Keluarga Mahasiswa Pusat
عائلة طالب المركزي

Natural Language Processing

Hands On with TensorFlow

Muhammad Dzaki Hanifa | Zildan Alfatih Agustian | Maulana Al Iqbal Widodo | Alif Rahmat Yudha

Articial Intelligence class A | Mr. Dimas Aryo Anggoro

What is a NLP Problem?

NLP (Natural Language Processing) adalah cabang dari kecerdasan buatan yang fokus pada interaksi antara komputer dan bahasa manusia. Tujuannya adalah memungkinkan komputer untuk memahami, memproses, dan menghasilkan bahasa manusia dengan cara yang bermanfaat.

How Actually NLP Works?

"Is this Tweet for a disaster or not?"

Diaster

Not Diaster

Actual output

```
[[0.22, 0.98, 0.02…],
[0.09, 0.55, 0.87…],
[0.53, 0.81, 0.79…],
…,
```

Numerical encoding
(Tokenization + Embedding)

Inputs

Machine Learning Algorithm

(often already exists, if not, you can build one)

Outputs

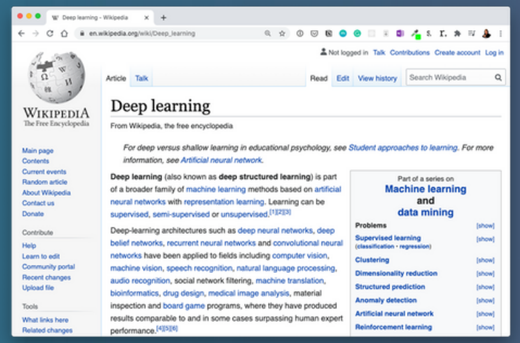
```
[[0.97, 0.03],
[0.81, 0.19],
…,
```

Predicted output

(comes from looking at lots of these)

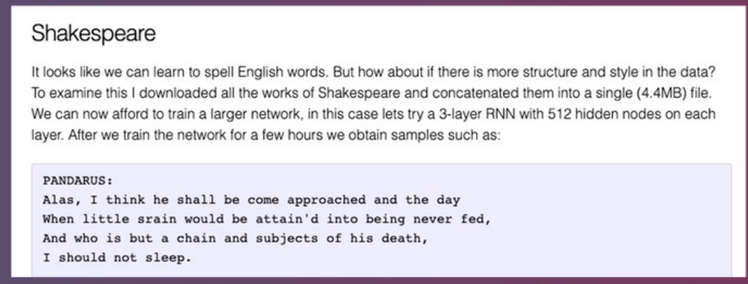
Example NLP problems

"What tags should this article have?"



Machine learning
Representation learning
Artificial intelligence

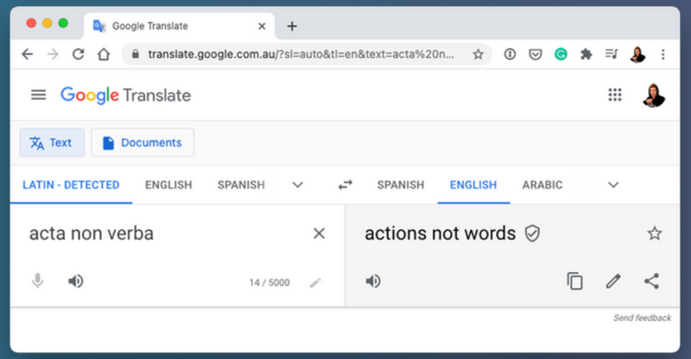
Classification



Source: http://karpathy.github.io/2015/05/21/rnn-e? ectiveness/

Text Generation

These are also referred to as sequence problems



Machine Translation

Voice Assistants

NLP dan Machine Learning

- **Supervised Learning**

  SkimLit adalah salah satu contoh penerapan NLP dengan algorimta supervised learning , RCT Dataset memiliki labeling yang menjadi acuan model untuk mempelajari patern dataset yang nantinya akan memberikan output sesuai dengan labeling pada training.

