Deep Learning Teil 2

Übersicht

- §1 Approximationssatz für Neural Networks
- §2 Convolutional Neural Network (CNN)
- §3 Generative Adversarial Network (GAN)
- §4 Moderne Theorie von Neural Networks

Übersicht

- **§1 Approximationssatz für Neural Networks**
- §2 Convolutional Neural Network (CNN)
- §3 Generative Adversarial Network (GAN)
- §4 Moderne Theorie von Neural Networks

1 Satz von Hahn-Banach

Satz: Sei V ein $\mathbb{R}$-Vektorraum. Ist $p : V \to \mathbb{R}$ eine sub-lineare Funktion und $\varphi : U \to \mathbb{R}$ ein lineares Funktional auf einem linearen Unterraum $U \subseteq V$ mit $\varphi(x) \leq p(x)\ \forall x \in U$. Dann existiert ein lineares Funktional $\psi : V \to \mathbb{R}$ mit $\psi(x) = \varphi(x)\ \forall x \in U$ und $\psi(x) \leq p(x)\ \forall x \in V$.

2 Korollar von Hahn-Banach

Korollar: Ist V ein normierter $\mathbb{R}$-Vektorraum und $U \subsetneq V$ ein linearer Unterraum mit $\bar{U} \subsetneq V$. Dann existiert ein beschränktes lineares Funktional $L \neq 0$ auf V mit $L\big|_{\bar{U}} = 0$.

2 Korollar von Hahn-Banach

Korollar: Ist V ein normierter $\mathbb{R}$-Vektorraum und $U \subsetneq V$ ein linearer Unterraum mit $\bar{U} \subsetneq V$. Dann existiert ein beschränktes lineares Funktional $L \neq 0$ auf V mit $L\big|_{\bar{U}} = 0$.

Beweis: Sei V ein normierter $\mathbb{R}$-Vektorraum und $U \subsetneq V$ ein linearer Unterraum mit $\bar{U} \subsetneq V$. Wähle $f_0 \in \bar{U}^C$. Sei $M = \langle \bar{U} \cup \{f_0\} \rangle$. Definiere $T : M \to \mathbb{R}$, $T(f + a * f_0) = a$, $f \in \bar{U}$. Es ist klar, dass T wohldefiniert und linear ist.

2 Korollar von Hahn-Banach

Korollar: Ist V ein normierter $\mathbb{R}$-Vektorraum und $U \subsetneq V$ ein linearer Unterraum mit $\bar{U} \subsetneq V$. Dann existiert ein beschränktes lineares Funktional $L \neq 0$ auf V mit $L\big|_{\bar{U}} = 0$.

Beweis: Sei V ein normierter $\mathbb{R}$-Vektorraum und $U \subsetneq V$ ein linearer Unterraum mit $\bar{U} \subsetneq V$. Wähle $f_0 \in \bar{U}^C$. Sei $M = \langle \bar{U} \cup \{f_0\} \rangle$. Definiere $T : M \to \mathbb{R}$, $T(f + a * f_0) = a$, $f \in \bar{U}$. Es ist klar, dass T wohldefiniert und linear ist.
Man muss noch zeigen, dass T stetig ist. Sei $\lim_{n\to\infty} f_n + a_n * f_0 = g$ mit $f_n \in \bar{U}$ and $a_n \in \mathbb{R}$ $\forall n \in \mathbb{N}$. Ist (a_n) unbeschränkt, so existiert eine Teilfolge (a_{n_k}) mit Grenzwert $\pm\infty$.

2 Korollar von Hahn-Banach

Korollar: Ist V ein normierter $\mathbb{R}$-Vektorraum und $U \subsetneq V$ ein linearer Unterraum mit $\bar{U} \subsetneq V$. Dann existiert ein beschränktes lineares Funktional $L \neq 0$ auf V mit $L\big|_{\bar{U}} = 0$.

Beweis: Sei V ein normierter $\mathbb{R}$-Vektorraum und $U \subsetneq V$ ein linearer Unterraum mit $\bar{U} \subsetneq V$. Wähle $f_0 \in \bar{U}^C$. Sei $M = \langle \bar{U} \cup \{f_0\} \rangle$. Definiere $T : M \to \mathbb{R}$, $T(f + a * f_0) = a$, $f \in \bar{U}$. Es ist klar, dass T wohldefiniert und linear ist.

Man muss noch zeigen, dass T stetig ist. Sei $\lim_{n\to\infty} f_n + a_n * f_0 = g$ mit $f_n \in \bar{U}$ and $a_n \in \mathbb{R}$ $\forall n \in \mathbb{N}$. Ist (a_n) unbeschränkt, so existiert eine Teilfolge (a_{n_k}) mit Grenzwert $\pm\infty$.

Aus $\lim_{k\to\infty} f_{n_k} + a_{n_k} * f_0 = g$ folgt durch Teilen von a_{n_k}: $\lim_{k\to\infty} f_{n_k}/a_{n_k} = -f_0$. Wiederspruch, da $\bar{U}$ abgeschlossen ist und damit eine Folge in $\bar{U}$ keinen Grenzwert $-f_0 \notin \bar{U}$ haben kann. Daher muss (a_n) beschränkt sein.

