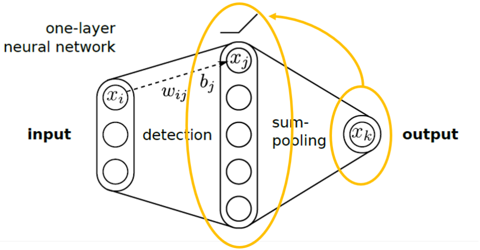
- Einfaches Netzwerk mit einem Hidden Layer, ReLU Aktivierung und Sum-Pooling als Output.
- Zusätzliche Voraussetzung: $b_j \leq 0$.



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

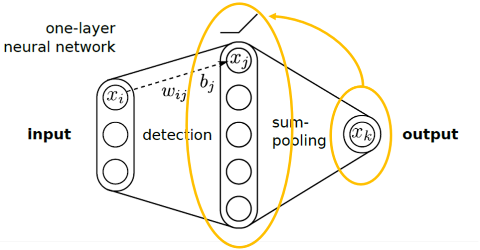
Deep Taylor Decomposition - Rückblick

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



Deep Taylor Decomposition - Rückblick

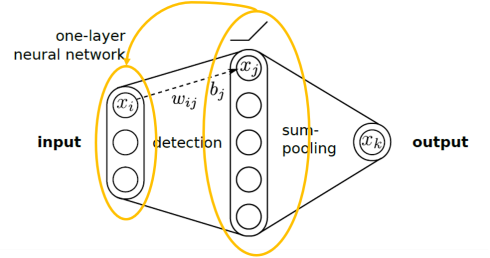
- Suche eine Nullstelle für die Taylorentwicklung von $R_k(\boldsymbol{x}) = \sum_j x_j$.



- Wg. ReLU Aktivierung im vorherigen Layer und $\sum_j x_j \stackrel{!}{=} 0$ ist $\tilde{\boldsymbol{x}} = \boldsymbol{0}$ 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
- $R_j = x_j = \max(0, \sum_i x_i w_{ij} + b_j)$

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

- z^+ Regel vom letzten Vortrag
- Idee der Regel: Wähle einen Punkt in $\mathbb{R}^l$, der bez. R_j auf 0 abbildet.
- Wähle $v_i^{(j)} = x_i - x_i \cdot \mathbb{1}_{w_{ij} \leq 0} = x_i \cdot \mathbb{1}_{w_{ij} > 0}$
- Einsetzen in die Gleichung liefert:

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

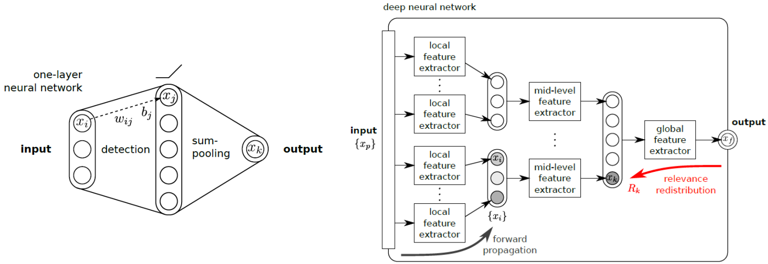
- Mit $z_{ij} = x_i \cdot w_{ij}$

Deep Taylor - Entwicklungspunkt

- z^B Regel herleiten.

Erweiterung auf tiefe Netze

- Bei tiefen Netzen, insbesondere Convolutional Layern ist die Relevanzfunktion nicht unbedingt explizit angegeben.



- Ein Feature kann vorhanden sein, aber bei der Bilderkennung keine Rolle spielen

Erweiterung auf tiefe Netze

- Gesucht ist eine Approximation der Relevanz Funktion, die leicht zu analysieren ist.
- Führe das Konzept **Relevanz-Modell** ein, um bei tieferen Netzen die Relevanzfunktion $R_j^{l+1}(\boldsymbol{x}^l)$ zu approximieren.
- Nehme an, die Relevanzfunktion R_j lässt sich in der Form $R_j = x_j \cdot c_j$ schreiben, wobei c_j konstant ist.
- Im Paper als "Training Free" Ansatz vorgestellt

Erweiterung auf tiefe Netze

- Nehme an, die Relevanzfunktion R_j lässt sich in der Form $R_j = x_j \cdot c_j$ schreiben, mit c_j konstant.
- Betrachte die generische Redistributionsregel

$$
\begin{aligned}
R_i &= \sum_j \frac{x_i \cdot \rho(w_{ij})}{\sum_i x_i \cdot \rho(w_{ij})} R_j \\
&= x_i \sum_j \frac{\rho(w_{ij})}{\sum_i x_i \cdot \rho(w_{ij})} x_j \cdot c_j \\
&= x_i \sum_j \rho(w_{ij}) \frac{\max\left(0, \sum_i x_i w_{ij}\right)}{\sum_i x_i \cdot \rho(w_{ij})} c_j
\end{aligned}
$$