- **Unsupervised Learning**

  Sentiment Analysis adalah mendeteksi sentimen dan opini dalam data media sosial dalam jumlah besar. Teknik seperti VADER (Valence Aware Dictionary and sEntiment Reasoner) mengklasifikasikan teks sebagai positif, negatif, atau netral tanpa data berlabel.

- **Reinforcement Learning**

  Chat GPT adalah text generated chat bot yang menggunakan Supervised Learning dan Reinforcement Learning sehingga Chat GPT mampu menambahkan input dari user sebagai pembelajaran baru bagi Chat GPT.

“ What and Why Using RNN for NLP Problem ”

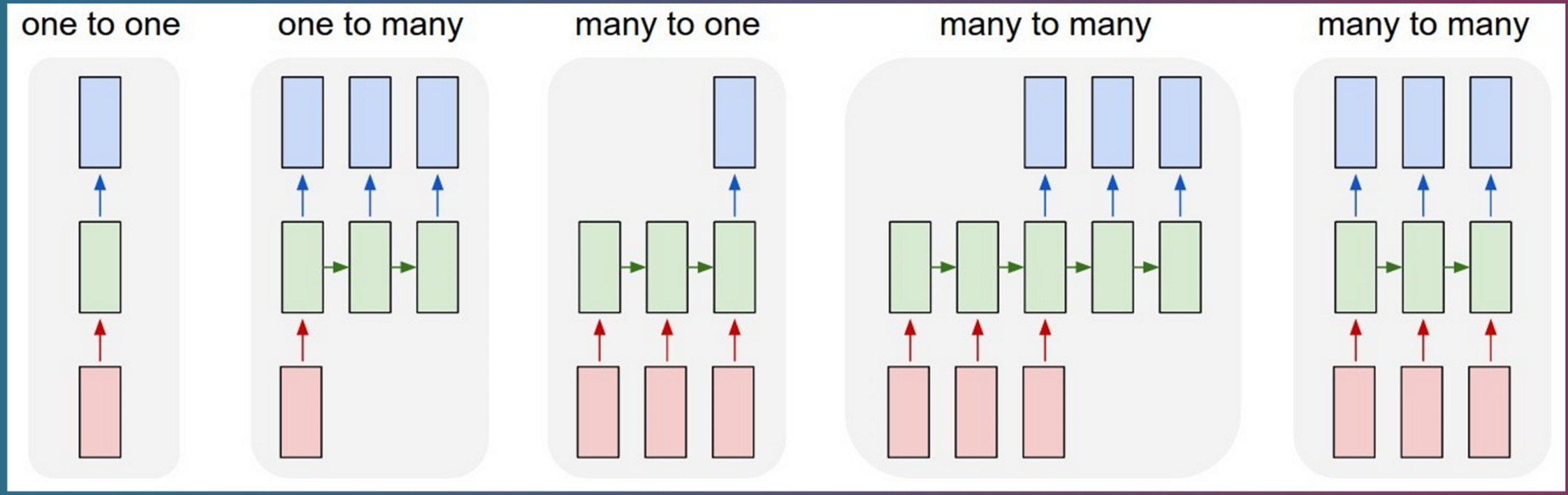
RNN (Recurrent Neural Networks) adalah jenis arsitektur jaringan saraf yang dirancang khusus untuk menangani data urutan, seperti teks atau waktu. RNN memiliki kemampuan untuk "mengingat" informasi sebelumnya dalam urutan, yang memungkinkannya untuk mengatasi masalah bahasa yang bersifat urutan, seperti analisis teks, terjemahan, dan penandaan pos kata.

Kenapa kita menggunakan RNN untuk NLP?

- **Sequential Data Handling :** RNN dapat mempertahankan konteks sebelumnya dalam urutan kata.
- **Language Modeling and Prediction :** RNN dapat dilatih sebagai model bahasa untuk memprediksi distribusi probabilitas kata berikutnya berdasarkan kata-kata sebelumnya dalam sebuah urutan. Kemampuan ini penting untuk tugas seperti pengisian otomatis, prediksi kata selanjutnya, atau menghasilkan teks yang kohesif.
- **Sequential Tasks :** Dalam NLP, banyak tugas yang melibatkan urutan data, seperti penandaan entitas, analisis sentimen, dan ringkasan teks. RNN menawarkan pendekatan yang kuat untuk menangani tugas-tugas ini.

