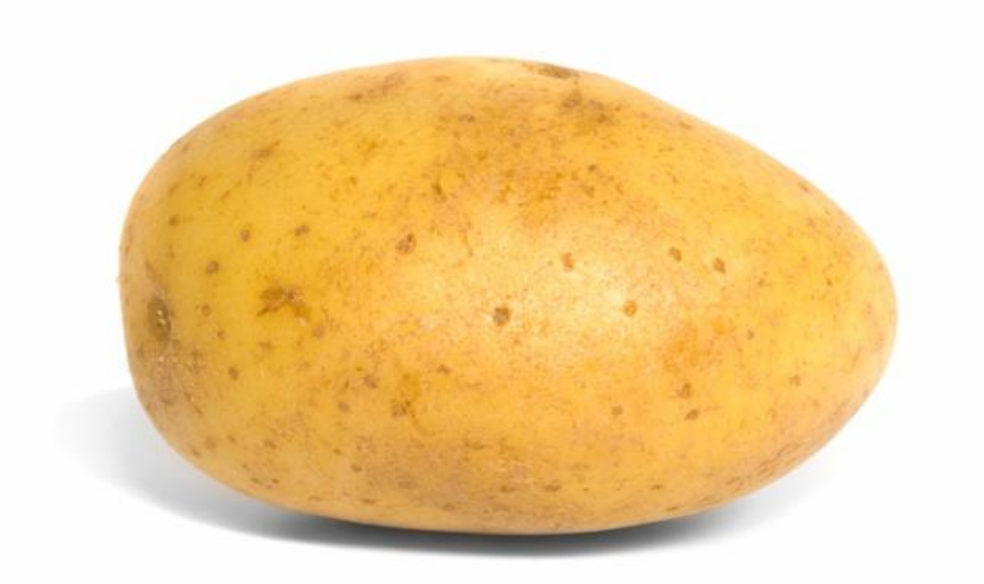
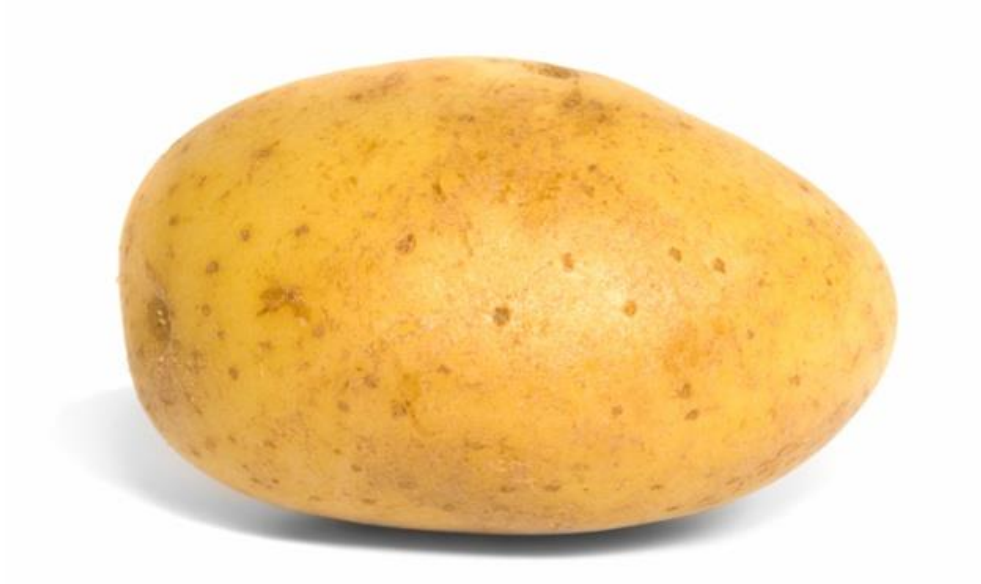
Sequence-to-Sequence

