AI Bootcamp

Introduction to Hugging Face Transformers

Module 21 Day 2

Class Objectives

By the end of class, you will be able to:

1. Describe the importance of transformers and what they do.
2. Become familiar with the different pre-trained transformer models.
3. Use transformers for NLP tasks such as text translation, generation, summarization, and question answering.
4. Apply a pre-trained transformer model for a specific task.

Instructor **Demonstration**

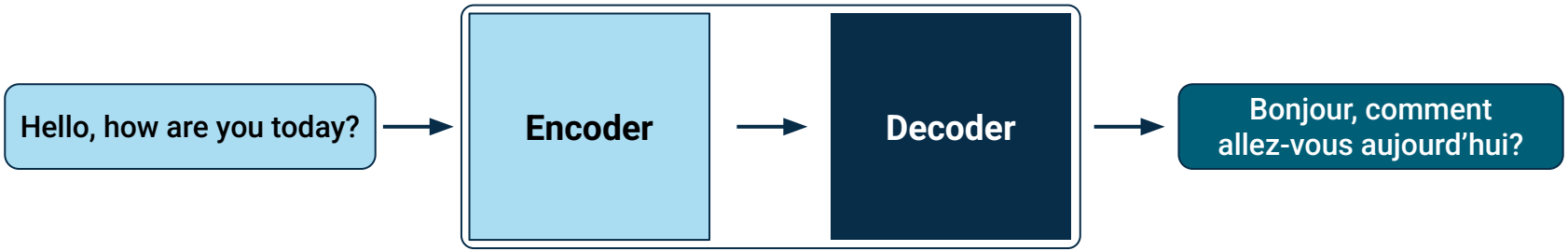
Introduction to Transformers and Pre-trained Models

Transformer Architecture

Encoder-decoders in translation LLMs

At a high level, the transformer architecture consists of an encoder-decoder.

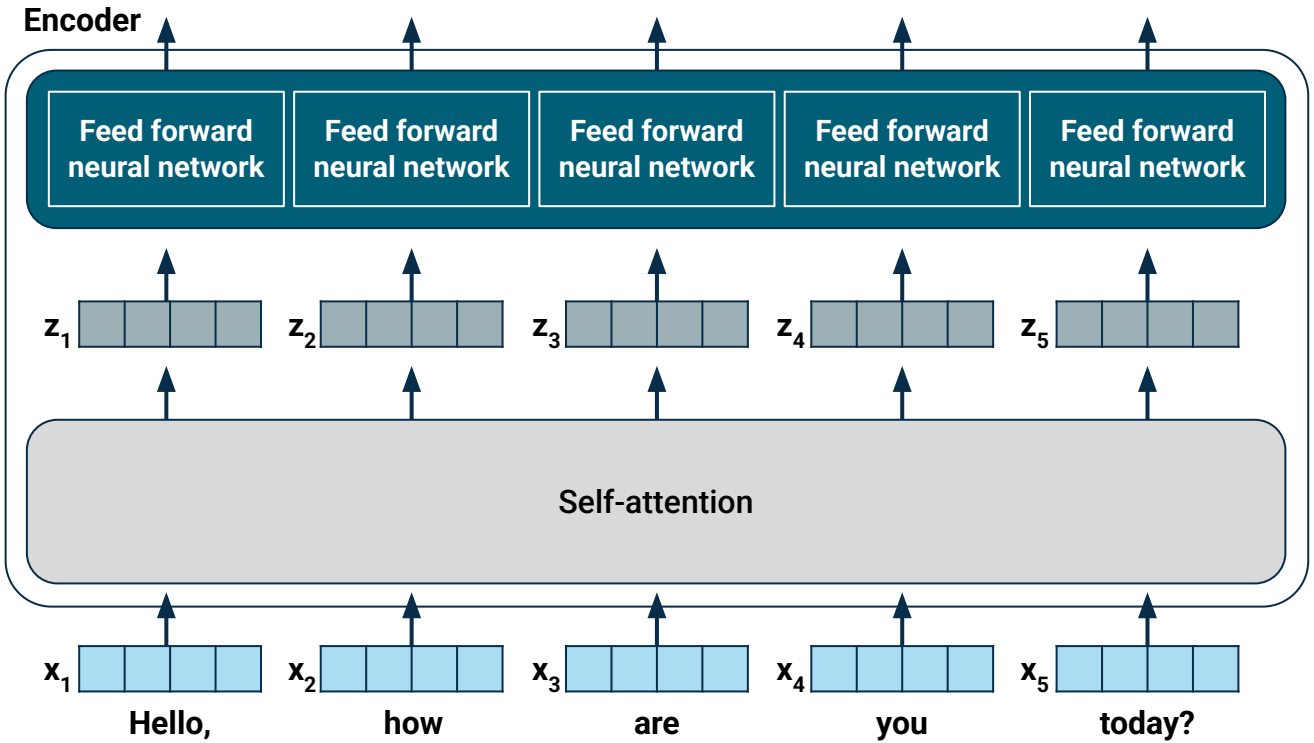
- The encoder will process an input sequence: a sentence in English.
- The encoded sentence will then pass through the decoder, which produces the output; the same sequence that has been translated into French.



