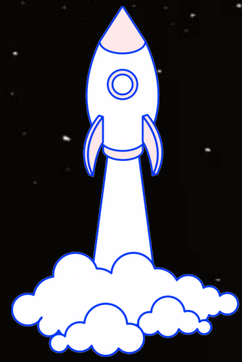
Language Models

UNSA

Tunisian Space Association

الجمعية التونسية للفضاء

NLP Zero to Hero



Natural language processing

Natural language processing (NLP) refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.

1. Tokenization


```
import tensorflow as tf
from tensorflow import keras
from tensorflow.keras.preprocessing.text import Tokenizer
```

```
sentences = [
    'I love my dog',
    'I love my cat'
]
```

```
tokenizer = Tokenizer(num_words = 100)
tokenizer.fit_on_texts(sentences)
word_index = tokenizer.word_index
print(word_index)
```

2.Sequencing - Turning sentences into data

```
sequences = tokenizer.texts_to_sequences(sentences)
```

```
[[4, 2, 1, 3], [4, 2, 1, 6], [5, 2, 1, 3], [7,
5, 8, 1, 3, 9, 10]]
```

```
test_data = [
    'i really love my dog',
    'my dog loves my manatee'
]

test_seq = tokenizer.texts_to_sequences(test_data)
print(test_seq)
```

```
[[4, 2, 1, 3], [1, 3, 1]]

{'think': 8, 'amazing': 10, 'my': 1, 'love': 2, 'dog': 3, 'is': 9,
'you': 5, 'do': 7, 'cat': 6, 'i': 4}
```

```
tokenizer = Tokenizer(num_words = 100, oov_token="<OOV>")
```

```
{'think': 9, 'amazing': 11, 'dog': 4, 'do': 8, 'i': 5, 'cat': 7,
 'you': 6, 'love': 3, '<OOV>': 1, 'my': 2, 'is': 10}
```

if we want to train a neural network, how to handle sequences of different lengths?

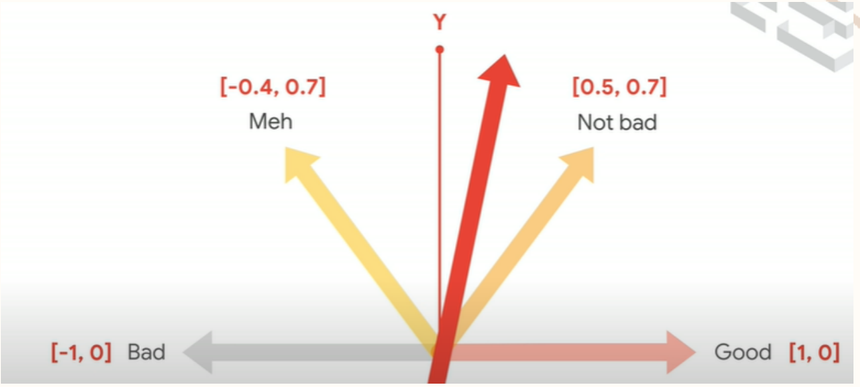
```
padded = pad_sequences(sequences, padding='post',
                                  truncating='post', maxlen=5)
```

3.training a model

```
model = tf.keras.Sequential([
    tf.keras.layers.Embedding(vocab_size, embedding_dim, input_length=max_length),
    tf.keras.layers.GlobalAveragePooling1D(),
    tf.keras.layers.Dense(24, activation='relu'),
    tf.keras.layers.Dense(1, activation='sigmoid')
])
model.compile(loss='binary_crossentropy',optimizer='adam',metrics=['accuracy'])
```

```
num_epochs = 30

history = model.fit(training_padded, training_labels, epochs=num_epochs,
                    validation_data=(testing_padded, testing_labels), verbose=2)
```



Recurrent Neural Networks

A recurrent neural network (RNN) is a class of artificial neural networks where connections between nodes can create a cycle.
RNNs can use their internal state (memory) to process variable-length sequences of inputs.

