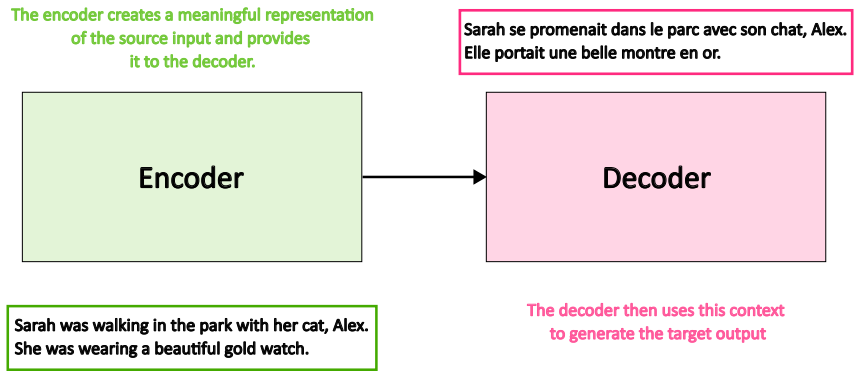
Transformers (Part 1)

Dr. Alireza Aghamohammadi

Encoder-Decoder Architecture

- The Transformer model utilizes an encoder-decoder architecture.
- This pattern is widely used in deep learning for tasks such as machine translation, question answering, and image captioning.
- The encoder processes the input and converts it into a **context representation**, while the decoder uses this context to generate a meaningful output.



Sequence to Sequence Model

- **Input:** $x = (x_1, x_2, x_3, \cdots, x_m)$
- **Output:** $y = (y_1, y_2, y_3, \cdots, y_n)$

$$y' = \underset{y}{\text{argmax}}\, P(y|x, \theta)$$
$$= \prod_{t=1}^{n} P(y_t|x, \theta, y_1, y_2, \cdots, y_{t-1})$$

- At time t, the model predicts a probability distribution $P(y_t|x, \theta, y_1, y_2, \cdots, y_{t-1})$