Recurrent Neural Networks (RNN)

Sequence Classification?



Source: http://karpathy.github.io/2015/05/21/rnn-e? ectiveness/

Keluarga Mahasiswa Pusat
عائلة طالب المركزي

“RNN Sequence Classification Example”

- **One to Many**

Input : Image

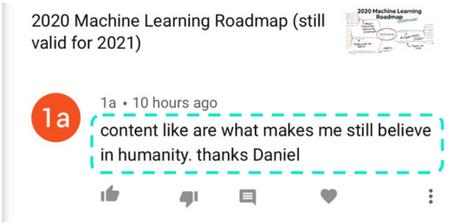
Output : String of Text

“A sledgehammer leaning up against a tire”

Image captioning

- **Many to One**

Input : Comments



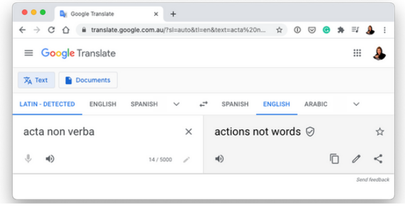
Output : String of Text

“Positive 👍”

Sentiment analysis

- **Many to Many**

Input : String of Text



Output : String of Text

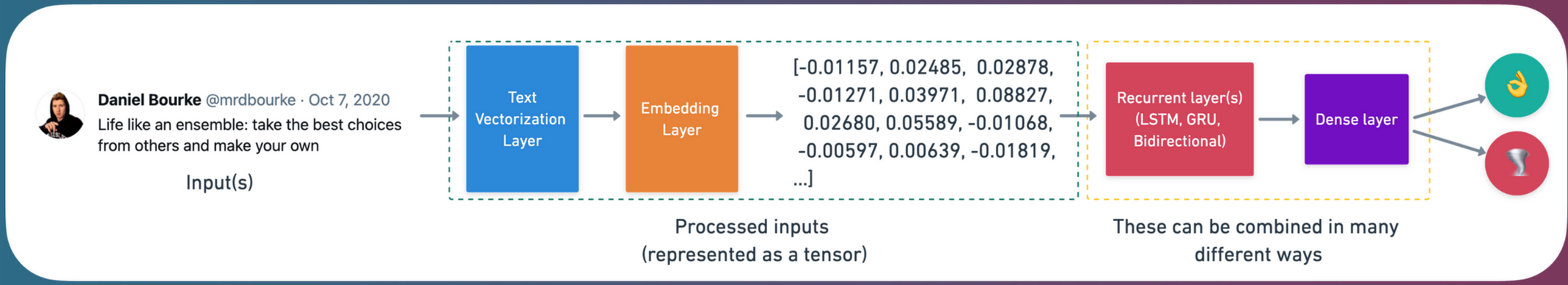
“Action not words”

Machine Translation

Architecture of an RNN

Hyperparameter/Layer type	What does it do?	Typical values
Input text(s)	Target texts/sequences you'd like to discover patterns in	Whatever you can represent as text or a sequence
Input layer	Takes in target sequence	`input_shape = [batch_size, embedding_size]` or `[batch_size, sequence_shape]`
Text vectorization layer	Maps input sequences to numbers	Multiple, can create with `tf.keras.layers.experimental.preprocessing.TextVectorization`
Embedding	Turns mapping of text vectors to embedding matrix (representaiton of how words relate)	Multiple, can create with `tf.keras.layers.Embedding`
RNN cell(s)	Finds patterns in sequences	SimpleRNN, LSTM, GRU
Hidden activation	Adds non-linearity to learned features (non-straight lines)	Usually Tanh (hyperbolic tangent) (`tf.keras.activations.tanh`)
Pooling layer	Reduces the dimensionality of learned sequence features (usually for Conv1D models)	Average (`tf.keras.layers.GlobaelAveragePooling1D`) or Max (`tf.keras.layers.GlobalMaxPool1D`)
Fully connected layer	Further refines learned features from recurrent layers	`tf.keras.layers.Dense`
Output layer	Takes learned features and outputs them in shape of target labels	`output_shape = [number_of_classes]` (e.g. 2 for Diaster, Not Diaster)
Output activation	Adds non-linearities to output layer	`tf.keras.activations.sigmoid` (binary classification) or `tf.keras.activations.softmax`

