Woche 10 Praktische Anwendungsbeispiele: Know Your Data & TensorFlow Playground

Skript

Erarbeitet von
Joana Grah

Lernziele

- Know Your Data benutzen können, um verschiedene Datensätze zu explorieren
- Durch den TensorFlow Playground eine bessere Intuition dafür erhalten, wie ein Neuronales Netzwerk und das Training funktionieren

Inhalt

Einstieg

Diese Woche soll es in dem Video zu Praktischen Anwendungsbeispielen für euch auch ganz praktisch werden. Auf der einen Seite möchten wir euch gerne eine Webseite vorstellen, mit Hilfe derer ihr euch Datensätze genauer angucken könnt und auf der anderen Seite sollt ihr euch ganz praktisch mit neuronalen Netzwerken beschäftigen und damit herumspielen können.

Projekt „KI für alle"
HeiCAD – Heine Center for Artificial Intelligence and Data Science

Es geht bei der Webseite heute exemplarisch um den MNIST-Datensatz, den ihr ja schon mittlerweile kennt. Der besteht ja aus Bildern von handgeschriebenen Ziffern und wird normalerweise zur Klassifikation benutzt. Es gibt also zehn Klassen für die zehn verschiedenen Ziffern.

Know Your Data

Schauen wir uns doch mal die Webseite an, die heißt nämlich „Know Your Data“. Hier sehen wir schon auf einen Blick verschiedene Datensätze. Wir sehen, wenn wir weiter runterscrollen, auch schon MNIST-ähnliche Datensätze und irgendwo hier sehen wir dann tatsächlich den MNIST-Datensatz. Wir sehen schon direkt auf den ersten Blick, dass der aus 70.000 verschiedenen Bildern in dem Fall besteht und eine kleine Vorschau davon, wie diese Bilder aussehen. Wenn wir jetzt hier auf „See dataset“ klicken würden, dann würden wir weitergeleitet werden zur entsprechenden Tensorflow-Website. Das machen wir jetzt nicht. Wir möchten gerne uns das hier näher anschauen auf dieser Website und deswegen klicken wir auf „Explore in Know Your Data“.

Was wir dann sehen, ist diese Webseite hier, also rechts sehen wir schon eine Vorschau von Bildern von handgeschriebenen Ziffern, die in dem Datensatz enthalten sind, in dem Fall für das Label „0“, also wir sehen verschiedene handgeschriebene Nullen. Und wir sehen, dass wir jetzt hier gerade bei dem Reiter „Stats“ sind.

Also was wir uns hier anschauen können, ist z. B. das: „Source features“. Wenn wir das öffnen, sehen wir, dass wir auf der einen Seite hier links auf den ersten Blick die verschiedenen Labels sehen können, also die Aufteilungen. Wir sehen: von den 70.000 Bildern, die in dem Datensatz enthalten sind, sind immer ca. 7.000 Bilder pro Ziffer enthalten, d. h. das ist ein sehr ausgewogener Datensatz, und so soll’s ja eigentlich auch sein. D. h. jede Klasse beinhaltet ungefähr die gleiche Anzahl an Bildern und dann auf der rechten Seite können wir uns z. B. den Split anschauen. Der Datensatz ist ja schon voraufgeteilt in Test- und Trainingsdaten und wir sehen, dass es 10.000 Testbilder gibt und 60.000 Trainingsbilder.

Was wir uns z. B. auch anschauen können, ist hier die „Image Quality“, die Bildqualität. Da kann man auch auf einen Blick sehen, wie denn da diese Bilder bewertet sind mit verschiedenen Scores, z. B. „Exposure Quality“ oder der „Sharpness Score“. Und dann gibt es noch andere verschiedene Möglichkeiten, wir wollen uns aber jetzt ein bisschen auf die rechte Seite hier noch konzentrieren.

Was wir nämlich beispielsweise machen können, ist das nicht nach „label“ auszuwählen, sondern z. B. nach „split“, und dann könnten wir uns nur Trainings- oder nur Testbilder anzeigen lassen. Was vielleicht interessant ist, ist die Bilder zu gruppieren, wie hier z. B. eben wieder nach einem von den zehn verschiedenen Labels. Also wenn ich jetzt sage, ich möchte z. B. sehen, wie die Bilder von den Fünfen aussehen, dann kann ich hier auf dieses Bild klicken und bekomme eben eine Übersicht über verschiedene Bilder, die in dem Datensatz enthalten sind, die die Ziffer „5“ zeigen. Und auch da kann ich das dann nochmal

in Trainings- und Testbilder splitten und sagen, ich möchte mir jetzt aber gerne z. B. die Testbilder anschauen. Dann klicke ich darauf.