Timo Wang

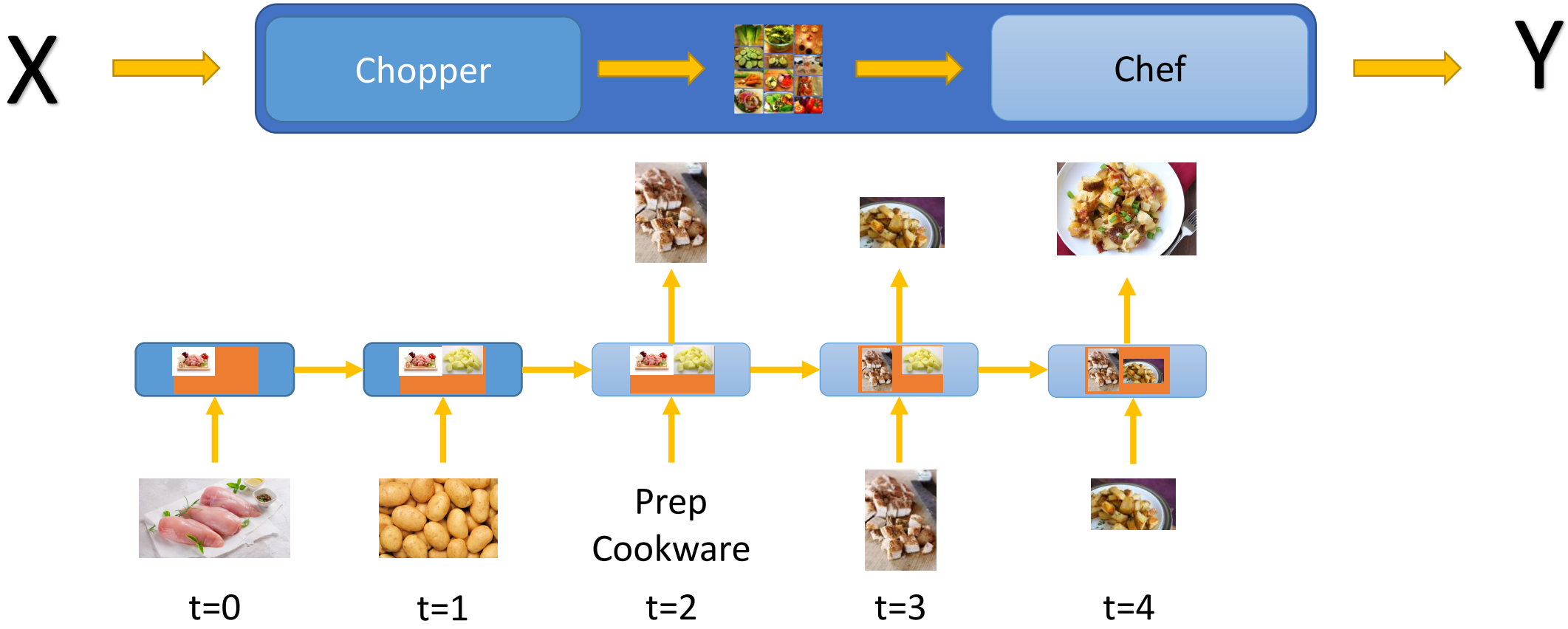





©TheSweetSeed.com


FOOD STUFF
GRAN GUSTO
VINAGRE DE JEREZ
LA DALIA

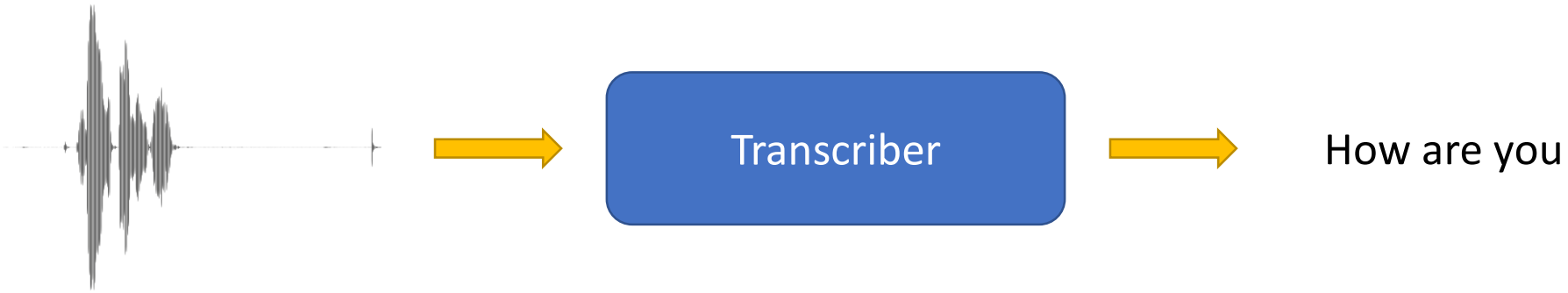
Chopper & Chef



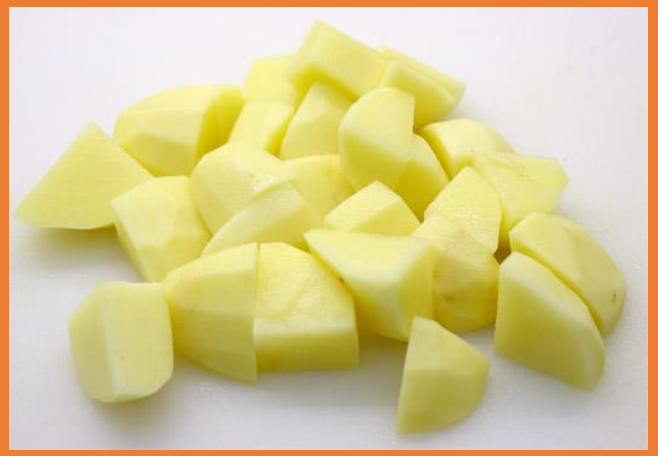
Seq2Seq


X
Chopper
Chef
Y


How are you
Translator
Wie geht’s

