**Kurze Erläuterung was Hugging Face ist**

Hugging Face ist ein Unternehmen, das sich auf künstliche Intelligenz (KI) im Bereich des Natural Language Processing (NLP) spezialisiert hat. Sie sind bekannt für ihre Open-Source-Plattform, die als "Transformers" bekannt ist. Diese Plattform bietet Zugriff auf eine Vielzahl von vortrainierten Modellen für verschiedene NLP-Aufgaben wie Textklassifikation, maschinelles Übersetzen, Generierung von Text und mehr.

Hugging Face ermöglicht es Entwicklern, auf diese Modelle zuzugreifen und sie in ihren eigenen Anwendungen zu nutzen. Die Plattform bietet auch Tools für das Feintuning von Modellen, um sie an spezifische Anforderungen anzupassen. Darüber hinaus fördert Hugging Face die gemeinsame Nutzung von Modellen und Ressourcen in der Entwicklergemeinschaft, was zu einer breiten Verfügbarkeit von fortschrittlichen NLP-Modellen führt.

**Grundlagen bei BERT**

**CLS] Token (Anfangs-Token):**

* Das [CLS] Token (Classification-Token) wird üblicherweise am Anfang einer Sequenz platziert. Es wird für Aufgaben wie Klassifikation verwendet, um dem Modell zu signalisieren, dass es eine bestimmte Aufgabe lösen soll.
* In der Kontext-Aufgabenstellung von BERT wird das [CLS] Token verwendet, um den gesamten Kontext zu repräsentieren und in Verbindung mit den Endpositionen von [SEP] die Antwort zu bestimmen.

**[SEP] Token (Trennungs-Token):**

* Das [SEP] Token wird verwendet, um Sequenzen voneinander zu trennen. Es wird am Ende jeder Sequenz, einschließlich der letzten Sequenz in einem Text, hinzugefügt.
* In BERT wird [SEP] verwendet, um den Anfang und das Ende von Textsegmenten anzuzeigen. Beim Fine-Tuning auf Frage-Antwort-Aufgaben wird [SEP] nach der Frage und nach der Antwort platziert.

**Tokenizer:** Ein Tokenizer wandelt Text in tokenisierte Form um, was für die Verarbeitung durch Modelle notwendig ist. Hugging Face bietet effiziente Tokenizer für viele Sprachen und Modelle.

**Konkretes Beispiel**

* **[CLS]:** Wird durch die Zahl 3 repräsentiert.
* **[SEP]:** Wird durch die Zahl 4 repräsentiert.

**Frage: Wie heißt du?**

[CLS] Wie heißt du? [SEP] Ich heiße Tim. [SEP]

* input\_ids: [3, 1316, 2018, 4547, 26972, 4, 1671, 22429, 5095, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
* token\_type\_ids: [0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
* attention\_mask: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

**Frage: Wie alt bist du?**

[CLS] Wie alt bist du? [SEP] Ich bin 20 Jahre alt. [SEP]

* input\_ids: [3, 1316, 3778, 23267, 4547, 26972, 4, 1671, 4058, 148, 573, 3778, 26914, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0]
* token\_type\_ids: [0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]
* attention\_mask: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]

**Frage: Was sind die Hauptbestandteile von Wasser?**

1. [CLS] Was sind die Hauptbestandteile von Wasser? [SEP] Die Hauptbestandteile von Wasser sind Wasserstoff und Sauerstoff. [SEP]
   * input\_ids: [3, 2019, 287, 30, 910, 4626, 4070, 88, 1400, 26972, 4, 125, 910, 4626, 4070, 88, 1400, 287, 21604, 42, 16334, 26914, 4]
   * token\_type\_ids: [0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

* attention\_mask: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

**token\_type\_ids**: Dieser Tensor zeigt an, welcher Teil der Eingabe zu welcher Sequenz gehört. In diesem Fall gibt es zwei Sequenzen: 0 für die Frage und 1 für die Antwort. Die restlichen 0en sind Paddings.

**attention\_mask**: Der Wert 1 bedeutet, dass das Modell diese Positionen in der Sequenz beachten soll. Der Wert 0 bedeutet, dass das Modell diese Positionen ignorieren soll.

**Begriffserklärungen**

**Pipeline:** Die Pipeline-API von Hugging Face ermöglicht es, Modelle für spezifische Aufgaben mit wenigen Zeilen Code zu verwenden, z.B. Textklassifikation, Named Entity Recognition (NER) usw.

**Pretraining:** Pretraining bezieht sich auf das Training eines Modells auf einem großen Korpus von Textdaten, bevor es für spezifische Aufgaben feinabgestimmt wird.

**Checkpoint:** Ein Modell-Checkpoint ist eine gespeicherte Version eines Modells zu einem bestimmten Zeitpunkt während des Trainings, die später geladen und verwendet werden kann.

**Embedding:** Embeddings sind darstellende Vektoren, die Wörter oder Token in einem NLP-Modell repräsentieren.

**Fine-Tuning:** Feinabstimmung ist der Prozess, bei dem ein vortrainiertes Modell auf spezifische Aufgaben oder Datensätze angepasst wird, um die Leistung zu verbessern.

**Wichtige Evaluationsmetriken und deren Bedeutung**

**HasAns (Has Answer):**

* **exact:** Der Prozentsatz der Fragen mit einer exakt korrekten Antwort.
* **f1:** Der F1-Score, der ein Gleichgewicht zwischen Precision und Recall darstellt.

**NoAns (No Answer):**

* **exact:** Der Prozentsatz der Fragen, für die das Modell korrekt festgestellt hat, dass keine Antwort vorhanden ist.
* **f1:** Der F1-Score für die Fragen ohne Antwort.

**Best Exact und Best F1:**

* Die besten Werte sowohl für exakte Übereinstimmung als auch für den F1-Score während der Evaluation.

**Epoch:**

* Die Anzahl der Durchläufe des Trainingsdatensatzes.

**Train Loss:**

* Der Verlust (Loss) auf dem Trainingsdatensatz.

**Train Runtime:**

* Die Gesamtzeit, die für das Training benötigt wurde.

**Train Samples und Train Samples Per Second:**

* Die Anzahl der Trainingsbeispiele und die Anzahl der verarbeiteten Trainingsbeispiele pro Sekunde.

**Train Steps Per Second:**

* Die Anzahl der Trainingschritte (Berechnungen) pro Sekunde.