HUMANE AI NET

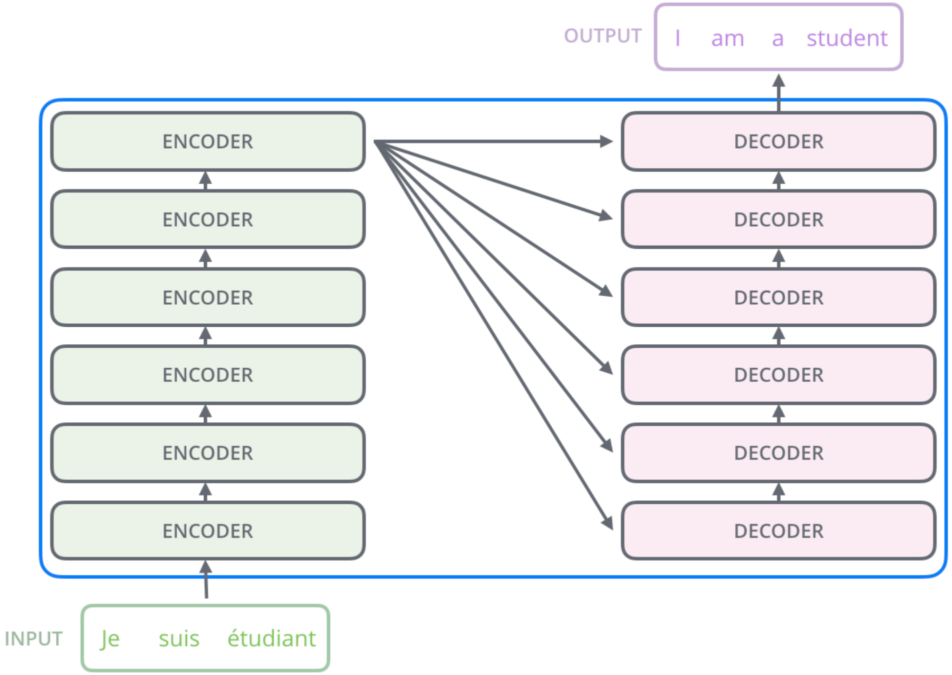
Multimodal Perception and Interaction with Transformers

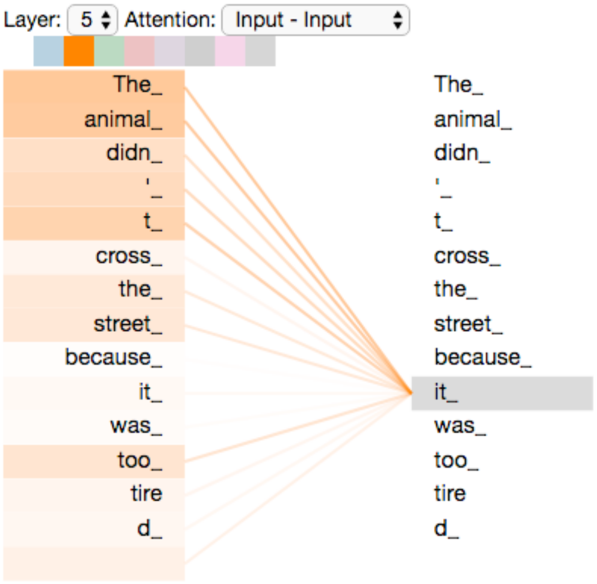
Francois Yvon, Camille Guinaudeau, Marc Evrard
Univ Paris Saclay (LISN CNRS)

James L. Crowley
Grenoble Institut Polytechnique, Univ Grenoble Alpes

Transformers use attention to associate mutually relevant entities

Transformers are stacked layers of Encoders and Decoders that use attention to associate mutually relevant entities.