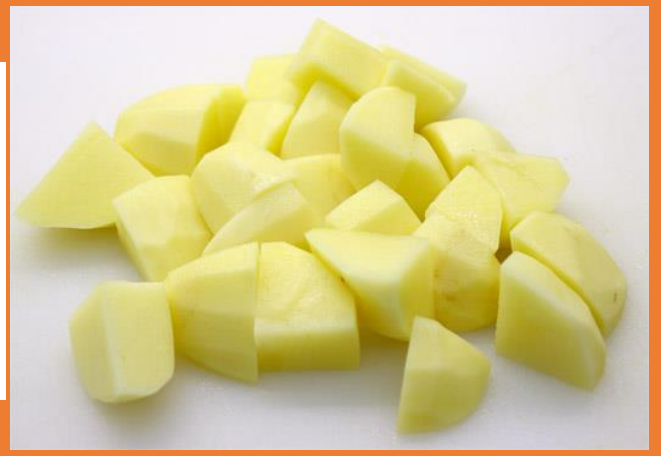

Transcriber
How are you

Limitations



Limitations

Limitations



Limitations

Improved Approach

Improved Approach

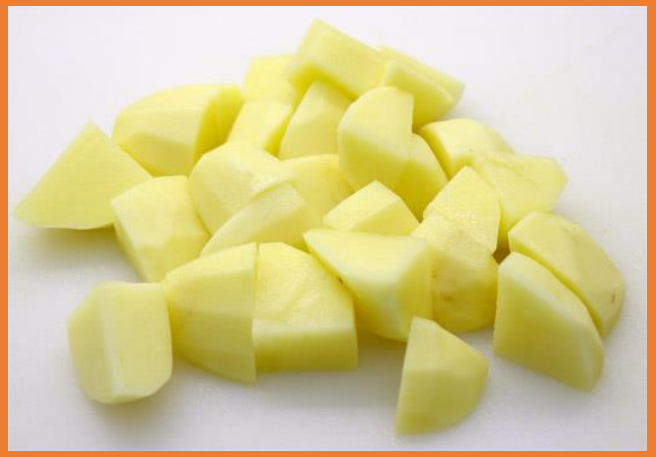
Improved Approach

Improved Approach







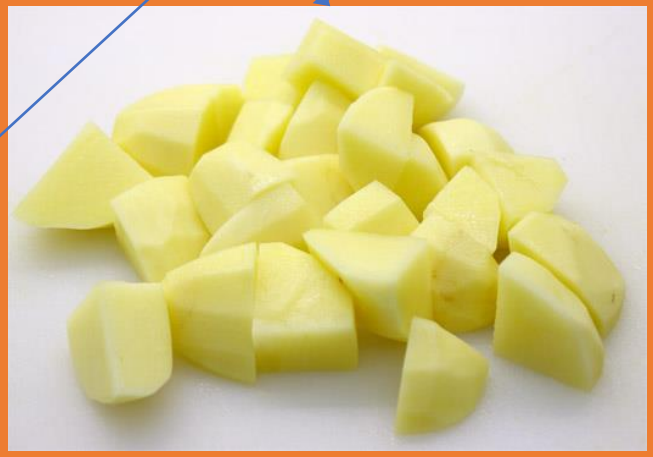


Even Better