Erweiterung auf tiefe Netze

- Nehme an, die Relevanzfunktion R_j lässt sich in der Form $R_j = x_j \cdot c_j$ schreiben, mit c_j konstant.
- Betrachte die generische Redistributionsregel

$$\begin{aligned} R_i &= \sum_j \frac{x_i \cdot \rho(w_{ij})}{\sum_i x_i \cdot \rho(w_{ij})} R_j \\ &= x_i \sum_j \frac{\rho(w_{ij})}{\sum_i x_i \cdot \rho(w_{ij})} x_j \cdot c_j \\ &= x_i \underbrace{\sum_j \rho(w_{ij}) \frac{\max\left(0, \sum_i x_i w_{ij}\right)}{\sum_i x_i \cdot \rho(w_{ij})} c_j}_{:\approx c_i} \end{aligned}$$

Erweiterung auf tiefe Netze

- Es gilt also

$$R_i = x_i \underbrace{\sum_j \rho(w_{ij}) \frac{\max\left(0, \sum_i x_i w_{ij}\right)}{\sum_i x_i \cdot \rho(w_{ij})} c_j}_{:\approx c_i} = \sum_j \frac{\rho(w_{ij}) \cdot x_i}{\sum_i x_i \cdot \rho(w_{ij})} R_j$$

- D.h. R_i^l lässt sich wieder schreiben als $x_i^l \cdot c_i^l$, mit c_i^l annähernd konstant.
- Ausgehend von der letzten Schicht kann die Relevanz somit auch gemäß der hergeleiteten Regeln zurück zum Input verteilt werden.
- Der Parameter c_i^l wird dabei durch die betrachtete Regel "induktiv" gebildet.

Zusammenhang mit LRP

- Die klassische $LRP - 0$ Formel kann als Deep Taylor Entwicklung im Nullpunkt gesehen werden
- Betrachte O.B.d.A. ein Neuron x_j mit $R_j > 0$.
- Die Suchrichtung $\boldsymbol{v}$ ist hierbei der Punkt $\boldsymbol{x}$ selbst, und somit gilt:

Zusammenhang mit LRP

- Die klassische $LRP - 0$ Formel kann als Deep Taylor Entwicklung im Nullpunkt gesehen werden
- Betrachte O.B.d.A. ein Neuron x_j mit $R_j > 0$.
- Die Suchrichtung $\boldsymbol{v}$ ist hierbei der Punkt $\boldsymbol{x}$ selbst, und somit gilt:

$$
\begin{aligned}
R^l_{i \leftarrow j} &= \frac{\partial R^{l+1}_j}{\partial x^l_i}(x_i - \tilde{x}_i) = w_{ij} \cdot c_j \cdot (x_i - \tilde{x}_i) \\
&= w_{ij} \cdot c_j \cdot \frac{\sum_i x_i w_{ij} + b_j}{\sum_i x_i w_{ij}} x_i \\
&= \frac{x_i \cdot w_{ij}}{\sum_i x_i w_{ij}} x_j \cdot c_j
\end{aligned}
$$

Zusammenhang mit LRP

- Die klassische $LRP - 0$ Formel kann als Deep Taylor Entwicklung im Nullpunkt gesehen werden
- Betrachte O.B.d.A. ein Neuron x_j^{l+1} mit $R_j > 0$.
- Die Suchrichtung $\boldsymbol{v}$ ist hierbei der Punkt $\boldsymbol{x}$ selbst, und somit gilt:

$$
\begin{aligned}
R_{i \leftarrow j}^{l} &= \frac{\partial R_j^{l+1}}{\partial x_i^l}(x_i - \tilde{x}_i) = w_{ij} \cdot c_j \cdot (x_i - \tilde{x}_i) \\
&= w_{ij} \cdot c_j \cdot \frac{\sum_i x_i w_{ij} + b_j}{\sum_i x_i w_{ij}} x_i \\
&= \frac{x_i \cdot w_{ij}}{\sum_i x_i w_{ij}} \underbrace{x_j \cdot c_j}_{R_j}
\end{aligned}
$$

Zusammenhang mit LRP

- Für die totale Relevanz von x_i gilt:

$$R_i^l = \sum_j R_{i\leftarrow j}^l = \sum_j \frac{x_i \cdot w_{ij}}{\sum_i x_i w_{ij}} R_j^{l+1} = \sum_j \frac{z_{ij}}{\sum_i z_{ij}} R_j^{l+1}$$

- Mit der gleichen Vorgehensweise können auch die $LRP - \varepsilon$ Regel sowie $LRP - \gamma$ hergeleitet werden.
- Deep Taylor Decomposition als Basis für *LRP*
- Wesentlicher Unterschied: Geforderte Konsistenz von Montavon et al. im Vergleich zu den *LRP* Regeln