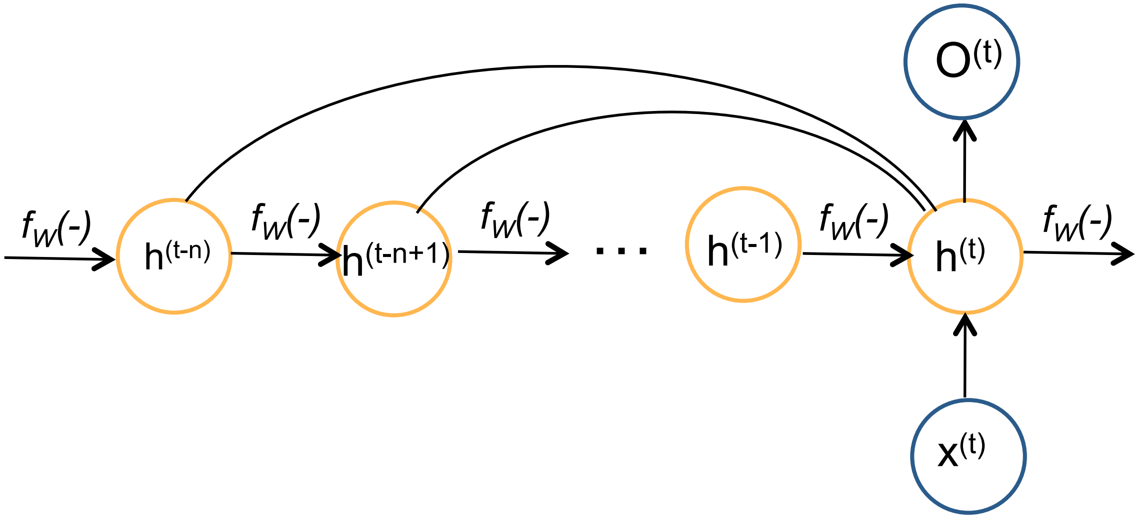



Layer: 5 Attention: Input - Input

Images from Jay Alammar, The Illustrated Transformer (http://jalammar.github.io/illustrated-transformer/)

Attention Extends Time for Recurrent Networks

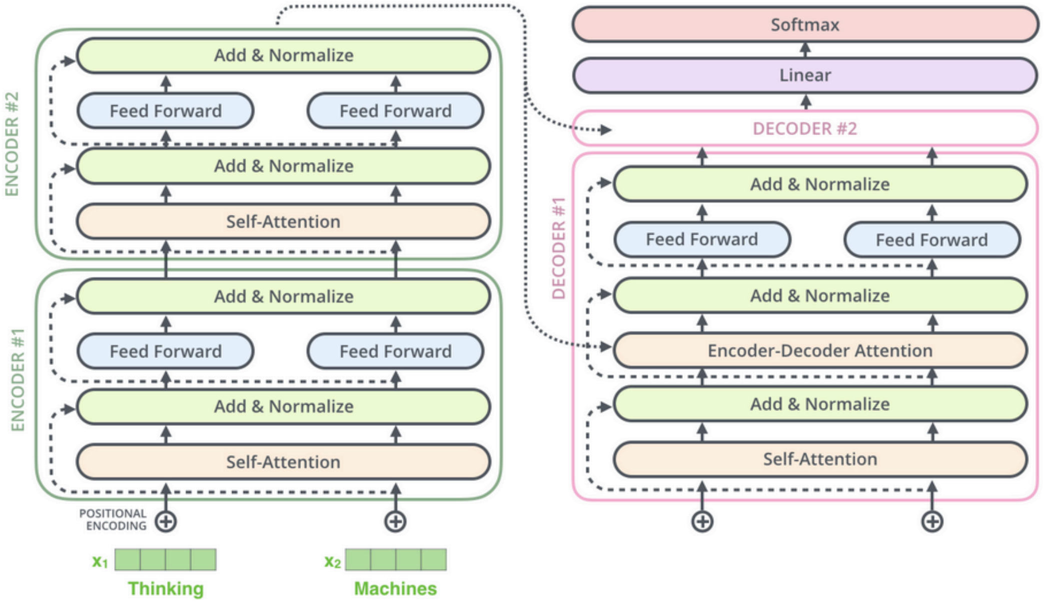
Attention was originally proposed as a soft search mechanism to extend the temporal range of Recurrent Networks (Bahdanau et al 2015).



D. Bahdanau, K. Cho, and Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In 3rd International Conference on Learning Representations, 2015

Attention is all you need

In 2017, a revolutionary paper by Vaswani et al [1] from Google showed that the deep convolutional and recurrent networks using layers of could be completely replaced with attention.