2 Korollar von Hahn-Banach

Korollar: Ist V ein normierter $\mathbb{R}$-Vektorraum und $U \subsetneq V$ ein linearer Unterraum mit $\bar{U} \subsetneq V$. Dann existiert ein beschränktes lineares Funktional $L \neq 0$ auf V mit $L\big|_{\bar{U}} = 0$.

Beweis: Sei V ein normierter $\mathbb{R}$-Vektorraum und $U \subsetneq V$ ein linearer Unterraum mit $\bar{U} \subsetneq V$. Wähle $f_0 \in \bar{U}^C$. Sei $M = \langle \bar{U} \cup \{f_0\} \rangle$. Definiere $T : M \to \mathbb{R}$, $T(f + a * f_0) = a$, $f \in \bar{U}$. Es ist klar, dass T wohldefiniert und linear ist.

Man muss noch zeigen, dass T stetig ist. Sei $\lim_{n\to\infty} f_n + a_n * f_0 = g$ mit $f_n \in \bar{U}$ and $a_n \in \mathbb{R}\ \forall n \in \mathbb{N}$. Ist (a_n) unbeschränkt, so existiert eine Teilfolge (a_{n_k}) mit Grenzwert $\pm\infty$.

Aus $\lim_{k\to\infty} f_{n_k} + a_{n_k} * f_0 = g$ folgt durch Teilen von a_{n_k}: $\lim_{k\to\infty} f_{n_k}/a_{n_k} = -f_0$. Wiederspruch, da $\bar{U}$ abgeschlossen ist und damit eine Folge in $\bar{U}$ keinen Grenzwert $-f_0 \notin \bar{U}$ haben kann. Daher muss (a_n) beschränkt sein.

Es gibt also eine konvergente Teilfolge (a_{n_k}) mit Grenzwert $a \in \mathbb{R}$. Daraus folgt $\lim_{k\to\infty} f_{n_k} + a_{n_k} * f_0 = f + a * f_0$ für ein $f \in \bar{U}$. Damit gilt $g = f + a * f_0$. Daraus wiederum folgt $\lim_{n\to\infty} a_n = a$. Damit gilt $T(g) = a = \lim_{n\to\infty} a_n = \lim_{n\to\infty} T(f_n + a_n * f_0)$. T ist also stetig und damit beschränkt auf M.

2 Korollar von Hahn-Banach

Korollar: Ist V ein normierter $\mathbb{R}$-Vektorraum und $U \subsetneq V$ ein linearer Unterraum mit $\bar{U} \subsetneq V$. Dann existiert ein beschränktes lineares Funktional $L \neq 0$ auf V mit $L\big|_{\bar{U}} = 0$.

Beweis: Sei V ein normierter $\mathbb{R}$-Vektorraum und $U \subsetneq V$ ein linearer Unterraum mit $\bar{U} \subsetneq V$. Wähle $f_0 \in \bar{U}^C$. Sei $M = \langle \bar{U} \cup \{f_0\} \rangle$. Definiere $T : M \to \mathbb{R}$, $T(f + a * f_0) = a$, $f \in \bar{U}$. Es ist klar, dass T wohldefiniert und linear ist.
Man muss noch zeigen, dass T stetig ist. Sei $\lim_{n\to\infty} f_n + a_n * f_0 = g$ mit $f_n \in \bar{U}$ and $a_n \in \mathbb{R}\ \forall n \in \mathbb{N}$. Ist (a_n) unbeschränkt, so existiert eine Teilfolge (a_{n_k}) mit Grenzwert $\pm\infty$.
Aus $\lim_{k\to\infty} f_{n_k} + a_{n_k} * f_0 = g$ folgt durch Teilen von a_{n_k}: $\lim_{k\to\infty} f_{n_k}/a_{n_k} = -f_0$. Wiederspruch, da $\bar{U}$ abgeschlossen ist und damit eine Folge in $\bar{U}$ keinen Grenzwert $-f_0 \notin \bar{U}$ haben kann. Daher muss (a_n) beschränkt sein.
Es gibt also eine konvergente Teilfolge (a_{n_k}) mit Grenzwert $a \in \mathbb{R}$. Daraus folgt $\lim_{k\to\infty} f_{n_k} + a_{n_k} * f_0 = f + a * f_0$ für ein $f \in \bar{U}$. Damit gilt $g = f + a * f_0$. Daraus wiederum folgt $\lim_{n\to\infty} a_n = a$. Damit gilt $T(g) = a = \lim_{n\to\infty} a_n = \lim_{n\to\infty} T(f_n + a_n * f_0)$. T ist also stetig und damit beschränkt auf M.
Wähle nun bei den Bezeichnungen vom Satz von Hahn-Banach $U = M, V = V, \varphi = T$ und $p : V \to \mathbb{R}, p(f) = \|T\|_{op} * \|f\|$. Dann folgt die Behauptung aus dem Satz von Hahn-Banach.

