Why deep neuronal networks are difficult to train

Mashine Learning Seminar

Christian Gommeringer

betreut durch Prof. Schreiber

Tübingen, den 7. Juni 2023

Motivation

In unserem heutigen Vortrag werden wir die ersten Schritte ausgehend von den bisher betrachteten flacheren Netzwerken mit nur einem Hidden Layer hin zu tiefen neuronalen Netzen mit vielen Layern, die in der Anwendung oft benutzt werden, unternehmen. Mit den bisherigen Netzwerken haben wir ja bereits gute Erebnisse erziehlt. Es lassen sich allerdings Gründe finden, wieso die Leistung von Deep Learning höher sein kann. Zum einen lässt sich das Konzept meherer auf einander aufbauender Level gut mit der natürlichen Herangehensweise identifizieren, mit der man ausgehend von grundlegenden Bausteinen sein Modell mit jedem Abstraktionsniveau weiter verfeinert.
Es lässt sich auch zeigen, dass durch Benutzung mehrer Layer von stückweise linearen Neuronen sich eine Outpunktfunktion generieren lässt, bei der die Anzahl an linearen Teilen, in die man die Funktion aufteilen kann, größer ist als bei flachen Netzwerken. Dadurch kann die Outpunktfunktion kompliziertere Funktionen annähern, was im Ende ja das Ziel von Mashine Learning ist (je komplizierter die Funktion ist, desto feinteiliger kann man unterscheiden). Der Beweis kann mit Hilfe der Geometrie von Schnitten von Räumen durch lineare Manigfaltigkeiten (oder so ähnlich :) geführt werden. Diesen Beweis werde ich nicht vorstellen. Ich möchte aber eine Vorstellung dafür vermitteln, in dem ich den Fall für die rectified Aktivierungfunktion, die im nächsten Kapitel auch des öfteren zur Anwendung kommt, untersuche. Die Aktivierungsfunktion rectified ist einfach die im Negativen abgeschnittene Identität.

$$id_{\text{rect}}(x) = \begin{cases} x & \text{wenn } x > 0 \\ 0 & \text{sonst} \end{cases}$$

Ich werde id_{rect} auch komponentenweise auf einen Vektor angewandt verwenden. Ich wähle hier die Größe von input, 1. hiddden und 2. hidden Layer gleich und bezeichne sie mit n. W_1, W_2, b_1 und b_2 sind die entsprechenden Gewichtsmatrizen und Vektoren. Untersuchen wir zunächst das Output a_1 des ersten hidden Layers, das sich folgendermaßen darstellt

$$a_1 = id_{\text{rect}}\ (W_1\ + b_1)$$

Wenn W_1 eine Invertierbare Matrix ist wird jeder Punkt im N-dimensionalen Vektorraum erreicht. Es ergeben sich verschiedene lineare Teilbereiche. Ein linearer Teilbereich des Inputraums ist der, bei welchem die Output Vektoren nur positive Komponenten besitzt. Die nächsten n Teilgebiete sind die

bei denen jeweils nur eine Komponente negativ ist. Dann gibt es noch die Teilgebiete, bei denen 2 Komponenten negativ sind und so weiter. Es gibt also insgesamt

$$N_1 = \sum_{k=0}^{n} \binom{n}{k} = \sum_{k=0}^{n} \binom{n}{k} 1^k \, 1^{n-k} = 2^n$$

lineare Teilgebiete. für diese linearen Teilmengen gilt die Linearitätsbedingung

$$a_1 \left(\lambda \, x + \mu \, x'\right) = \lambda \, a_1(x) + \mu \, a_1(x')$$

für

$$\lambda, \mu \geqslant 0$$

, da so die Eigenschaft der Komponenten, ob sie positiv oder negativ sind nicht verändert werden. Diese Mengen sind zusammenhängend, wessen man sich auf folgende Weise vergewissern kann. Die Funktion