Standard RNN



Keluarga Mahasiswa Pusat
عائلة طالب المركزي

Let’s code!

https://github.com/KM-Pusat/SkimLit_NLP.git

SkimLit model_0 Algoritma

To investigate the efficacy of @ weeks of daily low-dose oral prednisolone in improving pain , mobility , and systemic low-grade inflammation in the short term and whether the effect would be sustained at @ week in older adults with moderate to sever knee osteoarthritis (OA).

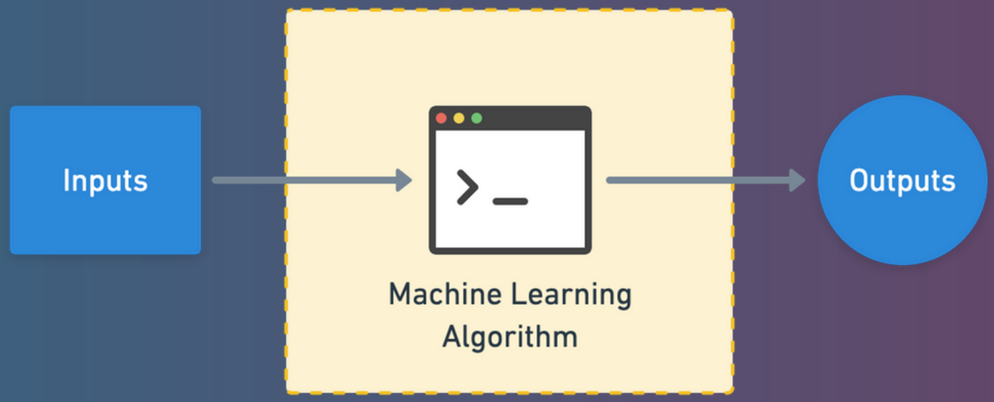
"What section should this sentence belong to?"

- BACKGROUND
- OBJECTIVE
- METHODS
- RESULTS
- CONCLUSION

Actual output

```
[[0.22, 0.98, 0.02…],
[0.09, 0.55, 0.87…],
[0.53, 0.81, 0.79…],
…,
```

Numerical encoding
(Tokenization + Embedding)



(often already exists, if not, you can build one)

```
[0.03, 0.82, 0.13, 0.02, 0.00],
…,
```

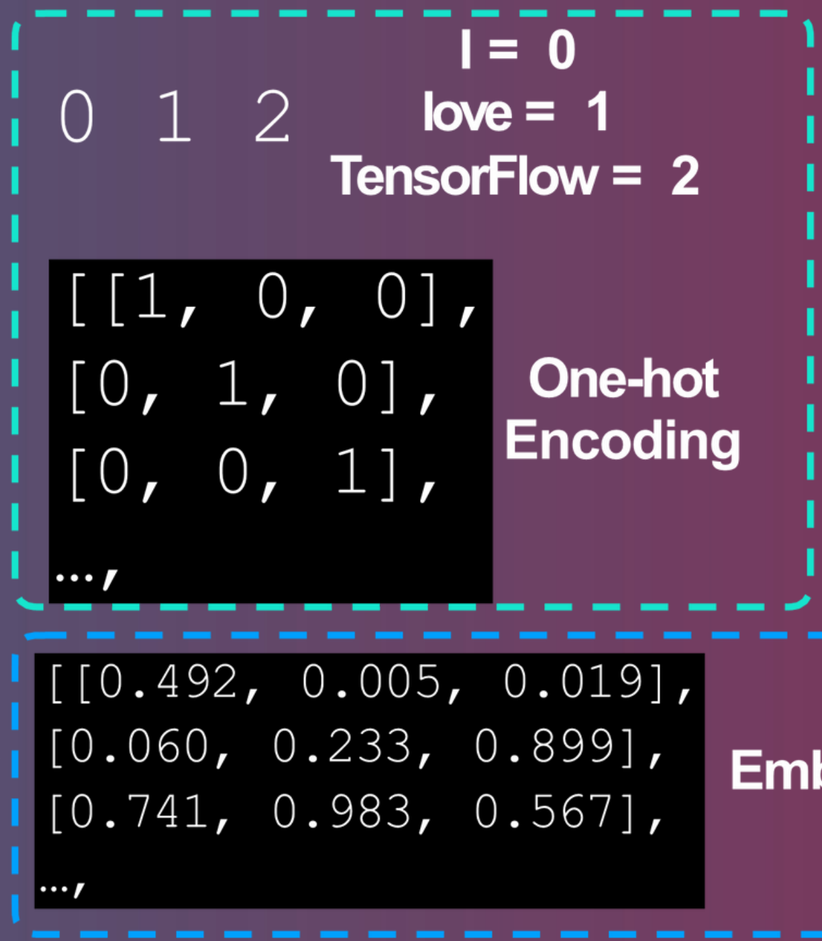
Predicted output
(comes from looking at lots of these)