3 Satz von Riesz-Markov-Kakutani für beschränkte, lineare Funktionale

Satz: Sei K ein kompakter Hausdorff-Raum und sei $L : C(K) \to \mathbb{R}$ ein beschränktes lineares Funktional. Dann existiert ein eindeutig bestimmtes reguläres, signiertes,endliches Borel-Maß μ auf K, sodass

$$\psi(f) = \int_K f(x) d\mu \; \forall f \in C(K)$$

4 Definitionen

Def: $I_n := [0,1]^n$(der n-dimensionale Einheitswürfel)
Def: $M(I_n) :=$ Raum der endlichen, signierten, regulären Borel-Maße auf I_n
Def: $\sigma : \mathbb{R} \to \mathbb{R}$ hat die diskriminierende Eigenschaft, wenn für ein Maß $\mu \in M(I_n)$ mit $\int_{I_n} \sigma(y^T x + \theta) d\mu(x) = 0 \; \forall y \in \mathbb{R}^n, \theta \in \mathbb{R}$ folgt, dass $\mu = 0$ ist.

5 Approximationssatz von G.Cybenko(1989)

Satz: Sei $\sigma : \mathbb{R} \to \mathbb{R}$ eine stetige diskriminierende Funktion. Dann sind die endlichen Summen der Form $G(x) = \sum_{j=1}^{N} \alpha_j * \sigma(w_j^T x + b_j)$ mit $N \in \mathbb{N}$ und $\forall$j = 1,...,n : $w_j \in \mathbb{R}^n$ und $\alpha_j, b_j \in \mathbb{R}$ dicht im Raum $C(I_n)$ ausgestattet mit der Supremumsnorm. In anderen Worten, für $f \in C(I_n)$ und $\epsilon > 0$, existiert ein $G(x)$ der obigen Form sodass $|G(x) - f(x)| < \epsilon \; \forall x \in I_n$.

5 Approximationssatz von G.Cybenko(1989)

Satz: Sei $\sigma : \mathbb{R} \to \mathbb{R}$ eine stetige diskriminierende Funktion. Dann sind die endlichen Summen der Form $G(x) = \sum_{j=1}^{N} \alpha_j * \sigma(w_j^T x + b_j)$ mit $N \in \mathbb{N}$ und $\forall$j = 1,...,n : $w_j \in \mathbb{R}^n$ und $\alpha_j, b_j \in \mathbb{R}$ dicht im Raum $C(I_n)$ ausgestattet mit der Supremumsnorm. In anderen Worten, für $f \in C(I_n)$ und $\epsilon > 0$, existiert ein $G(x)$ der obigen Form sodass $|G(x) - f(x)| < \epsilon \; \forall x \in I_n$.

Beweis: Sei $S = \{f(x) = \sum_{j=1}^{N} \alpha_j * \sigma(w_j^T x + b_j) | N \in \mathbb{N}, \forall$j = 1,...,n : $w_j \in \mathbb{R}^n$ und $\alpha_j, b_j \in \mathbb{R}\}$. Dies ist ein linearer Unterraum von $C(I_n)$. Ist der Abschluss von S schon $C(I_n)$ sind wir fertig.

5 Approximationssatz von G.Cybenko(1989)

Satz: Sei $\sigma : \mathbb{R} \to \mathbb{R}$ eine stetige diskriminierende Funktion. Dann sind die endlichen Summen der Form $G(x) = \sum_{j=1}^{N} \alpha_j * \sigma(w_j^T x + b_j)$ mit $N \in \mathbb{N}$ und $\forall$j = 1,...,n : $w_j \in \mathbb{R}^n$ und $\alpha_j, b_j \in \mathbb{R}$ dicht im Raum $C(I_n)$ ausgestattet mit der Supremumsnorm. In anderen Worten, für $f \in C(I_n)$ und $\epsilon > 0$, existiert ein $G(x)$ der obigen Form sodass $|G(x) - f(x)| < \epsilon \; \forall x \in I_n$.

Beweis: Sei $S = \{f(x) = \sum_{j=1}^{N} \alpha_j * \sigma(w_j^T x + b_j) | N \in \mathbb{N}, \forall$j = 1,...,n : $w_j \in \mathbb{R}^n$ und $\alpha_j, b_j \in \mathbb{R}\}$. Dies ist ein linearer Unterraum von $C(I_n)$. Ist der Abschluss von S schon $C(I_n)$ sind wir fertig.

Angenommen also $\bar{S} \neq C(I_n)$. Nach dem oben betrachteten Korollar des Satzes von Hahn-Banach existiert ein beschränktes lineares Funktional $L \neq 0$ auf $C(I_n)$ sodass $L\big|_{\bar{S}} = 0$. Nach dem Satz von Riesz-Markov-Kakutani existiert ein $\mu \in M(I_n)$ sodass $L(h) = \int_{I_n} h(x) d\mu(x) \; \forall h \in C(I_n)$. Für $h \in S$ folgt also $L(h) = 0$. Nach der diskriminierenden Eigenschaft von σ folgt daraus $\mu = 0$. Dies ist ein Wiederspruch zu $L \neq 0$.