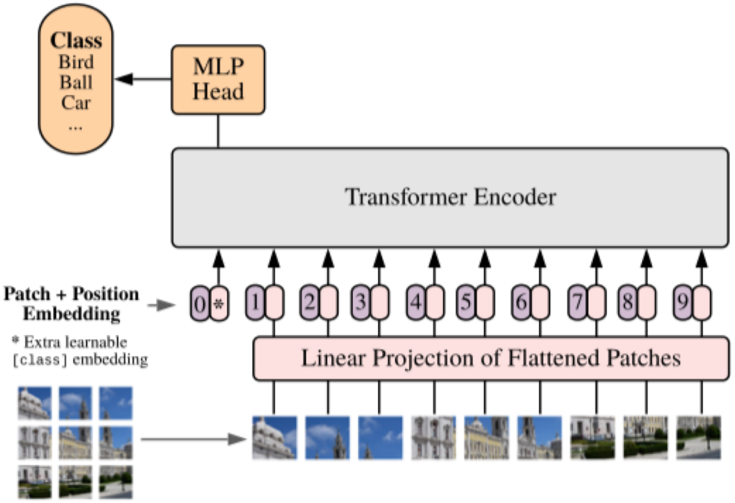


From Jay Alamar, The Illustrated Transformer: http://jalammar.github.io/illustrated-transformer/

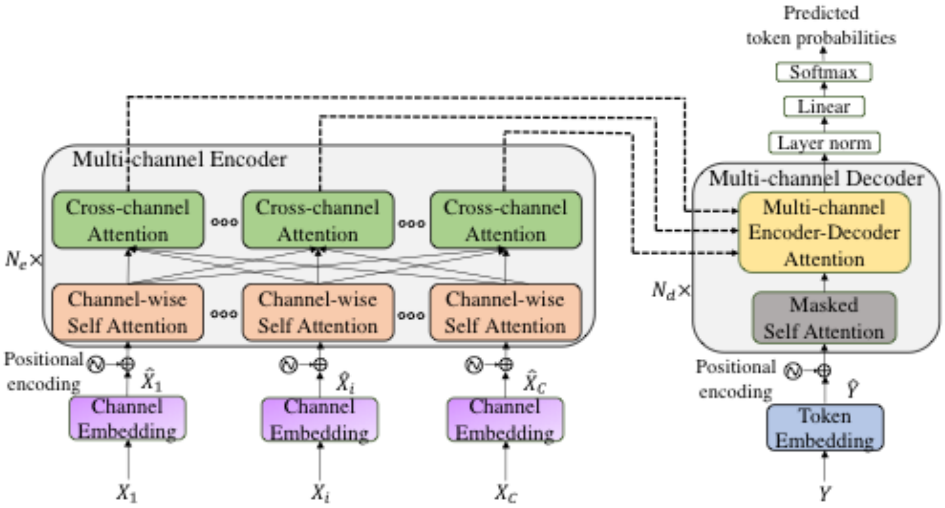
[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. . Attention is all you need. 2017

Extensions to Vision and Speech

Transformers are rapidly replacing Deep Recurrent Networks and Convolutional networks for **Speech Recognition** and **Computer Vision.**



Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0 *
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