Transformer Architecture

Encoder architecture

Each vector is fed into the self-attention layer and then fed into the feed forward layers of the encoder.



Self-attention

Self-attention allows the model to compare each word in the sequence to other words in the sequence to determine any useful information for encoding the word.

The flower that stood in the sun wilted because it was too bright.

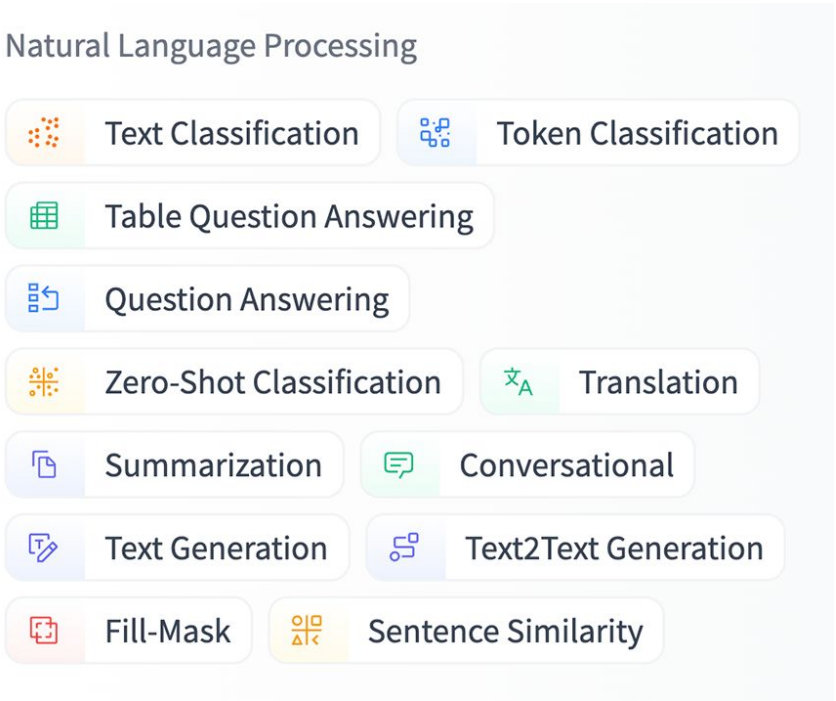
In English, the word “it” refers to the sun and not the flowers.

When the model is encoding the word “it,” the self-attention mechanism allows the model to encode useful information, such as how strongly the word “it” is related to “sun” and “flower.”

This is as close to understanding what “it” means in the sentence as a computational model will get.