Übersicht

- §1 Approximationssatz für Neural Networks
- **§2 Convolutional Neural Network (CNN)**
- §3 Generative Adversarial Network (GAN)
- §4 Moderne Theorie von Neural Networks

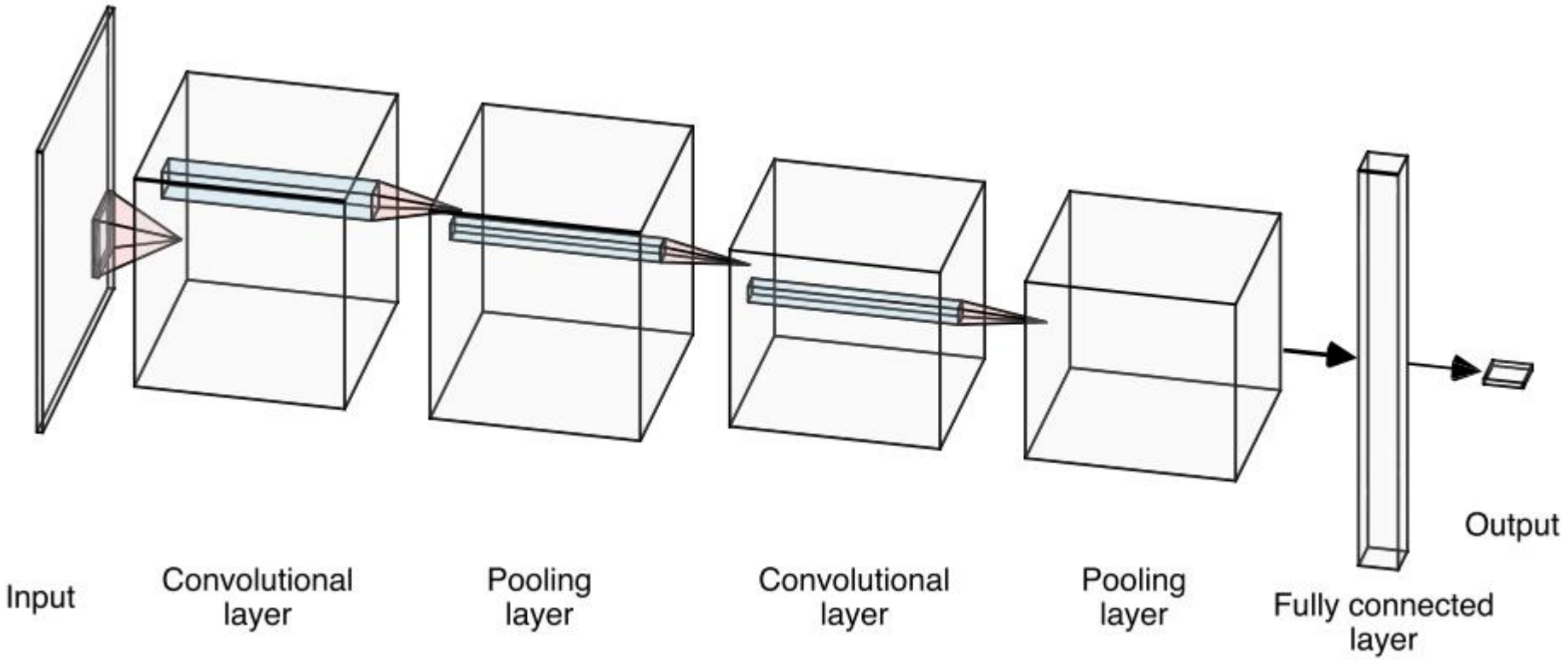
Einführung

- CNNs können räumliche Struktur des Inputs analysieren
- Ursprünglich entwickelt um mit Bildern zu arbeiten
- Besitzen meistens viel weniger Parameter als fully connected feedforward neural networks
- Dies macht effizientes Trainieren sehr tiefer CNNs möglich
- Input Space: $(\mathbb{R}^{n^d})^k$ mit $n, d, k \in \mathbb{N}$
- n heißt Größe, d Dimension und k Tiefe

Beispiel: Input sind Bilder

- Dimension d = 2
- n gibt Größe des Bildes an (n*n ist Pixelanzahl)
- Jedes Pixel wird durch drei Werte RGB bestimmt, damit ist k = 3

Struktur eines CNN



Quelle: Maria Han Veiga, François Gaston Ged, The Mathematics of Machine Learning, de Gruyter, 2024, S.126

Convolution Layer

- Seien $n_{in}, d_{in}, k_{in} \in \mathbb{N}$ die Größe, Dimension und Tiefe des Inputs
- Ein Filter ist ein Element aus $(\mathbb{R}^{n_{fil}^{d_{in}}})^{k_{in}}$ mit $n_{fil} \in \mathbb{N}$ und $n_{fil} < n_{in}$
  das mit den Eingaben gefaltet wird.
- Betrachte Filter g und Eingabe x mit $g \in (\mathbb{R}^{n_{fil}^{d_{in}}})^{k_{in}}$ und $x \in (\mathbb{R}^{n_{in}^{d_{in}}})^{k_{in}}$
- Sei g_j die j-te Tiefe von g
- g gefaltet mit x ergibt $y \in \mathbb{R}^{(n_{in}-n_{fil}+1)^{d_{in}}}$

