Parsing mit neuronalen Netzen: Training und Anwendung

Parsertraining

Implementieren Sie analog zur Aufgabe LSTM-Tagger eine Klasse **Vocab** in einer Datei `vocab.py`, deren init-Funktion die Trainingsdaten als Argument erhält und die Konstituentenlabels auf fortlaufende Zahlen beginnend mit 1 abbildet. Der Index 0 ist für die Spezialkategorie `No_Constituent` reserviert. Unbekannte Kategorien (bspw. aus den Development-Daten) werden auf -1 abgebildet. Erstellen Sie Dictionaries für die **Abbildung von Labels auf Zahlen** und umgekehrt. Erzeugen Sie ferner ein Dictionary für die **Abbildung von Buchstaben auf Zahlen**, wobei die Nummern 0 und 1 für das Paddingsymbol und das UnkChar-Symbol reserviert sind. Buchstaben, welche nur einmal aufgetreten sind, erhalten keinen eigenen Index sondern dienen zum Training des UnkChar-Symbols. Schreiben Sie eine Methode **char_id_tensor(words)**, welche eine Wortliste erhält und einen gepaddeten Buchstaben-Tensor und eine Liste der Wortlängen liefert. Schreiben Sie außerdem zwei Methoden **num_char_types()** und **num_label_types()**, welche die Zahl der Zeichen bzw. die Zahl der syntaktischen Labels liefern. Schreiben Sie dann noch eine Methode **store_parameters(filename)**, welche die Abbildungstabellen für Buchstaben und Labels in einer Datei speichert. Erweitern Sie die init-Funktion der Klasse so, dass sie prüft, ob ein String (=Dateiname) als Argument übergeben wurde, und in diesem Fall die Tabellen aus der Datei einliest und damit das Objekt initialisiert.

Erstellen Sie eine weitere Datei `train-parser.py`, welche das Training implementiert und Ihren Code aus der letzten Aufgabe importiert. Schreiben Sie die Hauptfunktion **train**. Diese liest mit Hilfe der Funktionen aus der letzten Aufgabe Trainingsdaten und Developmentdaten aus zwei Dateien ein, deren Namen als Kommandozeilenargumente übergeben werden. Sie speichert die Daten jeweils als eine Liste von Paaren, wobei jedes Paar aus einer Wortliste und einer Konstituentenliste besteht. Dann werden ein Vocab-Objekt und ein neuronales Modell erzeugt. Sie trainieren für **num_epochs** Epochen (bspw. `num_epochs=50`) auf den Trainingsdaten, die Sie zu Beginn jeder Epoche mit `random.shuffle` umordnen. `num_epochs` ist ein optionaler Kommandozeilenparameter mit Defaultwert. Verwenden Sie ggf. weitere Funktionen, um Ihren Code besser zu strukturieren.

Das neuronale Netzwerk aus der letzten Aufgabe gibt Ihnen einen 2-dimensionalen Score-Tensor zurück, der für jeden Span eine Liste von Scores liefert. An Position 0 des Vektors befinden sich die Bewertungen für den Span (0,1). Dann folgen (0,2),...,(0,n),(1,2),...,(1,n),...,(n-1,n), wobei n die Zahl der Wörter ist. Um das Loss für diesen Score-Tensor zu berechnen, müssen Sie einen Vektor **label_ids** erstellen, der für jeden Span die korrekte Label-ID enthält.

Gehen Sie dabei so vor, dass Sie zunächst einen 1-dimensionalen Tensor (Vektor) der richtigen Länge mit `NO_CONST_LABEL_ID` füllen und in **label_ids** speichern. Dann spalten Sie `label_ids` mit dem split-Befehl in eine Liste von Tensoren `label_ids_view` der Längen n, n-1, .., 1 auf. Nun iterieren Sie über alle Goldstandard-Konstituenten (l, s, e) des aktuellen Parsebaumes und Setzen das Label des Spans (s, e) mit dem Befehl `label_ids_view[s][e-s-1] = l` auf die Label-ID l. Da der split-Befehl nur Views erzeugt, wird `label_ids` dadurch ebenfalls modifiziert.