RNN vs CNN

	Convolutional neural network (CNN)	Recurrent neural network (RNN)
ARCHITECTURE	Feed-forward neural networks using filters and pooling	Recurring network that feeds the results back into the network
INPUT/OUTPUT	The size of the input and the resulting output are fixed (i.e., receives images of fixed size and outputs them to the appropriate category along with the confidence level of its prediction)	The size of the input and the resulting output may vary (i.e., receives different text and output translations—the resulting sentences can have more or fewer words)
IDEAL USAGE SCENARIO	Spatial data (such as images)	Temporal/sequential data (such as text or video)
USE CASES	Image recognition and classification, face detection, medical analysis, drug discovery and image analysis	Text translation, natural language processing, language translation, entity extraction, conversational intelligence, sentiment analysis, speech analysis

RNN types

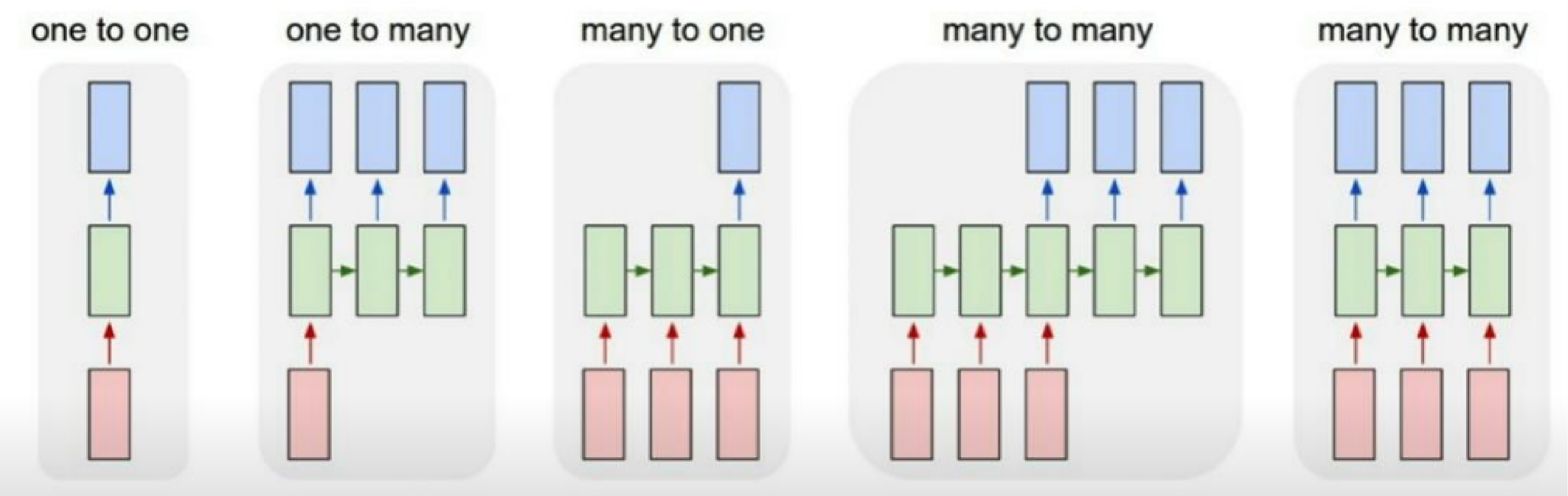
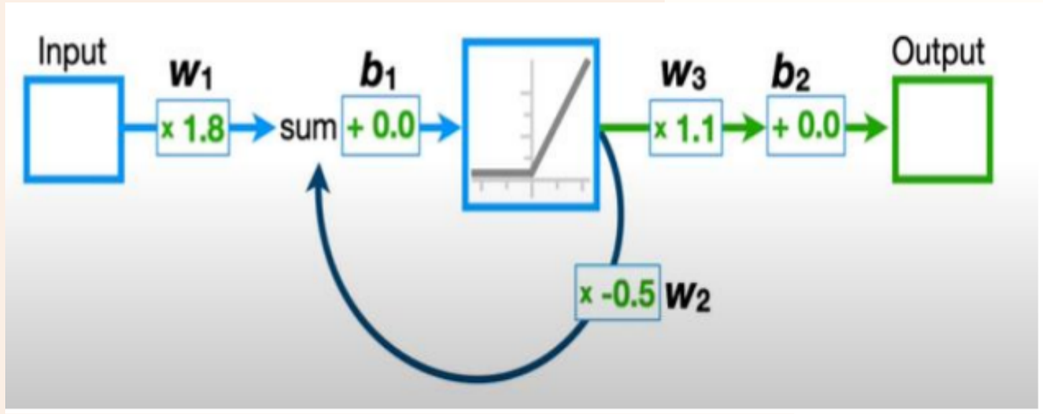