Convolution Layer

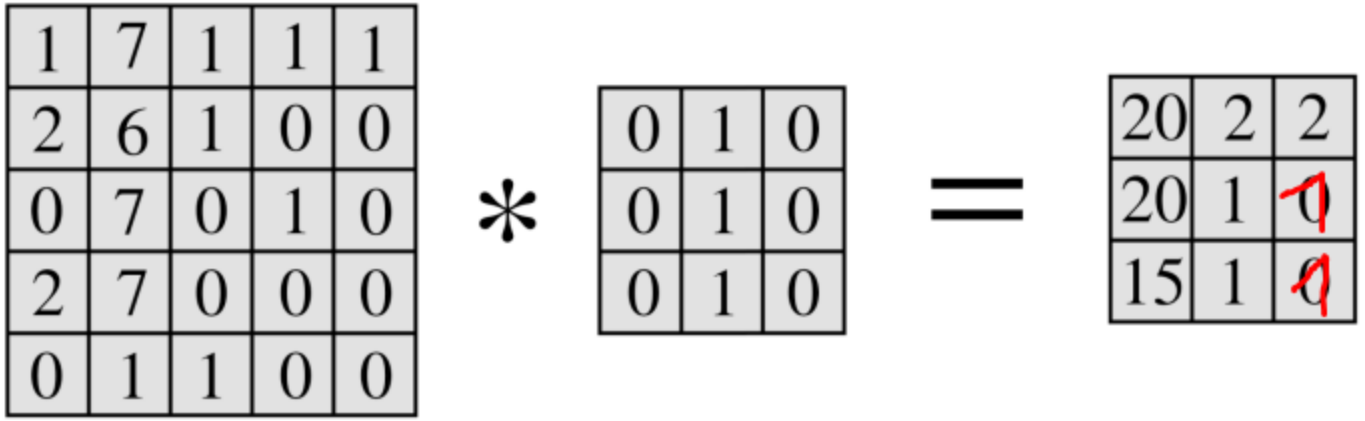
Für alle $i \in \{1, ..., n_{in} - n_{fil} + 1\}^{d_{in}}$ ist der Wert von y an Stelle i gegeben durch:

$$y(i) = (x * g)(i) = \sum_{j \leq k_{in}} \sum_{p_1,...,p_{d_{in}}=1}^{n_{fil}} x_j(i + p - I)g_j(p)$$

,wobei $p = (p_1, ..., p_{d_{in}})$ die Position des Filterpixels und I ein d_{in}-dimensionaler Vektor aus Einsen ist.

Ein Convolution Layer hat oft viele Filter $g^1, g^2, ..., g^l \in (\mathbb{R}^{n_{fil}^{d_{in}}})^{k_{in}}$. Dann ist die Ausgabe des Convolution Layers $y \in (\mathbb{R}^{n_{fil}^{d_{in}}})^l$.

Beispiel



1	7	1	1	1
2	6	1	0	0
0	7	0	1	0
2	7	0	0	0
0	1	1	0	0

$*$

0	1	0
0	1	0
0	1	0

$=$

20	2	2
20	1	0
15	1	0

Figure 8.2: A convolution is applied between a 5×5 pixel image and a 3×3 filter that detects vertical lines. The result of the convolution shows that a vertical line is present on the left of the image.

Eigene Darstellung nach Maria Han Veiga, François Gaston Ged, The Mathematics of Machine Learning, de Gruyter, 2024, S.127

Convolution Layer

- Die trainierbaren Parameter im CL sind die Werte in den Filtern
- #{trainierbaren Parameter} = #Filter*#(Pixel pro Filter)

Pooling Layer

- Zum Reduzieren von der Eingabegröße
- Bildet Pixel in Regionen der Eingabe auf einzelne Pixel ab

Pooling Layer

Sei n_{pool} ein Teiler von n_{in}, dieser wird Kachelgröße genannt.
Ein Pooling Layer teilt in der Tiefe j die $(n_{in})^{d_{in}}$ Pixel der Eingabe in $(n_{in}/n_{pool})^{d_{in}}$ d_{in}-dimensionale Quadrate der Größe $(n_{pool})^{d_{in}}$.
Jeder dieser Quadrate wird auf einen Pixel abgebildet.

