Relevance Propagation for Deep Neural Networks

Zwischenvortrag 2

Theo Conrads, Robin Kühling, Marc Bremser

13. Juni 2020

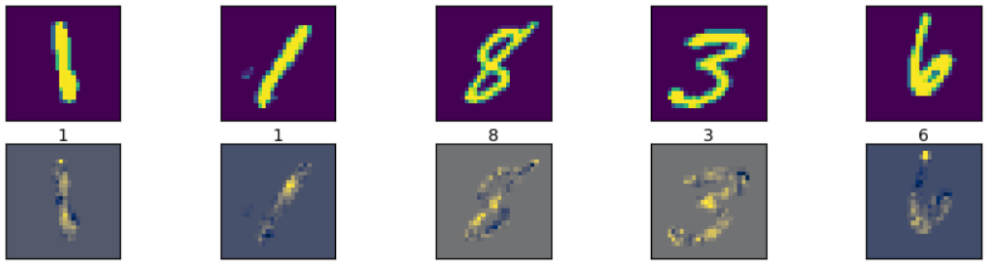
Überblick

Konzept der LRP

- Gemäss bestimmter Formeln soll die Relevanz einzelner Pixel durch eine "Rückrechnung" aus dem Output visualisiert werden.
- Einfachste Form:

$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_{i'} z_{i'j}} R_j^{(l+1)} \quad \text{mit} \quad z_{ij} = x_i^{(1)} w_{ij}^{(1,1+1)} \tag{1}$$

- Am Beispiel des MNIST Datensatz:



Vergleich LRP $\leftrightarrow$ Heatmap

- Ersetze den Wert der Neuronen x_i durch den Mittelwert der Schicht.
- Versuche, allgemeine Aussagen zu treffen.
- Am Beispiel des MNIST Datensatz:

Dense Layer ⇔ Convolutional Layer

Erweiterung: Mehrere Filter

- Grafik
-

Bilderkennung mittels CNNs - Implementierung

- Herausforderungen bei der Implementierung

Bilderkennung mittels CNNs - Implementierung

- Theo Abschnitt zur Implementierung
- Ergebnisse direkt dazu?

Taylor Decomposition für Neuronale Netze

- Problem bei der LRP bisher: Die Formeln machen Sinn und funktionieren, aber die theoretische Fundierung fehlt.
- Mit theoretischer Fundierung sind evlt. allgemeinere Aussagen möglich.
- Betrachte hierzu ein NN als Funktion $f : \mathbb{R}^p \to \mathbb{R}$, wobei p die Inputgrösse bezeichnet.
- Nehme an, dass $f(x) > 0$ Evidenz für das gesuchte Objekt bedeutet.
- Intuition: Verschiebe den Input, sodass $f(x) = 0$. Dimensionen, in die "weiter" verschoben wurde, haben offensichtlich mehr zur Klassifizierung beigetragen.

Taylor Decomposition

- Ansatz: Betrachte Taylor-Entwicklung von f.
- Wähle als Entwicklungspunkt $\hat{x}$ eine Nullstelle von f.
- Taylorentwicklung ist gegeben durch

$$\begin{aligned} f(\boldsymbol{x}) &= f(\hat{\boldsymbol{x}}) + \left(\frac{\partial f}{\partial \boldsymbol{x}} \bigg|_{\boldsymbol{x}=\hat{\boldsymbol{x}}} \right)^{\top} \cdot (\boldsymbol{x} - \hat{\boldsymbol{x}}) + \varepsilon \\ &= 0 + \sum_{p} \underbrace{\left| \frac{\partial f}{\partial x_p} \right|_{x=\hat{x}} \cdot (x_p - \hat{x}_p)}_{R_p(x)} + \varepsilon \end{aligned}$$

- $R_p(x)$ stellt somit die Relevanz der $p-$ten Komponente des Inputs dar.
- Genauigkeit hängt davon ab, ob Terme höherer Ordnung vernachlässigt werden können.

Taylor Decomposition

- Problem: Wahl eines geeigneten Punkts $\hat{x}$ für das ganze Netzwerk.
- Oft zu weit entfernt und gibt somit nicht so viele Informationen zur Entscheidungsfindung des Netzwerks.
- Gradient Shattering (-> Montavon Vortrag, nochmal genauer nachschauen)
- Abhilfe: Führe die Taylor Decomposition an jedem einzelnen Neuron mit einem eigenen Punkt $\hat{x}^{(j)}$ durch.
- Konzept der **Deep Taylor Decomposition**

Deep Taylor Decomposition

- Deep Taylor erklären

Zitate:

Zitat: [?]
Weiteres: [?]

Literaturliste