$$g : \mathbb{R}_{>0} \times \mathbb{R}_{>0} \to \mathbb{R}^n : (q, p) \mapsto \frac{1}{q+p} \, \left(q \, x_1 + p \, x_2\right)$$

für zwei Punkte $x_1,\ x_2$ im selben Linearitätsgebiet ist eine stetige Funktion, bei der das Output die obige Linearitätsbedingung erfüllt und für die gilt

$$g(1,0) = x_1$$
$$g(0,1) = x_2$$

Es gibt also einen stetigen Weg von x_1 nach x_2, der ganz im Linearitätsgebiet liegt, wodurch diese Menge wegzusammenhängend und somit auch zusammenhängend ist. Außerdem liegt für ein x auch die gesamte Strecke vom Ursprung ins Unendliche im Linearitätsgebiet (wieder aufgrund der obigen Linearitätsbedingung). Trennflächen zwischen den Gebieten können somit nur Ebenen sein, die den Ursprung schneiden. Mit dieser Vorstellung können wir weiter mit der Betrachtung des zweiten Layers fortfahren. Die Bildmenge des ersten hidden Layers ist die Teilmenge des $\mathbb{R}^n$, in der alle Komponenten größer gleich Null sind, wobei der Bereich, in dem alle Komponenten strikt größer als Null sind, ein einziges Linearitätsgebiet ist, und die die anderen Linearitätsgebiete Teile der Seitenflächen sind. Schauen wir uns als Beispiel

die Konfiguration

$$\begin{aligned} W_1 &= id \\ b_1 &= 0 \\ W_2 &= Rot_z(\alpha) \\ b_2 &= 0 \end{aligned}$$

für $n = 3$ an. Hier sind die Linearitätsgebiete die kanonischen achtel des $\mathbb{R}^3$. Das Output von Luyer 1 ist dann, wie zuvor im $\mathbb{R}^3_{\geqslant 0}$. Wenn wir im 2. Layer nun vor der rectified Aktivierung um einen kleinen Winkel um die z-Achse drehen, und dann mit rectify aktivieren, führen wir in der Nähe der yz-Ebene ein neues Linearitätsgebiet ein, während wir die bisherigen Linearitätsgebiete erhalten. Wir können innerhalb des zweiten Layers noch zusätzlich um weitere Achsen drehen, oder mögliche andere lineare Operationen durchführen, um weitere Linearitätsgebiete zu erzeugen. Auf so eine Art und weise könnte man in den nächsten Layer fortfahren und eine immer komplexere Funktion generieren. Wir betrachten uns zur Veranschaulichung die Vektoren

$$\begin{aligned} x_1 &= \begin{pmatrix} a \\ -\varepsilon \end{pmatrix} \\ x_2 &= \begin{pmatrix} b \\ c \end{pmatrix} \end{aligned}$$

mit a, b, c, $\varepsilon > 0$. Nach dem ersten Layer, das die Identität in der Aktivierungsfunktion enthält, werden die Vektoren folgendermaßen umgeformt.

$$\begin{aligned} a_1(x_1) &= (a, 0) \\ a_1(x_2) &= (b, c) \end{aligned}$$

und nach dem zweiten Layer

$$\begin{aligned} a_2(x_1) &= id_{\text{rect}}\left(\begin{pmatrix} \cos(\alpha)\ a \\ \sin(\alpha)\ a \end{pmatrix}\right) \\ a_2(x_2) &= id_{\text{rect}}\left(\begin{pmatrix} \cos(\alpha)\ b - \sin(\alpha)\ c \\ \sin(\alpha)\ b + \cos(\alpha)\ c \end{pmatrix}\right) \\ a_2(x_1) + a_2(x_2) &= id_{\text{rect}}\left(\begin{pmatrix} \cos(\alpha)\ a \\ \sin(\alpha)\ a \end{pmatrix}\right) + id_{\text{rect}}\left(\begin{pmatrix} \cos(\alpha)\ b - \sin(\alpha)\ c \\ \sin(\alpha)\ b + \cos(\alpha)\ c \end{pmatrix}\right) \end{aligned}$$