Hugging Face Transformers

Possible NLP applications



Source

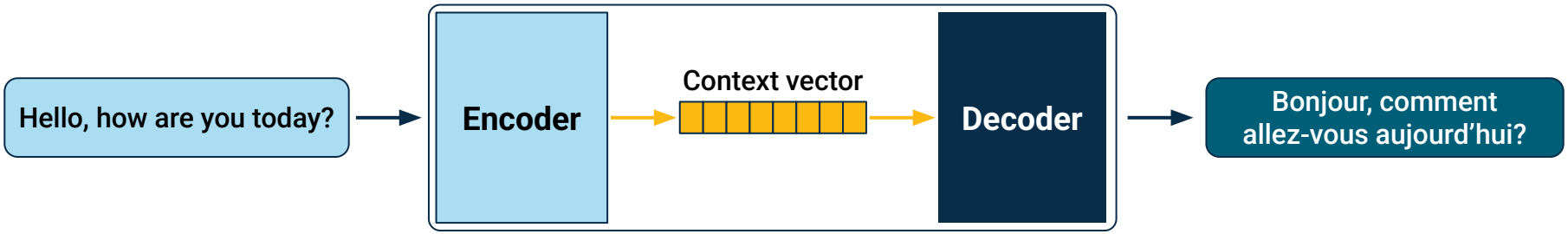
Instructor **Demonstration**

Language Translation

Language translation

Sequence-to-sequence LLMs

Sequence-to-sequence (Seq2Seq) models consist of an encoder and decoder.



The transformer model used in language translation tasks is the sequence-to-sequence (Seq2Seq) language model.

1. The encoder takes the source language sentence as input and encodes it into a fixed-length vector representation, often called a context vector.

2. Then, the decoder takes the context vector as input and generates the target language sentence word by word.

Activity:

News Headline Translation

In this activity, you will use a Hugging Face transformer pre-trained model to translate text into one or more languages.

Suggested Time:
15 Minutes

Time’s up!

Let’s review

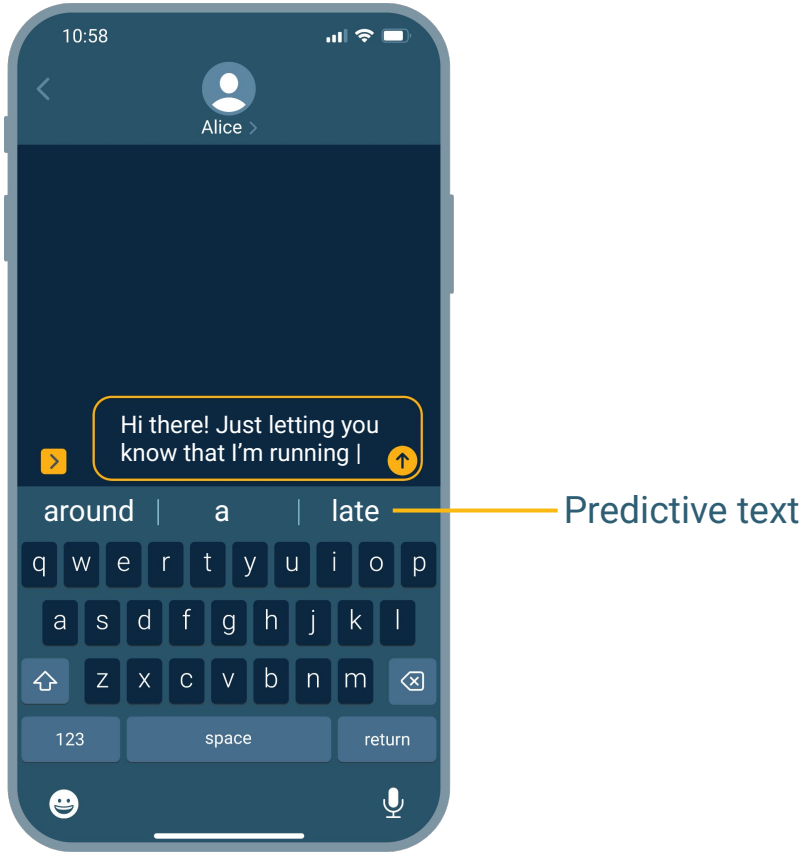
Questions?