Zwei Standardverfahren

- Max pooling:
  - Bilde Quadrat auf maximalen Wert im jeweiligen Quadrat ab
- Average pooling:
  - Bilde Quadrat auf Mittelwert der Werte im jeweiligen Quadrat ab

Beispiel

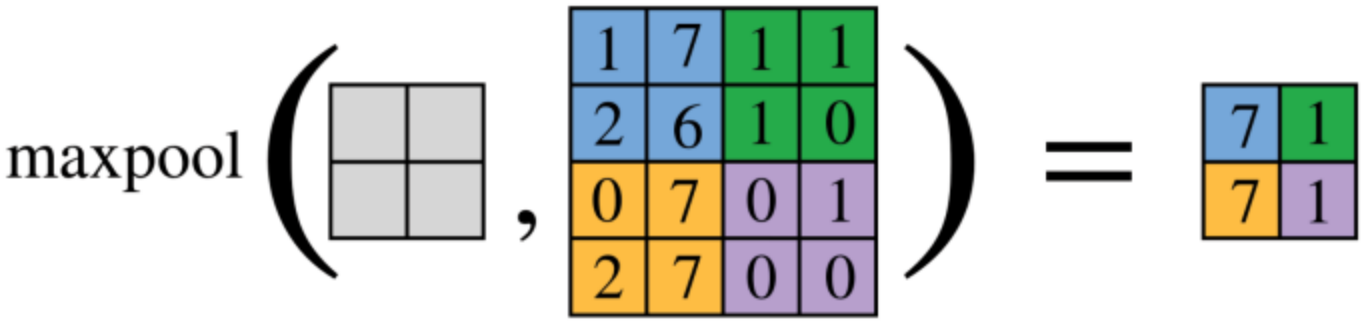


Figure 8.3: A 2×2 max pooling is applied to a 4×4 image, returning a 2×2 matrix with largest pixel value per element of the partition.

Quelle: Maria Han Veiga, François Gaston Ged, The Mathematics of Machine Learning, de Gruyter, 2024, S.128

Übersicht

- §1 Approximationssatz für Neural Networks
- §2 Convolutional Neural Network (CNN)
- **§3 Generative Adversarial Network (GAN)**
- §4 Moderne Theorie von Neural Networks

Frage

- Wenn Sie eine Menge von ähnlichen Objekten gegeben haben, können Sie künstlich ähnliche Objekte erzeugen?

GAN

- $S = \{(x_i, y_i), i = 1, \dots, m\} \subseteq$ X x Y wobei die x_i u.i.v. sind mit Wahrscheinlichkeitsverteilung *D*
- Sei Label y aus Y sowie GAN *G* gegeben
- Ziel: Erzeuge *G*(y) = x aus X, sodass x nach D(x|y) verteilt ist

Probleme

- 1.) *G* soll nicht auf Werte abbilden, die schon in *S* sind
- 2.) Welche Loss-Funktion kann man verwenden?

Vereinfachung

- Vergesse Label
- Reduziere das Problem darauf: Gegeben $S = \{x_i, i = 1, \dots, m\} \subseteq$ X , mit x_i u.i.v. nach *D* verteilt, dann soll *G*(z) nach *D* verteilt sein

Zu 1.)

- Wähle eine einfache, bekannte Wahrscheinlichkeits-Verteilung γ auf niedriger-dimensionalem Raum (d < n)
- Wähle γ absolutstetig bzgl. des Lebesgue-Maßes, damit nicht derselbe Wert doppelt gezogen wird
- Dann wird zufälliges Signal $z \sim \gamma$ an generatives Modell *G* übergeben
- Ziel von *G ist es die Verteilung* γ auf *D* abzubilden, d.h. $G(z) \sim D$

Zu 2.)

Die Diskriminatorfunktion f_{disc} ist ein Klassifizierer, der schätzen soll, ob eine Stichprobe x von der Verteilung D gezogen wurde, oder von G erzeugt wurde.
f_{disc} ist auch ein neural network.

Zu 2.)

Die Diskriminatorfunktion f_{disc} ist ein Klassifizierer, der schätzen soll, ob eine Stichprobe x von der Verteilung D gezogen wurde, oder von G erzeugt wurde.
f_{disc} ist auch ein neural network.
Wahre-Fehlerfunktion: $\mathcal{L}(G, f_{disc}) := \mathbb{E}_{x\sim D}[log f_{disc}(x)] + \mathbb{E}_{z\sim\gamma}[log(1 - f_{disc}(G(z)))]$.
Das GAN will folgendes minimax-Problem lösen: $min_G max_{f_{disc}} \mathcal{L}(G, f_{disc})$

Zu 2.)

- Optimisierung geschieht in 2.Schritten:
- 1.) Update den Diskriminator durch Gradientenaufstiegsverfahren auf den Parametern des Diskriminators
- 2.) Update das GAN durch Gradientenabstiegsverfahren auf den Parametern des GANs

Übersicht

- §1 Approximationssatz für Neural Networks
- §2 Convolutional Neural Network (CNN)
- §3 Generative Adversarial Network (GAN)
- **§4 Moderne Theorie von Neural Networks**

Einführung

- Wir konzentrieren uns hierbei auf überparametrisierte Neural Networks
- Insbesondere der Analyse von Initialisierung und Training
- Es wird sich zeigen, dass allein durch die Veränderung der Gewichtsskala bei Initialisierung unterschiedliche Trainingsregime entstehen