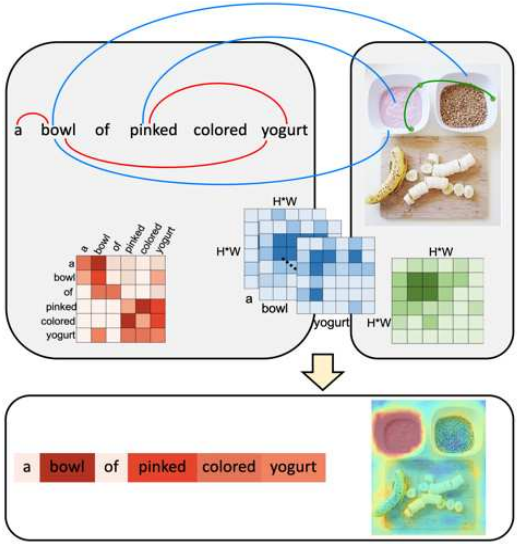


Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR, 2021

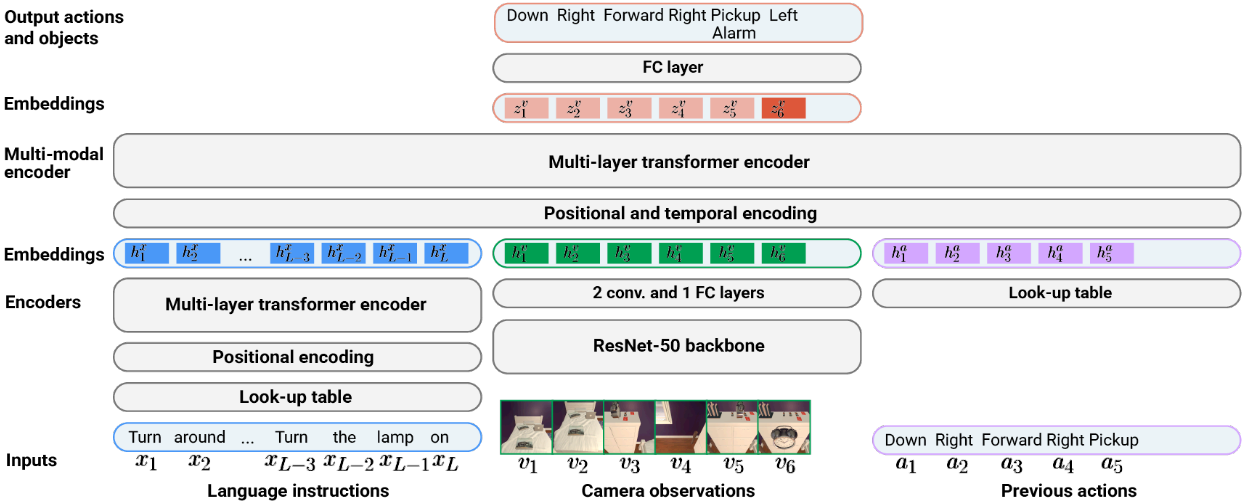
Chang, F. J., Radfar, M., Mouchtaris, A., King, B., & Kunzmann, S. (2021, June). End-to-End Multi-Channel Transformer for Speech Recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5884-5888). IEEE, 2021

Multimodal Perception with Transformers

Recent results indicate that Transformers are well adapted for **multi-modal Perception, Robotics and Human-Computer Interaction**


a bowl of pinked colored yogurt
H*W
a
bowl
of
pinked
colored
yogurt
a bowl of pinked colored yogurt

Ye et. Al. “Cross-Modal Self-Attention Network for Referring Image Segmentation”, CVPR 2019, June, 2019.


Output actions and objects
Down Right Forward Right Pickup Left Alarm
FC layer
Embeddings
Multi-modal encoder
Multi-layer transformer encoder
Positional and temporal encoding
Embeddings
Encoders
Multi-layer transformer encoder
Positional encoding
Look-up table
2 conv. and 1 FC layers
ResNet-50 backbone
Look-up table
Inputs
Turn around ... Turn the lamp on
x_1 x_2 x_{L-3} x_{L-2} x_{L-1} x_L
Language instructions
v_1 v_2 v_3 v_4 v_5 v_6
Camera observations
Down Right Forward Right Pickup
a_1 a_2 a_3 a_4 a_5
Previous actions

Pashevich, A., Schmid, C. and Sun, C., Episodic Transformer for Vision-and-Language Navigation, Int. Conf. on Computer Vision, ICCV 2021, Oct. 2021.

Multimodal Perception with Transformers

Plan:

Transformers in Natural Language Processing (François Yvon, 1h30)

- Text classification and language models
- The Transformer architecture
- Encoder-Decoder architecture for Neural Machine translation

Transformers in Speech (Marc Evrard, 45 minutes)

- Speech Recognition
- Attention for Speech Recognition
- Transformers for Speech Recognition

Transformers in Vision (Camille Guinaudeau, 45 minutes)

- From CNN to Vision Transformer
- Vision Transformers
- Multi-Modal Transformer and Temporal encoding

Conclusions (James Crowley, 15 minutes)

- Research Challenges, Data Sets and Open Problems