Was vielleicht auch noch interessant ist, ist hier oben dieses Tab „Item“. Wenn ich jetzt auf ein Bild draufklicke, dann sehe ich hier links eben ein paar Metadaten über das Bild, also z. B., dass es im PNG-Format abgespeichert ist, oder – was wir mittlerweile ja auch schon wissen – dass die Höhe und Breite jeweils 28 Pixel beträgt.

TensorFlow Playground

Die zweite Website, die wir euch heute vorstellen möchten, ist der „TensorFlow Playground“ und da ist der Name auch schon quasi Programm, das ist nämlich tatsächlich eine Art Spielwiese, die ihr verwenden könnt, um mit neuronalen Netzwerken herumzuspielen und das mal auszuprobieren in der Praxis.

Also, wie ist die Seite aufgebaut? Wir sehen vielleicht schon auf den ersten Blick, dass das hier in der Mitte ein bisschen aussieht wie ein neuronales Netzwerk, also wir sehen in dem Fall jetzt keine runden, aber quadratischen Neuronen, die verbunden sind. Auf der rechten Seite sehen wir höchstwahrscheinlich Datenpunkte, also wir sehen ein Koordinatensystem, in dem Punkte aufgezeichnet sind. Und hier auf der linken Seite und hier oben können wir ein paar Parameter verstellen.

Fangen wir vielleicht einfach mal direkt hier oben links an. Das spezifiziert das Datenset, mit dem wir uns beschäftigen, also wir haben jetzt hier einmal dieses und haben eigentlich vier verschiedene zur Auswahl. Ich klicke jetzt nochmal auf die anderen. Wir sehen, dass sich das dann hier rechts verändert, also es gibt diese Punktwolke, dann gibt es noch das Datenset mit zwei Klassen, und außerdem dieses.

Ich habe jetzt schon das Wort Klassen benutzt. Wenn wir jetzt hier einmal nach oben rechts wechseln, können wir nämlich auch den Problemtyp auswählen. Also ich habe jetzt hier ein Klassifikationsproblem ausgewählt, wir können aber auch z. B. eine Regression auswählen. Dann gibt es allerdings nur zwei Datensets, die zur Verfügung stehen. Ich möchte mir aber jetzt gerne hier mit euch ein Klassifikationsproblem anschauen, und zwar bleibe ich mal direkt bei diesem Problem hier oben links, bei dem ersten Beispiel.

Hier drunter sehen wir auch ein paar Sachen, die wir verstellen können. Das erste müsste euch bekannt vorkommen. Das ist nämlich wieder das Verhältnis zwischen Trainings- und Testdaten. Das ist momentan auf 50 % gesetzt, das kann ich auch niedriger oder höher stellen. Dann gibt es hier noch Noise oder Rauschen, also hier rechts sehe ich z. B., wenn ich das höher stelle, dann sieht man, dass sich die Punkte der Klassen doch irgendwie ziemlich miteinander mischen. Das stelle ich jetzt wieder auf 0.

Und dann so ein paar Begriffe wie z. B. Batch size, Epoch, Learning rate werdet ihr nächste Woche kennenlernen und dann könnt ihr ja vielleicht nochmal zurück zu der Webseite gehen und euch das näher anschauen. Was wir schon kennen, ist hier die

Aktivierungsfunktion. Da wählen wir jetzt die aus, die wir kennen, nämlich die ReLU, die Rectified Linear Unit. Die Regularisierung oder Regularisation ignorieren wir jetzt im Moment auch noch, das ist jetzt gerade nicht so wichtig für uns, das lassen wir noch ausgestellt.

Dann können wir uns jetzt mal mit dem neuronalen Netzwerk an sich beschäftigen. Hier auf der linken Seite sehen wir die verschiedenen Features, die man als Input für das Netzwerk auswählen kann, also z. B. x_1 und x_2 und wir sehen ja hier, dass das auf der rechten Seite ja ein Koordinatensystem ist, d. h. das entspricht x_1 und das entspricht x_2, und wir sehen hier in der Legende auch, dass die Farben negativen und positiven Werten entsprechen, also in dem Falle steht Orange für -1 und Blau für $+1$. Wenn ich jetzt hier über dieses Feature x_1 fahre, dann sehen wir, dass logischerweise die ganze linke Seite negativ ist, also -1, und die ganze rechte Seite positiv ist, also 1. D. h. also x_1 ist eben links negativ und rechts positiv. Genauso können wir das natürlich auch für x_2 machen, dann ist das Ganze horizontal aufgeteilt. Wir sehen, in der Mitte ist es weiß, weil da der Wert 0 ist, aber oberhalb haben wir ein positives x_2, also ist es blau gefärbt, und unterhalb ein negatives x_2, also ist der Hintergrund orange gefärbt. Dann können wir das Ganze noch machen mit ${x_1}^2$ als Input, dann ergibt sich natürlich nur eine positive Zahl, also $+1$, genauso für ${x_2}^2$. Dann kann man auch noch $x_1 x_2$ als Feature auswählen oder $\sin x_1$ oder $\sin x_2$.