während

$$a_1(x_1 + x_2) = (a + b, c - \varepsilon)$$

für ε klein genung, und

$$a_2(x_1 + x_2) = id_{\text{rect}} \left(\begin{pmatrix} \cos(\alpha)\ (a + b) - \sin(\alpha)\ (c - \varepsilon) \\ \sin(\alpha)\ (a + b) + \cos(\alpha)\ (c - \varepsilon) \end{pmatrix} \right)$$

Man erkennt hier zum einen, dass bestehende Linearitätsgebiete nicht wieder zusammen geführt wurden, da sich $a_2(x_1 + x_2)$ von $a_2(x_1) + a_2(x_2)$ durch den ε-Term unterscheidet. Zum andern ist zu sehen, dass sich ein neues Linearitätsgebiet ergibt. Für eine groß genuge Drehung α gibt es nähmlich Punkte

$$x_1 = (a, b)$$
$$x_2 = (a', b')$$

mit a, b, a', $b' > 0$, für die gilt

$$Rot_\alpha(a_1(x_1))_0 > 0$$
$$Rot_\alpha(a_1(x_2))_0 < 0$$

Es unterscheiden sich also die Vorzeichen der 0-Komponenten vor der Anwendung der Aktivierungsfunktion des 2. Layers, wodurch diese beiden Vektoren nach dem 2. Layer in unterschiedliche Linearitätsgebiete fallen.

Verschidene Verfahren zur Verbesserung des Lernverhaltens tiefer neuronaler Netze

In diesem Abschnitt stelle ich zunächst den Vorteil zweier verbesserterter Optimierungsverfahren vor, der Hessian-free Optimierung und Nesterov's Accelerated Gradient (NAG). Danach stelle ich kurz eine leicht verbesserte Inintialisierungsmethode vor und zum Ende gehe ich noch ein wenig auf unsupervised pre-training ein.
Beginnen wir mit dem Ablauf der Hessian-free Methode und mit einer Betrachtung zu dessen Vorteil bei allgemeinen Optimierungsproblemen. Die Kernidee dieses Verfahrens ist es die Funktion bis zur zweiten Ordnung zu approximieren

$$f(x) \approx T_2(x) = f(p) + \nabla f(p) \cdot x + \frac{1}{2} x^T H x$$

und dann die Funktion T_2 zu minimieren, was die Lösung einer Gleichung der Form

$$\nabla f(p) = H\,x$$

erfordert. Es sei noch erwähnt, dass in der Praxis H oft noch etwas modifiziert wird, um die Leistung noch zu steigern. Die Schwierigkeit zur Lösung dieser Gleichung, besteht auch aber nicht haupsächlich in der Bestimmung von H, für die jedoch relativ gute Algorithmen vorhanden sind. Sie besteht zu einem großen Teil darin, dass gewöhnliche Algorithmen oft zur Lösung dieses Problems die Matrix H zu invertieren, was viel kostet. Der Hessian-free Algorithmus umgeht diese Probleme, indem er zunächst

$$H\,x = \frac{\nabla f(p + \varepsilon\,x) - \nabla f(p)}{\varepsilon} \tag{1}$$

für ein kleines ε approximiert und dann die Lösung der quadratischen Gleichung der Minimierung von T_2 mittels des konjugierten Gradientenabstiegsverfahrens mit einigen wenigen Iterationen annähert. Die Lösung der Gleichung ist somit nicht exakt, genügt jedoch um die Vorteile eines Hessverfahrens gegenüber eines einfachen Gradientenabstiegs zu erhalten. Diese liegen nämlich in der Analyse des Krümmungsverhaltens durch den Algorithmus. Nehmen wir an die Lösung von Gleichung (1) liegt in Richtung eines Eigenvektors d. Dann lässt sich die die Lösung schreiben als

$$x = \frac{\nabla f(p)}{|\frac{\partial^2 f(p)}{\partial\, d^2}|}$$

Im Update wird der Gradient also mit der Krümmung in diese Richtung gewichtet.