Instructor **Demonstration**

Text Generation

Text generation models

Text generation is one of the more common applications of transformer models. An example of text generation is using predictive text in text or email messages.



Examples of Prompts for Text Generation

01

A user might provide an incomplete sentence and ask the model to finish it, or ask the model to write a story, giving it a setting, characters, or a few sentences as a starting point.

02

A user could even describe a programming problem and ask the model to generate code that solves the problem in a programming language of the user's choice.

03

To generate the output text from the prompt, the transformer uses a decoder and its knowledge of language patterns and context to predict (i.e., generate) the next word in a sequence, given the previous ones.

Generative models can sometimes produce inaccurate, offensive, or biased responses. **Use them with caution**.

Activity:

Text Generation Playground

In this activity, you will test three different pre-trained Hugging Face transformer models to generate text and compare the outputs from the models.

Suggested Time:
15 Minutes

Time’s up!

Let’s review

Questions?

Break

15 mins

Instructor **Demonstration**

Question and Answering

Question and answering models

Encoder-only transformer architecture

Welcome to our store!
Is there anything I can help you with today?

BERT-based models

Bidirectional encoder representations from transformers

- The BERT architecture allows for bidirectional information flow.
- The model uses past and future tokens to understand meaning.

Is it true that male bees get kicked out of the hive during winter?

Yes, male bees, or drones, are forcibly removed from the hive when resources are scarce to protect the population of the worker bees and the queen, who are all female. This includes removing them from the hive during winter.

What is the biggest whale on the planet?

The blue whale (*Balaenoptera musculus*) is the largest whale on the planet. They can grow up to 100 feet or more.

Do bees choose their queen by giving her purple honey?

No, bees do not choose a queen based on honey color. They select a queen by feeding a few larvae with a nutritious substance known as royal jelly. All larvae are initially fed royal jelly and are then weaned off until only a few receive the special substance as food. The nutrients cause the favored few to develop reproductive organs and grow larger in size, which is characteristic of a queen bee. But royal jelly is generally white in color not purple.

Activity:

Q & A Testing

In this activity, you will work with a partner to create a question and answering system using a pre-trained Hugging Face transformer model.

Suggested Time:
15 Minutes

Time’s up!

Let’s review

Questions?

Instructor **Demonstration**

Text Summarization

Activity:

Text Summarization

In this activity, you will use a pre-trained Hugging Face transformer model to summarize text of your choosing.

Suggested Time:
15 Minutes

Time’s up!

Let’s review

Questions?

Question 1:

Your friend wants to create a generative text application for their business that searches the company database based on customer prompts. They ask you for advice on which Hugging Face pre-trained model to use. Which of the following Hugging Face models would you recommend?

1. facebook/bart-large-cnn
2. openai-gpt
3. EleutherAI/gpt-neo-2.7B
4. EleutherAI/gpt-neo-125M

Question 1: Answers

Your friend wants to create a generative text application for their business that searches the company database based on customer prompts. They ask you for advice on which Hugging Face pre-trained model to use. Which of the following Hugging Face models would you recommend?

1	facebook/bart-large-cnn	This is incorrect. This model is ideal for text summarization.
2	openai-gpt	This is incorrect. This model is ideal for question answering, semantic similarity, and text classification.
3	EleutherAI/gpt-neo-2.7B	**Correct.** This model has been trained on 3 billion parameters and is ideal for generating text from prompts. The text generated from this model might contain more accurate text than smaller models.
4	EleutherAI/gpt-neo-125M	Incorrect. Although this model has been trained on 125 million parameters and is suitable for generating text from prompts, text generated from this model might generate more nonsensical or hallucinatory text than a larger model.