Wiederholung: Gauß-Prozess

Definition(Gauß-Prozess): Sei $\mathcal{X}$ eine nichtleere Menge, sei $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ ein positiv definiter Kern und sei $\mu : \mathcal{X} \to \mathbb{R}$ eine reellwertige Funktion. Dann ist eine zufällige Funktion $f : \mathcal{X} \to \mathbb{R}$ ein Gauß-Prozess (GP) mit Mittelwertfunktion μ und Kovarianzkern K, dessen Verteilung mit $GP(\mu, K)$ bezeichnet wird, wenn für jede endliche Menge $X = (x_1, \ldots, x_m) \in \mathcal{X}^m$ von jeder Größe $m \in \mathbb{N}$, der zufällige Vektor $f_X = (f(x_1), \ldots, f(x_m))^T \in \mathbb{R}^m$ einer mehrdimensionalen Normalverteilung $\mathcal{N}(\mu_X, \underline{K}_{XX})$ folgt, mit Kovarianzmatrix $\underline{K}_{XX} = (K(x_i, x_j))_{i,j=1}^m \in \mathbb{R}^{m \times m}$ und Mittelwertvektor $\mu_X = (\mu(x_1), \ldots, \mu(x_m))^T \in \mathbb{R}^m$.

Wiederholung: Gauß-Prozess

- Definition(RBF-Kern): Der RBF-Kern (radial basis function kernel) ist definiert durch $K_{RBF}(x, x') = exp(-\frac{\|x-x'\|}{2l^2})$ mit der Längenskalierung $l \in \mathbb{R}^+$.
- Für Punkte, die nahe beieinander liegen, wertet der Kern zu einem Wert nahe 1 aus.
- Für Punkte, die weit voneinander liegen, wertet der Kern zu einem Wert nahe 0 aus.
- $\Rightarrow$ Werte für $f \sim GP(0, K_{SE})$ bei Punkten, die nahe beinander liegen, sind stark korreliert
- $\Rightarrow$ Werte für $f \sim GP(0, K_{SE})$ bei Punkten, die weit voneinander liegen, sind schwach korreliert

Modellannahmen

- Betrachte Neural Network $f_w : \mathbb{R} \to \mathbb{R}$
- Definiert durch $f_w(x) := \frac{1}{d^\gamma} w_2 \sigma(w_1 x + b)$
- $w_1, b \in \mathbb{R}^{d \times 1}$ und $w_2 \in \mathbb{R}^{1 \times d}$ als trainierbare Parameter
- $\gamma > 0$ als Hyperparameter
- σ als Aktivierungsfunktion
- d ist Weite der verborgenen Schicht

Modellannahmen

- Betrachte Neural Network $f_w : \mathbb{R} \to \mathbb{R}$
- Definiert durch $f_w(x) := \frac{1}{d^\gamma} w_2 \sigma(w_1 x + b)$
- $w_1, b \in \mathbb{R}^{d \times 1}$ und $w_2 \in \mathbb{R}^{1 \times d}$ als trainierbare Parameter
- $\gamma > 0$ als Hyperparameter
- σ als Aktivierungsfunktion
- d ist Weite der verborgenen Schicht
- Zur Initialisierung sind Komponenten von w_1, w_2 und b unabhängig, identisch standardnormalverteilt
- Schreibe Vektoren aus: $f_w(x) := \frac{1}{d^\gamma} \sum_{k=1}^{d} w_{2,k} \sigma(w_{1,k} x + b_k)$

Initialisierungsskala

Satz: Sei f_w ein Neural Network bei der Initialisierung, welches wie oben definiert ist, mit $\gamma = 1/2$. Sei außerdem σ Lipschitz-stetig. Dann konvergiert die Verteilung von f_w für $d \to \infty$ nach $GP(0, K^{(2)})$, wobei der Kern $K^{(2)}$ gegeben ist durch:

$$K^{(2)}(x, x') := \mathbb{E}_{g \sim GP(0, K^{(1)})}[\sigma(g(x))\sigma(g(x'))]$$

$$\text{mit } K^{(1)}(x, x') := xx' + 1.$$

Initialisierungsskala

- Damit wissen wir, dass unendlich breite Neural Networks der obigen Form und Initialisierung Gauß-Prozesse sind
- Damit wird das a-priori-Wissen des Modells charakterisiert

Neural Tangent Kernel

- Betrachte das oben definierte Neural Network f_w mit $\gamma = 1/2$
- w_t notiert die Parameter des Neural Networks zur Zeit t des Trainings
- Die Schritte des Gradientenabstiegsverfahrens sind gegeben durch:

$$w_{t+1} = w_t - \eta \frac{1}{m} \sum_{i=1}^{m} \partial_{\hat{y}} \ell(y_i, \hat{y})\big|_{\hat{y}=f_{w_t}(x_i)} \nabla_w f_{w_t}(x_i)$$

- $\eta > 0$ ist die Lernrate
- Frage: Wie entwickelt sich f_{w_t} mit der Zeit ?