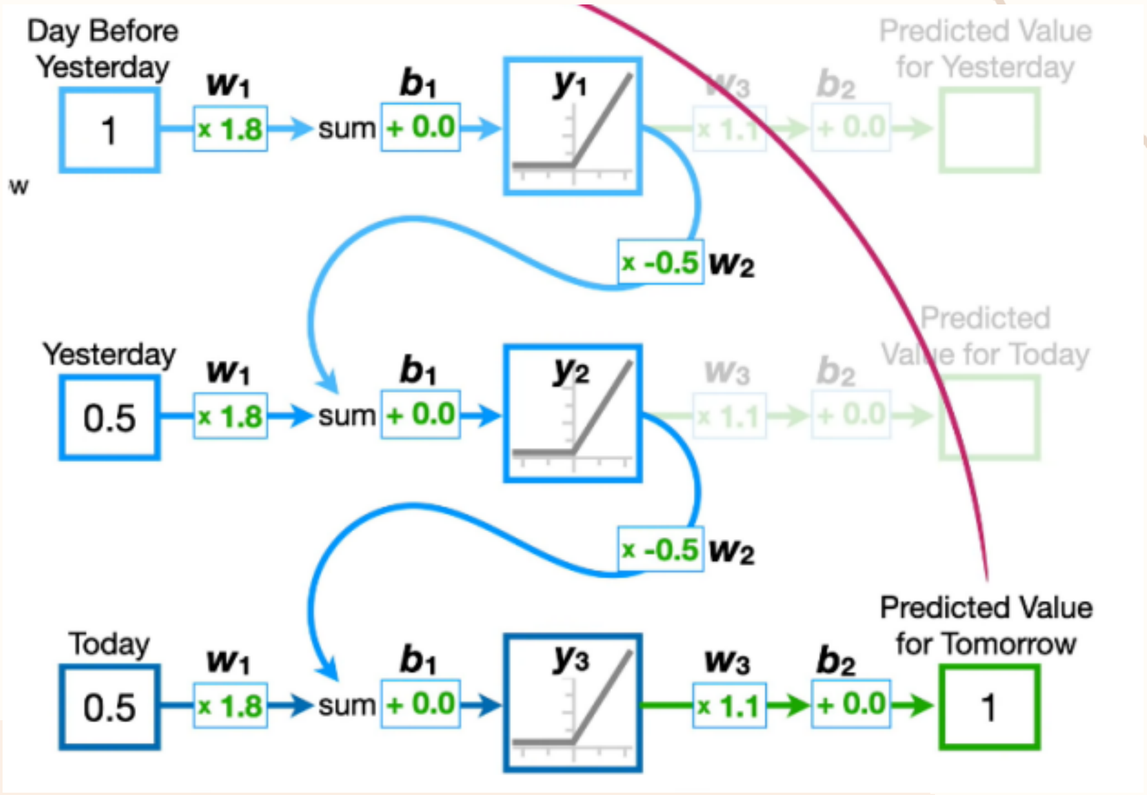
Image to sequence of words

sentiment of text/video

translation

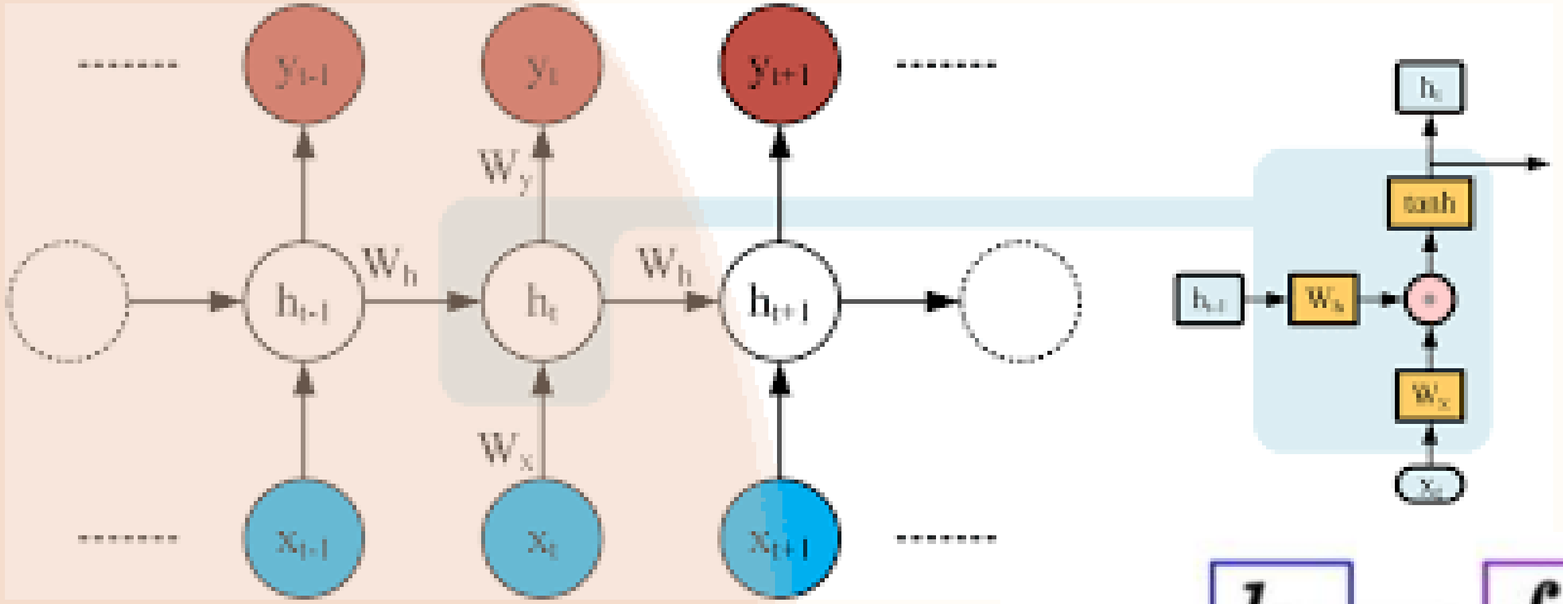
Feedback loop





architecture

Markov Assumption: At each time step, the input to the hidden layer is simply the last hidden state and current input.