Question 2:

You want to build an application that summarizes webpages. What is the best model to use for this process?

1. facebook/bart-large-cnn
2. openai-gpt
3. EleutherAI/gpt-neo-2.7B
4. t5-base

Question 2: Answers

You want to build an application that summarizes webpages. What is the best model to use for this process?

1	facebook/bart-large-cnn	**Correct.** This model is ideal for text summarization.
2	openai-gpt	Incorrect. This model is ideal for question answering, semantic similarity, and text classification.
3	EleutherAI/gpt-neo-2.7B	Incorrect. This model has been trained on 3 billion parameters and is ideal for generating text from prompts. The text generated from this model might contain more accurate text than smaller models.
4	t5-base	Incorrect. This model is ideal for translation between these languages: English, French, Romanian, and German.

Question 3:

Consider the following line of code from a Seq2Seq translation model:

```
translation_model = TFAutoModelForSeq2SeqLM.from_pretrained("t5-base")
```

There are two pre-trained models that are used to create the translation model. Fill in the blanks to describe what role each model plays: The `TFAutoModelForSeq2SeqLM` model _____ while the `t5-base` model _____.

1. performs the tokenization; is imported for its stored weights
2. performs the tokenization; houses the transformer architecture
3. houses the transformer architecture; gives the model a performance advantage since it is trained on millions of parameters
4. houses the transformer model architecture; is imported for its stored weights

Question 3: Answers

Consider the following line of code from a Seq2Seq translation model:

```
translation_model = TFAutoModelForSeq2SeqLM.from_pretrained("t5-base")
```

There are two pre-trained models that are used to create the translation model. Fill in the blanks to describe what role each model plays: The `TFAutoModelForSeq2SeqLM` model _____ while the `t5-base` model _____.

1. performs the tokenization; is imported for its stored weights

   Incorrect. We use a model like `Autotokenizer` to take care of the tokenization.

2. performs the tokenization; houses the transformer architecture

   Incorrect. We use a model like `Autotokenizer` to take care of the tokenization.

3. houses the transformer architecture; gives the model a performance advantage since it is trained on millions of parameters

   Incorrect. We use a model like `Autotokenizer` to take care of the tokenization.

4. houses the transformer model architecture; is imported for its stored weights

   Correct. The `TFAutoModelForSeq2SeqLM` houses the transformer architecture and the `t5-base` model is imported for its stored weights.

Question 4:

If your model has trained on a large enough dataset, you can trust that it will always generate accurate, appropriate outputs.

1. True
2. False

Question 4: Answers

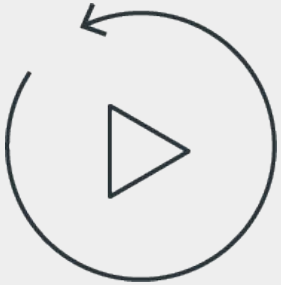
If your model has trained on a large enough dataset, you can trust that it will always generate accurate, appropriate outputs.

1 **True**

Incorrect. Although large datasets tend to perform better and generate more accurate outputs, they are still prone to inaccuracy and bias.

2 **False**

Correct, well done. Although large datasets tend to perform better and generate more accurate outputs, they are still prone to inaccuracy and bias.



Let’s **recap**

Recap

After today's lesson, you are able to:

1. Describe the importance of transformers and what they do.
2. Become familiar with the different pre-trained transformer models.
3. Use transformers for NLP tasks such as text translation, generation, summarization, and question answering.
4. Apply a pre-trained transformer model for a specific task.

Next

In the next lesson, we will begin to work with Gradio, which allows us to build graphic user interfaces (GUIs) that users can interact with so that they can use your applications.

Questions?

The End