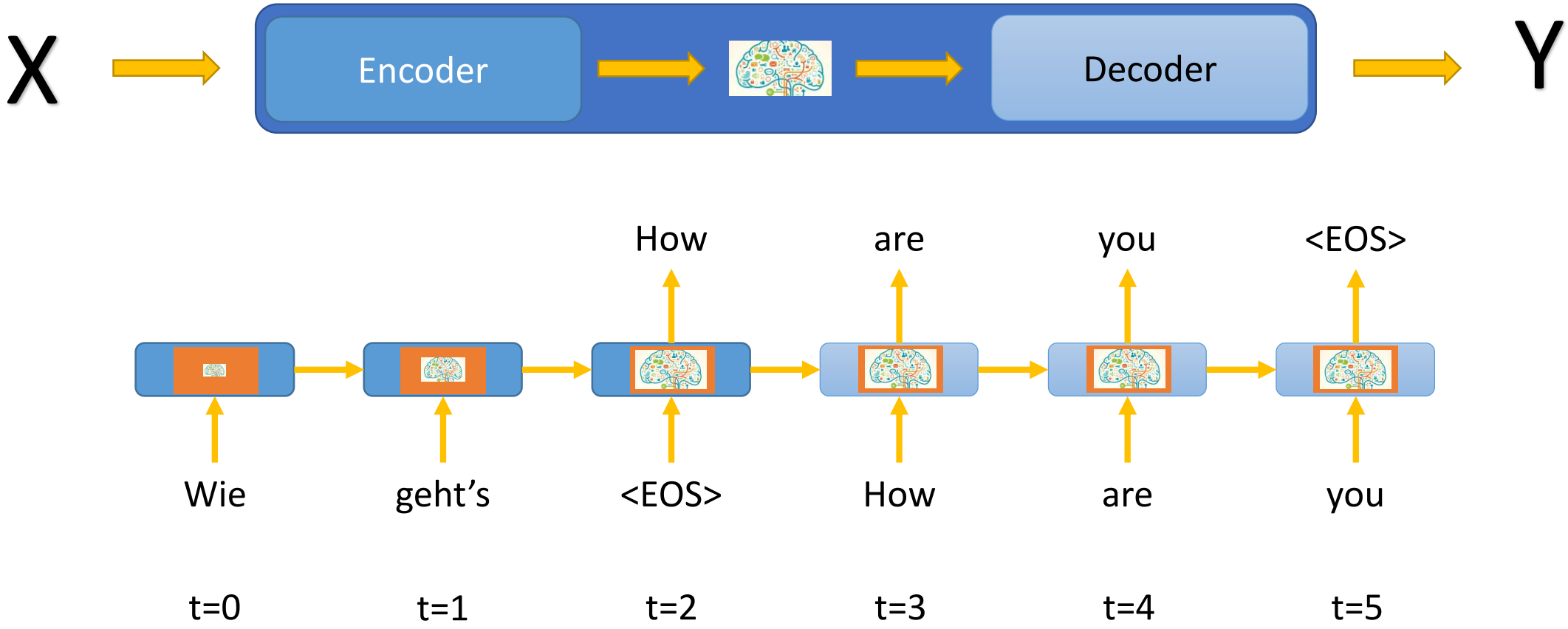






Even Better

Encoder & Decoder



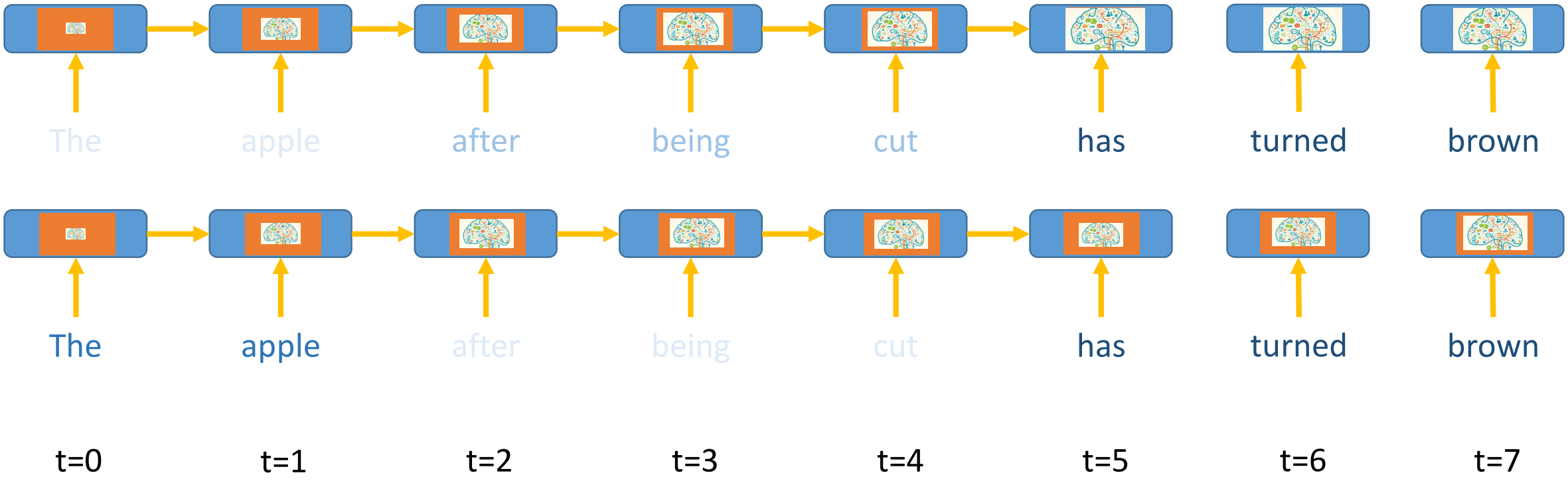
Limitations

The apple, after being cut into pieces, has turned brown.

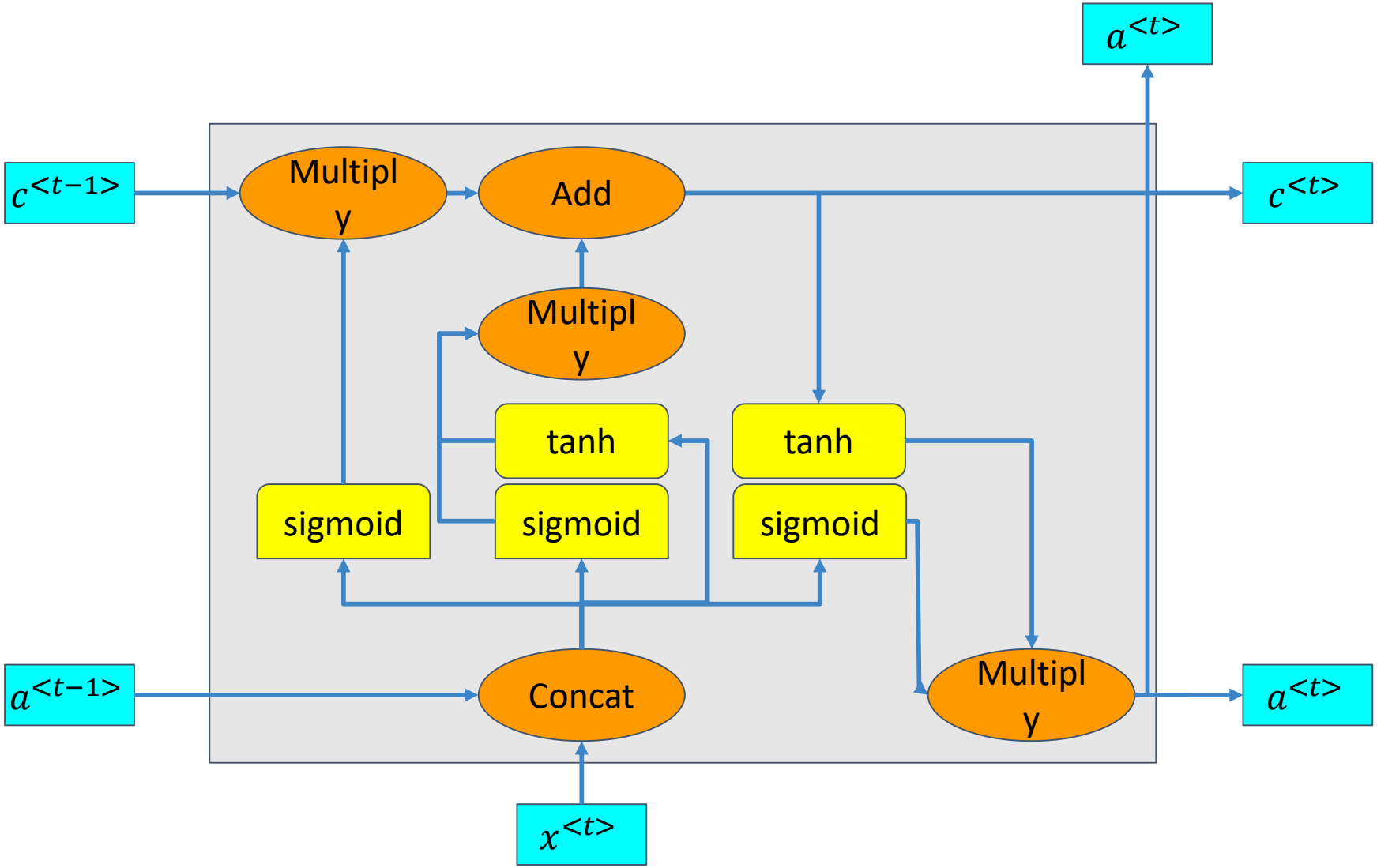
The apple, after being cut into pieces, has turned brown.