Tokenization vs Embedding

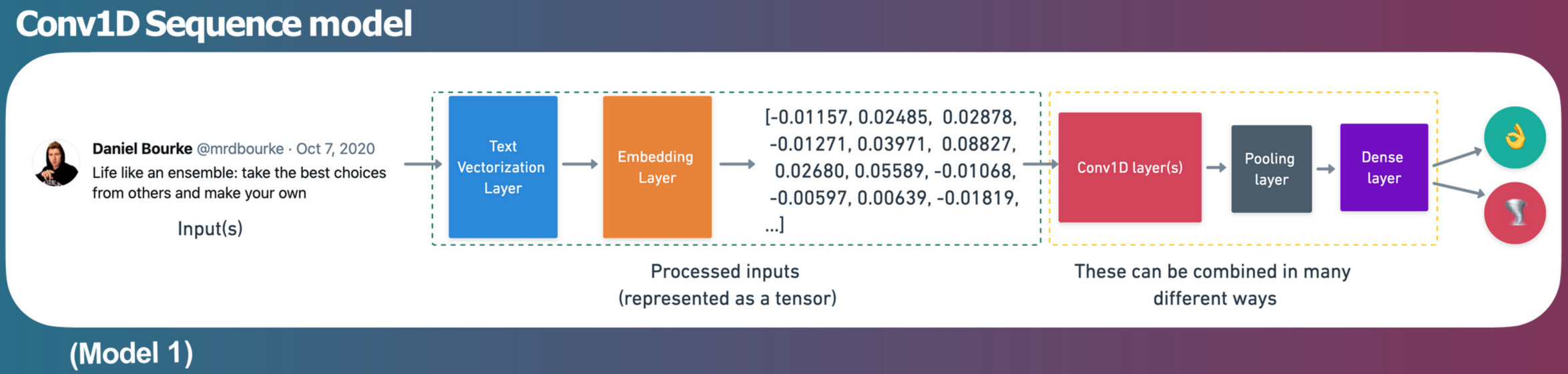
I love TensorFlow

Tokenization — straight mapping from token to number (can be modelled but quickly gets too big)

Embedding — richer representation of relationships between tokens (can limit size +can be learned)



Architecture of a Sequence Model_1 Conv1D



(Model 1)

Improving a model

```python
# 1. Create the model (specified to your problem)
model = tf.keras.Sequential([
  tf.keras.layers.Flatten(input_shape=(28, 28)),
  tf.keras.layers.Dense(4, activation="relu"),
  tf.keras.layers.Dense(10, activation="softmax")
])

# 2. Compile the model
model.compile(loss=tf.keras.losses.BinaryCrossentropy(),
              optimizer=tf.keras.optimizers.Adam(lr=0.001),
              metrics=["accuracy"])

# 3. Fit the model
model.fit(X_train_subset, y_train_subset, epochs=5)
```

