Attention + Transformers

$X^i_t \rightarrow \text{Encoder} \rightarrow \text{Decoder} \rightarrow y_t^i$

$X^i_t \rightarrow \text{Encoder} \rightarrow \text{Decoder} \rightarrow y_t^i$

Increase feature complexity

Output appropriate sequence based on context

Bottlenecked problem: context might have to encode a very long input sequence.
This new context is very large.
So, an "attention" mechanism was added.

I am a student

<table>
<thead>
<tr>
<th>Context</th>
<th>Hidden State #1</th>
<th>Hidden State #2</th>
<th>Hidden State #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Je</td>
<td>Hidden State #1</td>
<td>Hidden State #2</td>
<td>Hidden State #3</td>
</tr>
<tr>
<td>suis</td>
<td>Hidden State #2</td>
<td>Hidden State #2</td>
<td>Hidden State #3</td>
</tr>
<tr>
<td>étudiant</td>
<td>Hidden State #3</td>
<td>Hidden State #2</td>
<td>Hidden State #3</td>
</tr>
</tbody>
</table>

Increase feature complexity
Output appropriate sequence based on context
Contain states from past time steps
Image Capturing Example

- CNN ➔ Image Features ➔ Alignment ➔ Attention
- Decoder RNN
- Decoder RNN
- Decoder RNN

Person ➔ wearing ➔ a hat [END]
Transformer Models
(Not an RNN)

Encoder Block

Je suis étudiant

I am a student

Attention

Feed forward NN