The apple, after being cut into pieces, has turned brown.

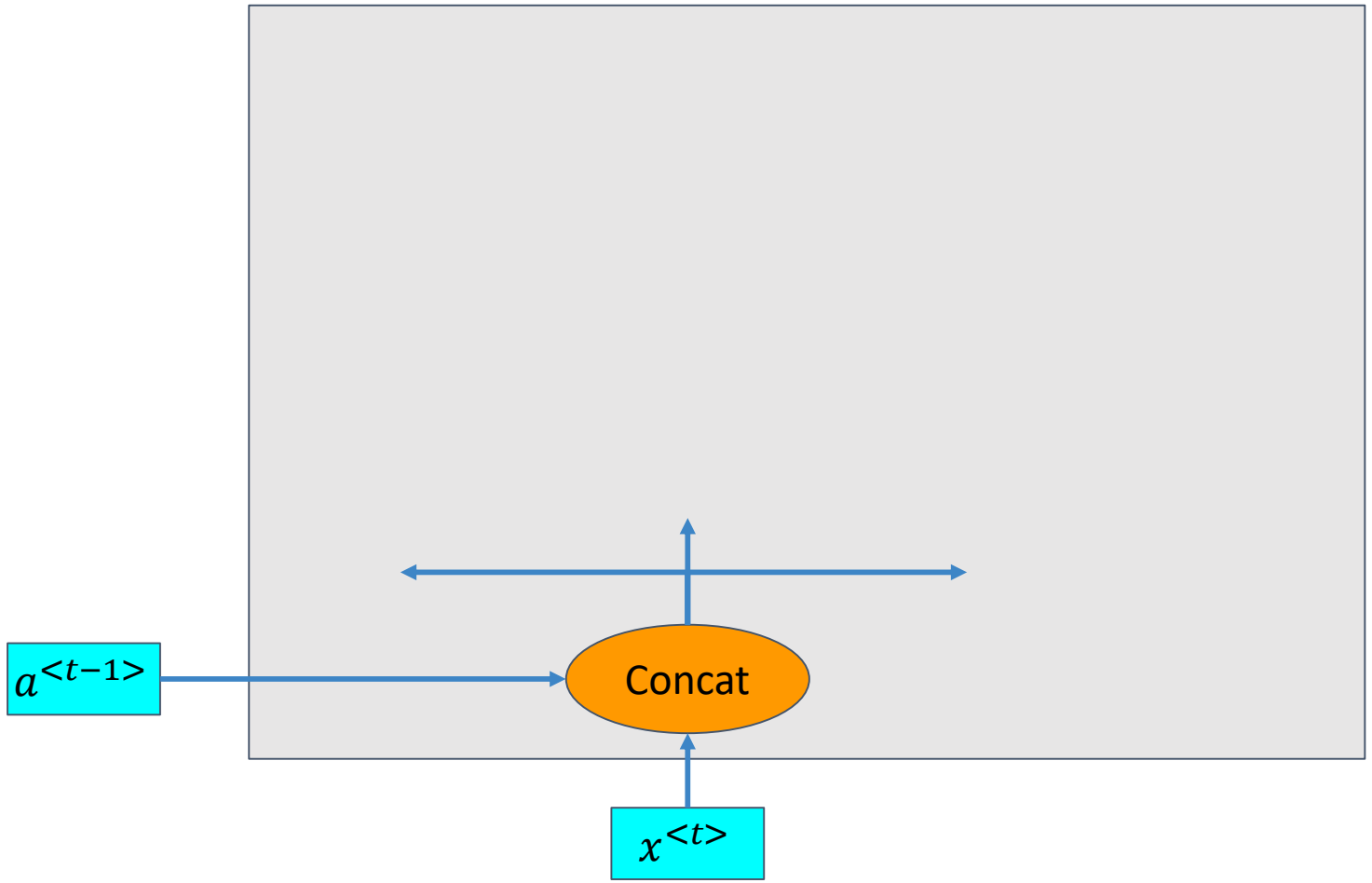
LSTM Concept