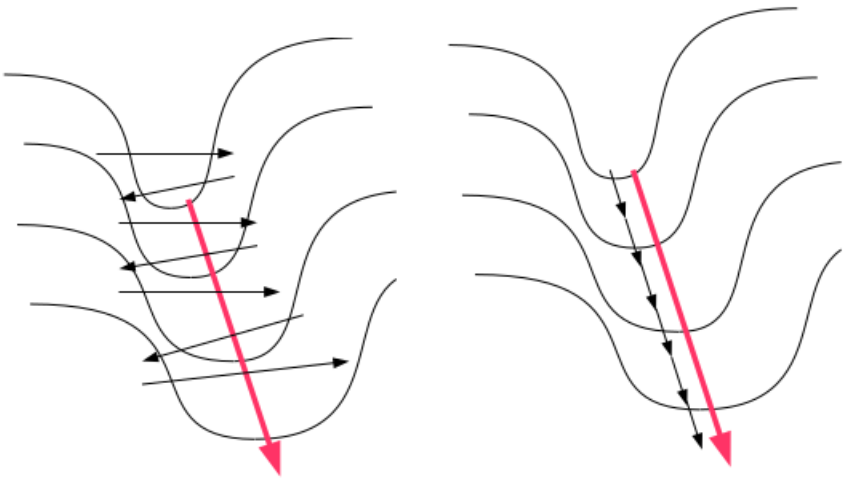

Abbildung 1

Wenn wir obiges Bild betrachten, sehen wir eine Situation, in der ein Hess Verfahren von Vorteil ist. Es handelt sich hierbei um ein steil eingeschnittenes Tal, das auf seiner Sohle flach aber auf einer großen Entfernung abfällt. Die Steigung vertikal zum Verlauf des Tals ist hier größer, weshalb der gewöhnliche Gradientenabstieg einen größeren Schritt in diese Richtung macht als in die parallele und deshalb oft von den Talseiten hin und herspringt. Der Hessian-free Algorithmus wird zusätzlich mit der Krümmung gewichtet und macht daher größere Schritte entlang des Tals, wodurch er in weniger Schritten zum Minimum findet. Wir können uns plausibel machen, dass solche Situationen nicht selten sind, da es sinnvoll erscheint, dass eine weniger gektrümmte Steigung einen nachhaltigeren Abfall bewirkt. Die Hessian-free Optimierung ist für das Trainieren tiefer neuronaler Netze im besonderen von Vorteil, weil er auch beim Kampf gegen das unstable Gradien Problem hilft. Wir erinnern uns, dass der Grund darür die Rekursion zur berechnung der Gradienten verschiedener Layer ist, bei der man viele potentiell kleine Zahlen mit einander multipliziert.
Mit der Notation aus dem Buch für die Aktivierung nach dem l-ten Layer a^l und

$$z^l = W^l \, a^{l-1} + b^l$$

lässt sich die Ableitung der Kostenfunktion schreiben als

$$\frac{\partial C}{\partial z^i} = \frac{\partial C}{\partial z^L} \prod_{k=L-1}^{i} \frac{\partial z^{k+1}}{\partial z^k}$$

für die zweite Ableitung ergibt sich daraus

$$\begin{aligned}\frac{\partial^2 C}{\partial z^j \partial z^i} =& \frac{\partial^2 C}{\partial z^j \partial z^L} \prod_{k=L-1}^{i} \frac{\partial z^{k+1}}{\partial z^k} \\ &+ \frac{\partial C}{\partial z^L} \sum_{\max\{j,i\}}^{L-1} \left[\left(\prod_{k=L-1,\, k\neq l}^{i} \frac{\partial z^{k+1}}{\partial z^k} \right) \frac{\partial^2 z^{l+1}}{\partial z^j \partial z^l} \right]\end{aligned}$$