In der Mitte sehen wir die Hidden Layers. Im Moment sind das zwei Schichten. Wir können die Anzahl aber auch verringern oder beliebig erhöhen. Ich stelle das jetzt mal wieder auf zwei Hidden Layers. Und dann können wir jeweils in jedem Hidden Layer auch die Anzahl der Neuronen entweder erhöhen oder verringern. Ich wähle jetzt einfach mal für die erste Schicht fünf Neuronen aus und für die zweite Schicht wähle ich mal drei Neuronen aus.

Wir sehen hier auch diese Verbindungslinien zwischen den Neuronen. Das ist genauso wie wir das kennengelernt haben eine Visualisierung für die Gewichte oder Weights zwischen den Neuronen. Auch hier entsprechen die Farben wieder negativen bzw. positiven Werten, also orange ist alles, was negativ ist, und blau sind positive Gewichte. Und auch die Stärke dieser gestrichelten Linie entspricht der Größe des Gewichts. Also z. B. hier sehen wir, das ist eine relativ dick gezeichnete Linie, da ist das Gewicht $0{,}43$, und diese dünnere Linie hier hat nur ein Gewicht von $0{,}13$.

Hier zwischen in den Hidden Layers sehen wir auch schon – also das erscheint dann immer jeweils hier rechts – das jeweilige Output aus diesem einen Neuron. Dann wird es ja verbunden bis zum letzten Hidden Layer und dann führen nochmal Verbindungen, die uns dann den Output geben, und der ist jetzt hier rechts immer im Hintergrund gekennzeichnet.

Also ich kann das jetzt auch einfach mal starten, das Netzwerk. Das kann ich nämlich machen, indem ich hier oben links auf den Play- oder Run-Button klicke. Das mache ich jetzt einfach mal. Wir sehen dann hier schon, dass diese Loss Function, diese Verlustfunktion, die hier visualisiert ist, sehr rapide sinkt, also dass der Wert sinkt, und dass sie sich sehr nah an den Wert 0 annähert, und dann, dass diese Klassifikation hier rechts eigentlich schon nahezu perfekt ist. Also wir sehen, dass alles, was an Punkten hier als blau gelabelt ist, auch

im Hintergrund, also als Vorhersage, als Prediction, blau ist, und alles, was orange ist, wird auch im Hintergrund als orange angezeigt. Ich kann jetzt nochmal die Testdaten mir dazu anzeigen lassen, wo wir sehen, dass sie sich ja ungefähr auch dort befinden, wo die Trainingsdaten auch schon sind, also hier ist auch alles korrekt klassifiziert. Was ich auch machen kann, ist, ich kann diesen Output – wir sehen ja, dass es hier auch so weiße Werte gibt, da ist sich dann das Modell unsicher, ob es jetzt positiv oder negativ ist, diesen Output kann ich einmal diskretisieren, indem ich hier unten drauf klicke. Und dann bekomme ich eben ein Entweder-oder, -1 oder 1, als Klassifikation.

Das war jetzt natürlich sehr leicht, weil ich hier ein sehr mächtiges Modell ausgewählt habe, mit vielen Neuronen. Ich kann das jetzt auch mal ein bisschen schwieriger machen und ein paar weniger Neuronen wählen für die Hidden Layers und kann z. B. mal das Verhältnis von Trainings- und Testdaten ein bisschen reduzieren. Dann sollte eigentlich unser Modell nicht ganz so gut sein. Ich drücke jetzt nochmal auf Run. Wir sehen, dass auch hier die Loss Function sinkt, vielleicht nicht mehr ganz so schnell wie vorhin, aber eigentlich ist das Ergebnis auch wieder sehr, sehr gut. D. h., was können wir noch machen, um es schlechter zu machen? Ich nehme nochmal ein paar Neuronen weg, nehme nochmal ein paar von den Trainingsdaten weg und klicke nochmal auf Run. Und dann sehen wir jetzt hier schon – ok, es sinkt nicht mehr ganz so stark. Und hier sehen wir jetzt tatsächlich eine Vorhersage, die nicht ganz so gut ist. Also wir sehen, dass auch orangene Punkte irgendwie blau klassifiziert werden und umgekehrt. Und wenn ich jetzt die Testdaten zeige und das diskretisiere, wird es vielleicht nochmal ein bisschen sichtbarer.

Take-Home Message

Jetzt könnt ihr natürlich gerne selber mal ausprobieren, wie ihr ein neuronales Netzwerk trainieren würdet, und könnt mit den Parametern herumspielen. Und wie gesagt – insbesondere nächste Woche, wenn es um die Optimierung geht, könnt ihr euch danach auch nochmal gerne diese Seite anschauen und versteht dann auch ein bisschen besser, was die Learning Rate, die Epochs und auch die Batch size bedeuten, und könnt damit auch rumspielen. Also, viel Spaß damit!

Disclaimer