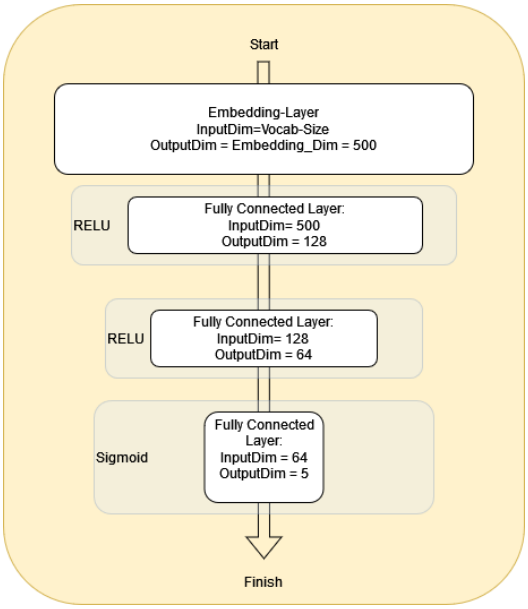
NLP – Übung 4: Domenic Bersch, 6582399

Aufgabe 1:

1.1) Skizze meiner Architektur:



Erklärung:

Bei unserem BagOfWords-Classifier handelt es sich um ein simples Feed-Forward Neural Network mit einem Embedding-Layer, und 3 Fully-Connected Layern.

Der Input vom Embeddinglayer entspricht der Größe vom gesamten Vokabular. Das liegt daran, dass im pre-processing Schritt ein Bag-Of-Words Modell aufgebaut wurde, bei dem alle vorkommenden Wörter gesammelt wurden und zu einem Vokabular sortiert wurden. Die Bag-Of-Words Repräsentation eines Satzes ist dann ein Vektor in der Größe dieses Vokabulars, mit Werten an den Stellen im Vektor, wo das Wort aus dem Satz auch im Vokabular erscheint.

Solch ein großer Vektor ist der Input für unser Netzwerk.

Das Embeddinglayer nimmt diesen Vektor entgegen und transformiert diesen in einen Embeddingvektor, welcher der Input für die darauffolgenden Layer wird.

Danach führen die Knoten in ein Fully-Connected Layer, welches durch eine ReLU gezogen wird, und schließlich zwei weitere Fully Connected Layer. Das letzte hat als Output lediglich 5 Knoten, da wir auch nur 5 Labels zu vergeben haben. Statt ReLU nutzen wir hier Sigmoid als Aktivierungsfunktion um die Wahrscheinlichkeitsverteilung auf den jeweiligen labels besser deuten zu können.

1.2: Laden Sie Trainingsdaten und implementieren Sie notwendige Methoden für das Pre-Processing

- Trainingsdaten wurden implementiert
- Ein Dictionary mit allen vorkommenden Wörtern wurde erstellt durch Zusammensetzung aller Daten (Für das Training wurden aber nur die Trainignsdaten verwendet)
- Klasse „Dataset“ erstellt einen Bag-of-Words Vektor
- Preprocessing-Schritte:
    - Entfernen von stopwords
    - Lowercase
    - Spezielle Zeichen entfernen
    - Nummern entfernen

1.3: Trainingsskript schreiben

- Habe ich gemacht aber leider komme ich auf maximal 29% accuracy ☹
- Ich habe echt alles probiert, viel gegoogelt, bessere Optimierungen gewählt aber wurde trotzdem nicht besser…

1.4.: Trainieren Sie das Modell auf dem Trainingsdatensatz

- Habe ich gemacht

1.5: Ergebnisse:

Ich habe verschiedene Dimensionen ausprobiert, verschiedene Vokabulargrößen (durch cropping seltener auftauchender Wörter), verschiedene Ansätze fürs Pre-Processing für die Sätze. Ich habe auch durch viel Googlen nach weiteren Ansätzen geschaut um meine Accuracy zu verbessern, jedoch bin ich nicht über maximal 29% hinweggekommen. Ich verstehe auch wirklich nicht warum, ich habe echt alles probiert aber es wird einfach nicht besser…

Man muss zwar sagen, dass die Accuracy aus den Folien (um die 40-45%) nicht viel besser ist, aber 29% ist nur 9% besser als Random (bei 5 Klassen = 20%).

Accuracy: `0.2864253393665158`