Diese Ausdrücke schätze ich ab in dem alle Gewichte ungefähr von der gleichen Größenordnung sind und schreibe.

$$\frac{\partial C}{\partial z^i} = O(w^{L-i})$$

die zweite Ableitung ist in diesem Fall auch in einer ähnlichen Größenordnung.

$$\frac{\partial^2 C}{\partial z^j \partial z^i} = \sum_{k=i}^{L} O(w^{L-i+\max\{0,k-j\}}) = O(z^{L-j})$$

Hier ist zu sehen, dass erste und zweite Ableitung z.b. nach dem ersten hidden Layer von gleicher Größenordnung bezüglich der Gewichte sind (und beinahe gleicher Größenordnung mit berücksichtigung nicht stückweise linearer Aktivierungsfunktionen). Da beim Hessian-free Algorithmus noch durch die zweite Ableitung geteilt wird, wirkt das dem unstable Gradient Problem entgegen.

Als zweiten vorteilhaften Algorithmus möchte Nesterov's Accelerated Gradient Algorithmn vorstellen. Es handelt sich hierbei um eine leichte veränderung des Gradientenabstiegs mit Momentum. Das Update Schema lautet

$$\begin{aligned} v_{t+1} &= \mu\, v_t + \epsilon\, \nabla f(x_t + \mu\, v_t) \\ x_{t+1} &= x_t + v_{t+1} \end{aligned}$$

Man kann sich an dem Talbeispiel von vorhin veranschaulichen, dass dieses Schema auch die Krümmung der Landschaft berücksichtigen kann. Das Update Schema erlaubt es dem Algorithmus nämlich ein wening voraus zu schauen und der Gradient in Richtung senkrecht zum Talverlauf würde durch Eingbeziehung der anderen Seite des Tal, auf der er im nächsten Schritt landen würde, Korriegiert. Oszillationen können sich auf diese Weise nicht konstruktiv aufsummieren. Auch dem vorherschenden vanishing Gradient Problem kann dieser Algorithmus entgegenwirken, da er erlaubt, dass sich die Gradienten über die Zeit aufsummieren. Auch hier erlangen nachhaltige Abstiege ein größeres Gewicht und unnötige Oszillationen summieren destruktiv. Ein großer Vorteil dieses Algorithmus ist, dass er weniger aufwändig ist, da er mit viel weniger Gradienten Auswertungen auskommt. In der Praxis hat dieser Algorithmus dennoch gute Ergebnisse erbracht.

Initialisierung

Wie schon im Vortrag von Peter und David, besprochen ist es wichtig für das anfängliche Lernen des Systems, wie die Gewichte am Anfang verteilt sind. Für tiefe neuronale Netzwerke muss die Varianz der Anfangsverteilung leicht angepasst werden im Vergleich zu flachen Netzen. Wie Bengio und Glorot zeigten lässt sich die fortgepflanzte Varianz der Gradienten der Gewichte -als Kennzeichen deren Größe- schreiben als

$$Var[\frac{\partial C}{\partial W^i}] = (n\,Var[W])^L\,Var[x]\,Var[\frac{\partial C}{\partial z^L}]$$

wobei die n die Zahl der Neuronen in einem Layer ist, die hier als Gleich für alle Layer angenommen wird. Um einen verschwindenden oder explodierenden Gradient für großes L zu verhindern, ist zu wünschen, dass

$$n\,Var[W] = 1$$

ist. Um dies zu erreichen, kann W^i aus einer gleicförmigen Verteilung aus dem Intervall

$$\left[-\sqrt{\frac{6}{n_i + n_{i+1}}}, \sqrt{\frac{6}{n_i + n_{i+1}}}\right]$$

gezogen werden.