Neural Tangent Kernel

- Durch Taylor-Entwicklung von $f_{w_{t+1}}$ ergibt sich:

$$f_{w_{t+1}}(x) \approx f_{w_t}(x) + (w_{t+1} - w_t) \cdot \nabla_w f_{w_t}(x)$$

- Durch die beiden früheren Gleichungen ergibt sich:

$$f_{w_{t+1}}(x) - f_{w_t}(x) \approx -\eta \frac{1}{m} \sum_{i=1}^{m} \partial_{\hat{y}} \ell(y_i, \hat{y})\big|_{\hat{y}=f_{w_t}(x_i)} \nabla_w f_{w_t}(x_i) \cdot \nabla_w f_{w_t}(x)$$

Neural Tangent Kernel

- Durch Taylor-Entwicklung von $f_{w_{t+1}}$ ergibt sich:

$$f_{w_{t+1}}(x) \approx f_{w_t}(x) + (w_{t+1} - w_t) \cdot \nabla_w f_{w_t}(x)$$

- Durch die beiden früheren Gleichungen ergibt sich:

$$f_{w_{t+1}}(x) - f_{w_t}(x) \approx -\eta \frac{1}{m} \sum_{i=1}^{m} \partial_{\hat{y}} \ell(y_i, \hat{y})\big|_{\hat{y}=f_{w_t}(x_i)} \nabla_w f_{w_t}(x_i) \cdot \nabla_w f_{w_t}(x)$$

- Das Skalarprodukt auf der rechten Seite der obigen Gleichung definiert einen von der Zeit t und der Breite d abhängigen Kern:

$$\Theta_t^{(d)}(x, x') := \nabla_w f_{w_t}(x) \cdot \nabla_w f_{w_t}(x')$$

- $\Theta_t^{(d)}$ ist der Neural Tangent Kernel des Neural Networks zur Zeit t
- Der NTK bei Initialisierung $\Theta_0^{(d)}$ ist ein zufälliger Kern, da die Initialisierung zufällig ist

Neural Tangent Kernel

Satz: Für obige Modellannahmen und σ Lipschitz und zweimal differenzierbar mit beschränkter zweiter Ableitung existiert ein deterministischer Kern $\Theta^{(\infty)} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, sodass $\forall x, x' \in \mathbb{R}$ und $\forall T \in \mathbb{R}^+$ fast sicher gilt:

$$lim_{d\to\infty} \sup_{t\leq T} |\Theta_t^{(d)}(x, x') - \Theta^{(\infty)}(x, x')| = 0.$$

Der deterministische Kern hat folgende Darstellung:

$$\Theta^{(\infty)}(x, x') = K^{(1)}(x, x')\dot{K}^{(2)}(x, x') + K^{(2)}(x, x'),$$

wobei $K^{(1)}$ und $K^{(2)}$ die Kerne aus dem Satz über die Initialisierungsskala sind. $\dot{K}^{(2)}$ ist wie $K^{(2)}$ definiert mit der Ableitung σ' anstelle von σ in der Definition.

Neural Tangent Kernel

- Ein weiterer bedeutender Satz kommt von Jacot et al. (2018) und zeigt:

$$\partial_t f_{w_t}(x) = -\frac{1}{m} \sum_{i=1}^{m} \partial_{\hat{y}} \ell(y_i, \hat{y})\big|_{\hat{y}=f_{w_t}(x_i)} \Theta_t^{(d)}(x_i, x).$$

- Für den Grenzwert $d \to \infty$ wird die Trainingsdynamik gegeben durch:

$$\partial_t f_{w_t}(x) = -\frac{1}{m} \sum_{i=1}^{m} \partial_{\hat{y}} \ell(y_i, \hat{y})\big|_{\hat{y}=f_{w_t}(x_i)} \Theta^{(\infty)}(x_i, x).$$

Neural Tangent Kernel

- Angenommen die Loss-Funktion ist: $\ell(y, y') = \frac{1}{2}(y - y')^2$
- Angenommen $(\Theta^{(\infty)}(x_i, x_j))_{1 \leq i,j \leq m}$ ist positiv definit
- Dann besitzt die Differentialgleichung:

$$\partial_t f_{w_t}(x) = -\frac{1}{m} \sum_{i=1}^{m} \Theta^{(\infty)}(x, x_i)(f_{w_t}(x_i) - y_i)$$

  eine explizite Lösung

- Insbesondere konvergiert f_{w_t} für $t \to \infty$ gegen:

$$f_{w_\infty}(x) = f_{w_0}(x) - \sum_{i,j=1}^{m} \Theta^{(\infty)}(x, x_i)(\Theta^{(\infty)})^{-1}(x_i, x_j)(f_{w_0}(x_j) - y_j)$$

- f_{w_∞} interpoliert die Trainingsdaten exakt

Quellen

- Maria Han Veiga, François Gaston Ged, The Mathematics of Machine Learning, de Gruyter, 2024
- G. Cybenko, Approximation by superpositions of a sigmoidal function, Mathematics of Control, Signals and Systems, 2(4), S. 303-314, 1989