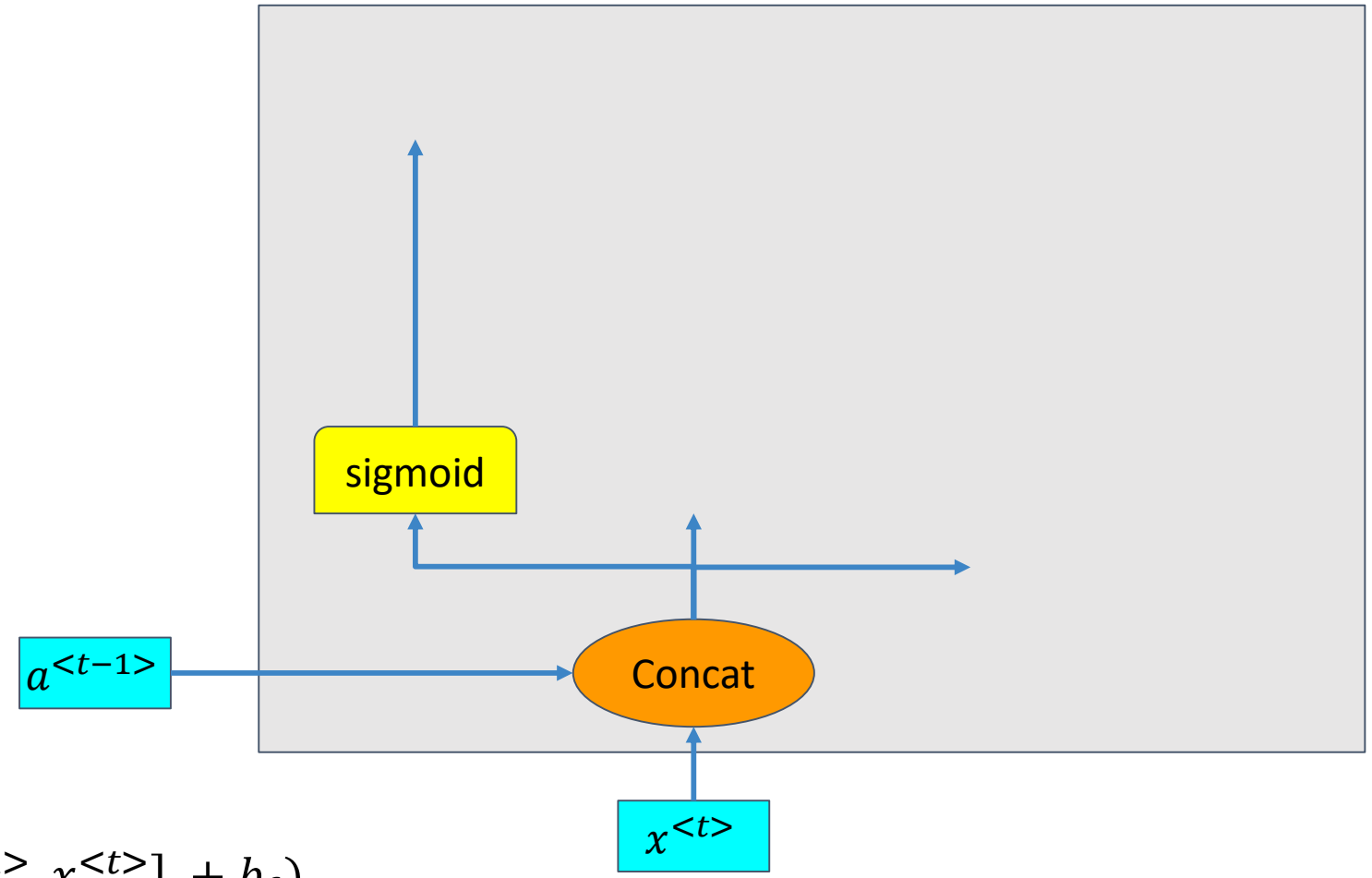
Long Short-Term Memory Unit



LSTM Unit

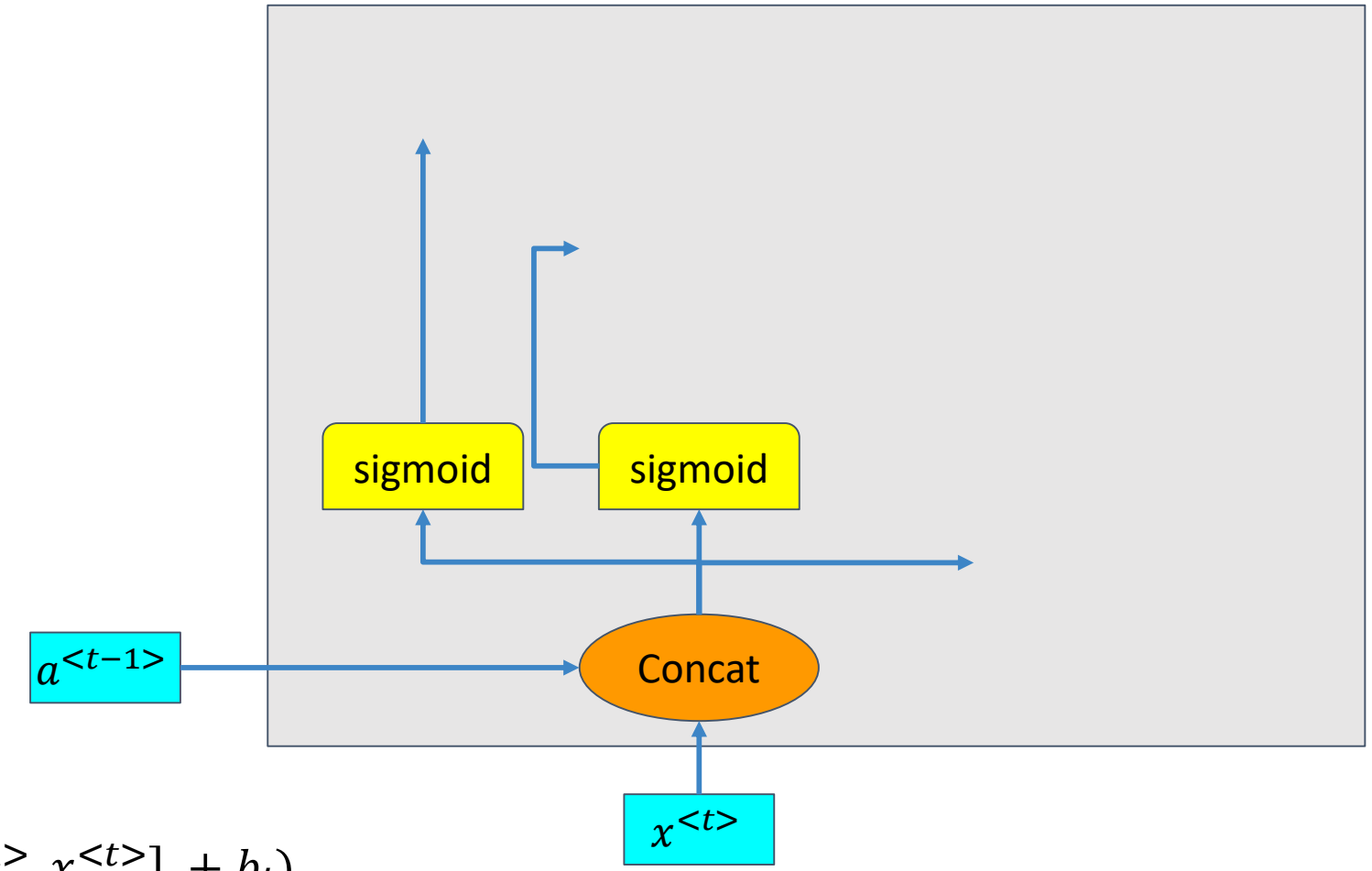