Smaller model


```python
# 1. Create the model (specified to your problem)
model = tf.keras.Sequential([
  tf.keras.layers.Flatten(input_shape=(28, 28)),
  tf.keras.layers.Dense(100, activation="relu"),
  tf.keras.layers.Dense(100, activation="relu"),
  tf.keras.layers.Dense(100, activation="relu"),
  tf.keras.layers.Dense(10, activation="softmax")
])

# 2. Compile the model
model.compile(loss=tf.keras.losses.BinaryCrossentropy(),
              optimizer=tf.keras.optimizers.Adam(lr=0.0001),
              metrics=["accuracy"])

# 3. Fit the model
model.fit(X_train_full, y_train_full, epochs=100)
```

Larger model

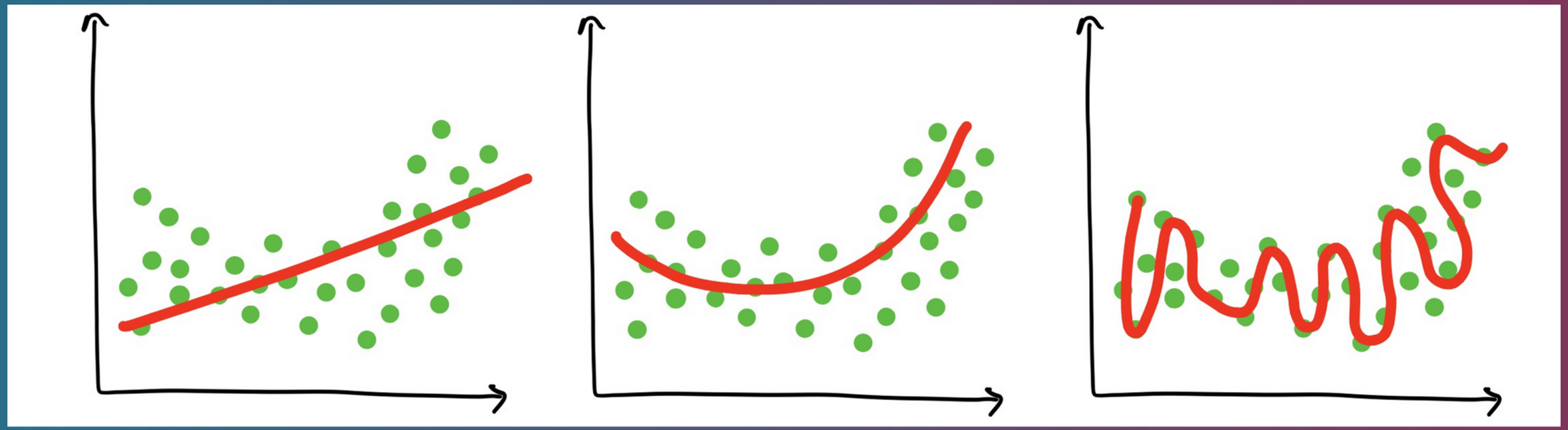
Cara umum untuk meningkatkan Deep Learning model

- Menambahkan layers
- Meningkatkan number of hidden units
- Mengubah activation functions
- Mengubah optimization function
- Mengubah learning rate
- Menyesuaikan dengan lebih banyak data
- Menyesuaikan lebih lama

What is overfitting?

Overfitting - ketika sebuah model mempelajari pola-pola dalam kumpulan data tertentu dan tidak dapat menggeneralisasi data yang tidak terlihat.

Misalnya, seorang siswa yang mempelajari materi pelajaran terlalu keras dan kemudian tidak dapat mengerjakan ujian akhir dengan baik. Atau mencoba mempraktikkan pengetahuan mereka di tempat kerja dan mendapati bahwa apa yang mereka pelajari tidak ada hubungannya dengan dunia nyata.



Underfitting | **Balanced** (goldilocks zone) | **Overfitting**

Keluarga Mahasiswa Pusat
عائلة طالب المركزي

Terima Kasih

شكرًا لك

بركة العلم في الجماعة

" Berkahnya Ilmu Terletak pada Kebersamaan."
Sayyid Muhammad bin Alawi al-Maliki al-Hasani

See our next project

https://github.com/KM-Pusat