Als Trainingskriterium verwenden Sie das `CrossEntropyLoss` von PyTorch. Wenden Sie **Gradient Clipping** mit einer Gradient-Norm von 1 an, um sehr große Gradienten zu vermeiden. Nach jeder Epoche berechnen Sie die Zahl der falsch gelabelten Spans in den Development-Daten. Wenn die Zahl der Fehler die bisher kleinste war, speichern Sie das Netzwerk mit der Methode `torch.save`. Außerdem müssen Sie die Methode `vocab.store_parameters` aufrufen. Speichern Sie die Tabellen und das Netzwerk in separaten Dateien mit gleichem Basisnamen und unterschiedlichen Dateiendungen (.io bzw. .pth).

Bei einer guten Implementierung kann die Zahl der falsch gelabelten Konstituenten in den Development-Daten im Laufe des Trainings unter 6000 sinken. Planen Sie genug Zeit für das Training ein, da es auch auf der GPU mehrere Stunden dauern kann.

Aufruf: python train-parser.py train-parses dev-parses parfile

Parsing

Nun implementieren Sie den eigentlichen Parser. Ihr Programm liest das gespeicherte neuronale Netzwerk ein und analysiert damit Sätze aus einer Datei. Jede Zeile der Datei enthält einen bereits tokenisierten Satz (keinen Parsebaum!).

Erstellen Sie eine Datei **parser-analyze.py** für das Parsen neuer Sätze, in der Sie die Klasse `Model` aus der letzten Aufgabe und die Klasse `Vocab` aus dieser Aufgabe importieren.

Implementieren Sie eine Funktion **parse**, welche eine Wortliste als Eingabe erhält und den besten Parsebaum in Klammernotation zurückgibt. `parse` erzeugt mit der Methode `vocab.char_id_tensor(words)` den Eingabetensor und Längenvektor. Damit wird das neuronale Netz aufgerufen, um die Logits für alle Spans und Labels zu berechnen. Dann wird für jeden Span das wahrscheinlichste Label und sein Logit-Wert berechnet. Der Viterbi-Algorithmus berechnet den besten Parsebaum. Dieser maximiert die Summe der Label-Scores. Der Viterbi-Algorithmus liefert eine Liste von Konstituenten-Tripeln. Diese wird mit der entsprechenden Methode aus der letzten Aufgabe in einen Parsebaum in Klammernotation umgewandelt und zurückgegeben. Bei der Implementierung des Viterbi-Algorithmus orientieren Sie sich am Pseudocode aus den Vorlesungsfolien. Versuchen Sie, die beste Zerlegung einer Konstituente in Tochterkonstituenten mittels einer einzigen PyTorch-Operation zu berechnen. Dazu speichern Sie die Viterbi-Scores in einem zweidimensionalen Tensor `vitscore[s,e]`.

Schreiben Sie nun das Hauptprogramm. Dieses liest zuerst das Vokabular mit den Abbildungstabellen ein, dann das trainierte Netzwerk, und schließlich die Sätze. Dann werden die Sätze geparst und die Parsebäume in Klammernotation auf den Bildschirm ausgegeben. Erstellen Sie zum Testen eine Textdatei mit einem tokenisierten Satz pro Zeile und parsen Sie die Sätze mit ihrem Parser.

Aufruf: python parser-analyze.py parfile sentences

Vorüberlegungen

- Wie sollte die Funktion *parse* vorgehen?

Schicken Sie mir bitte den kompletten ausführbaren Code mit den trainierten Parameterdateien, damit ich das Programm direkt testen kann. Schicken Sie außerdem eine Datei *num-errors.txt* mit der Anzahl der Fehler nach jeder Trainingsepoche.