LSTM Unit



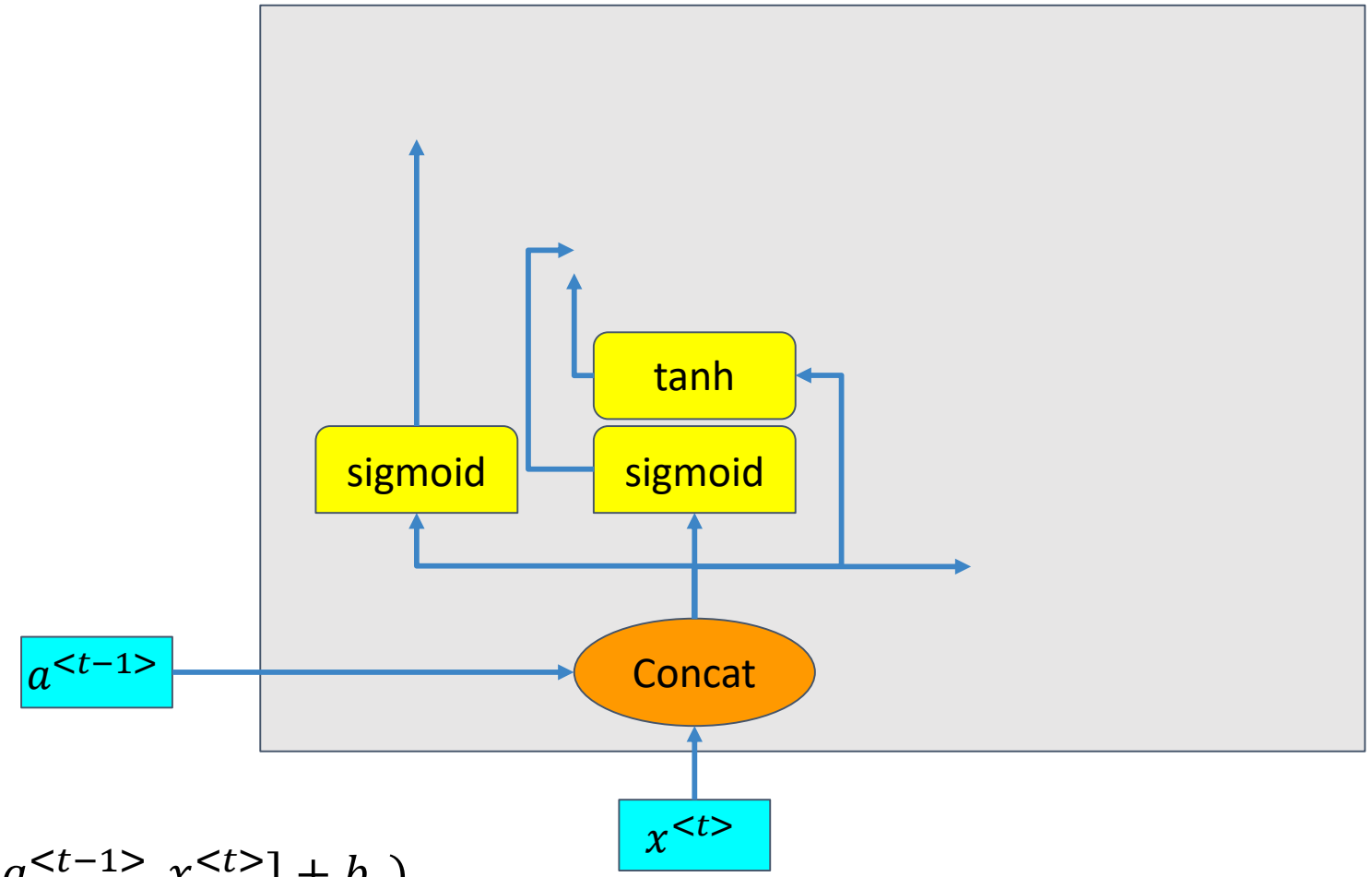
$$\Gamma_f = \sigma(W_f[a^{<t-1>}, x^{<t>}] + b_f)$$