$$h_t = f_W(h_{t-1}, x_t)$$

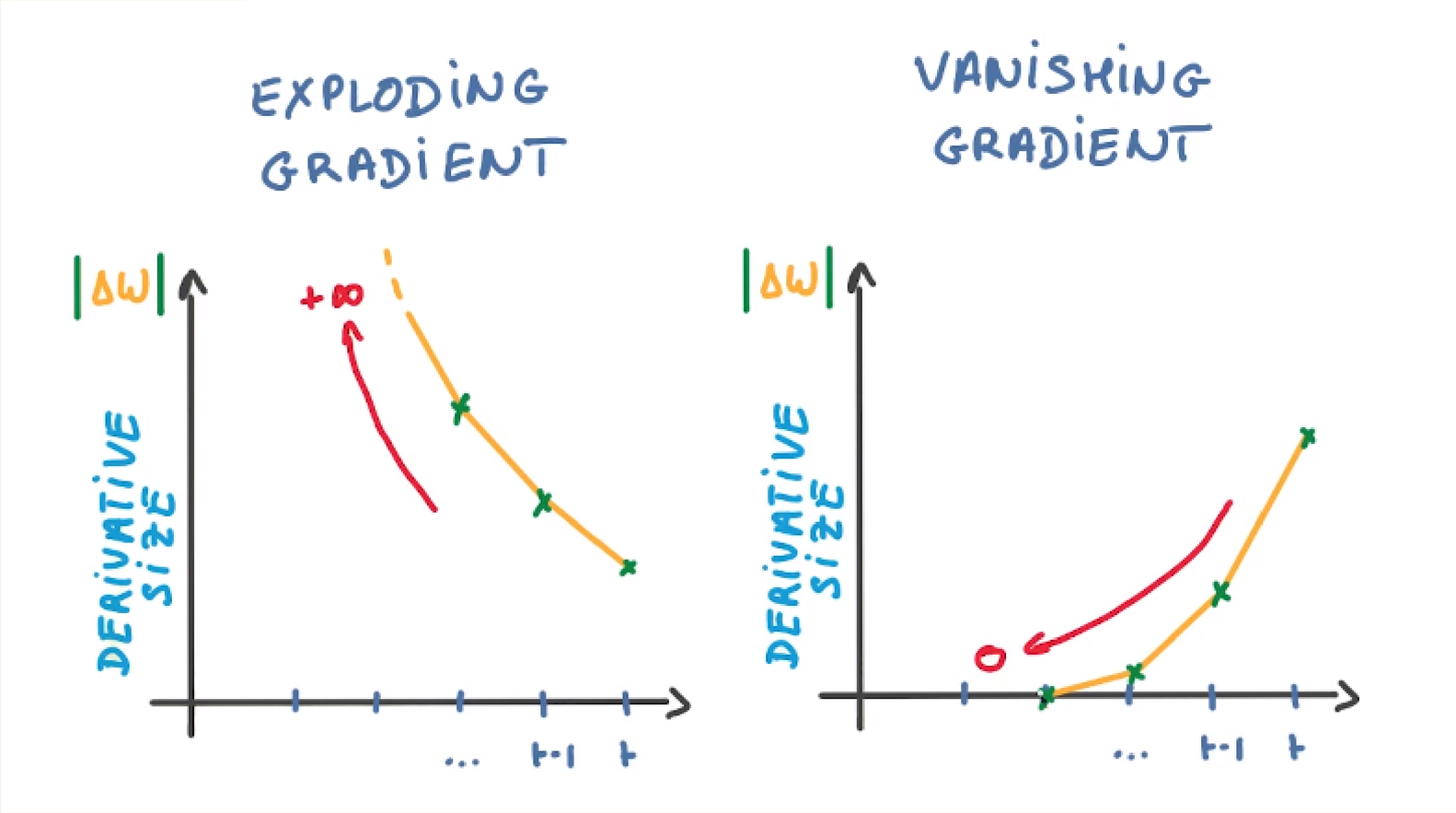
new state

old state

input vector at some time step

some function with parameters W

limitations

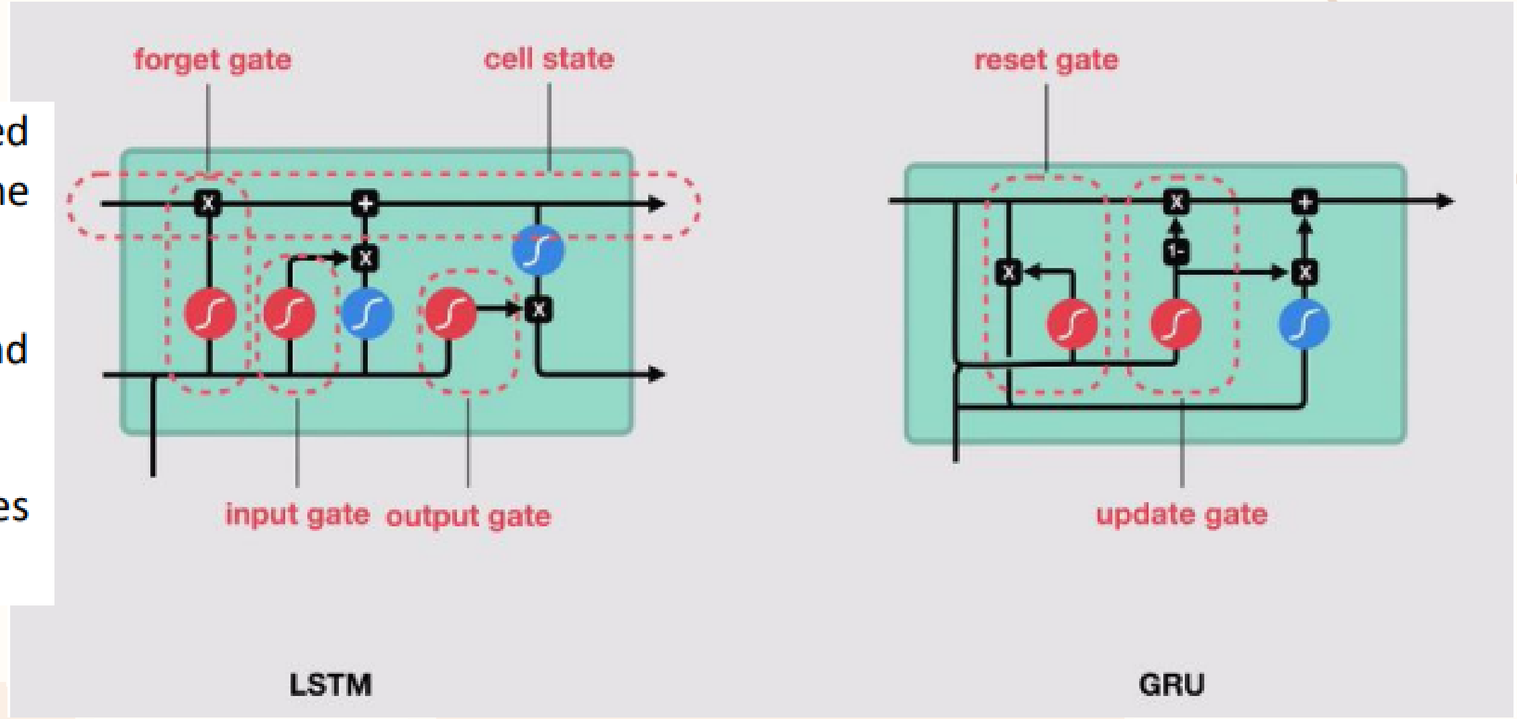


LTSM & GRU

Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) were designed to address the vanishing gradient problem in RNNs.

LSTMs use three gates called input, output, and forget gate.

GRU have a reset and update gate. These gates determine which information to retain.



ssakerna el NLP

transformers?