LSTM Unit



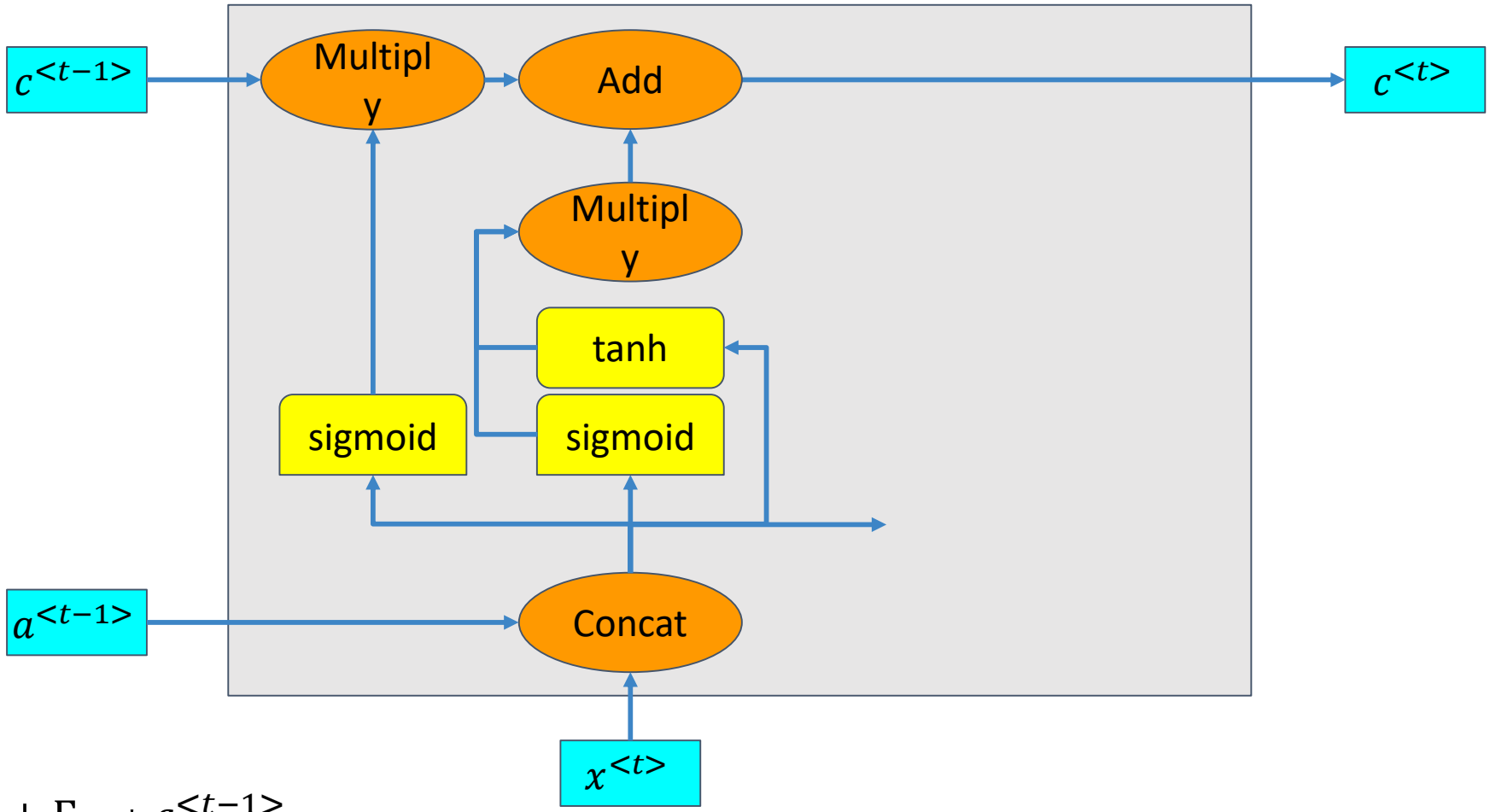
$$\Gamma_i = \sigma(W_i[a^{<t-1>}, x^{<t>}] + b_i)$$

LSTM Unit



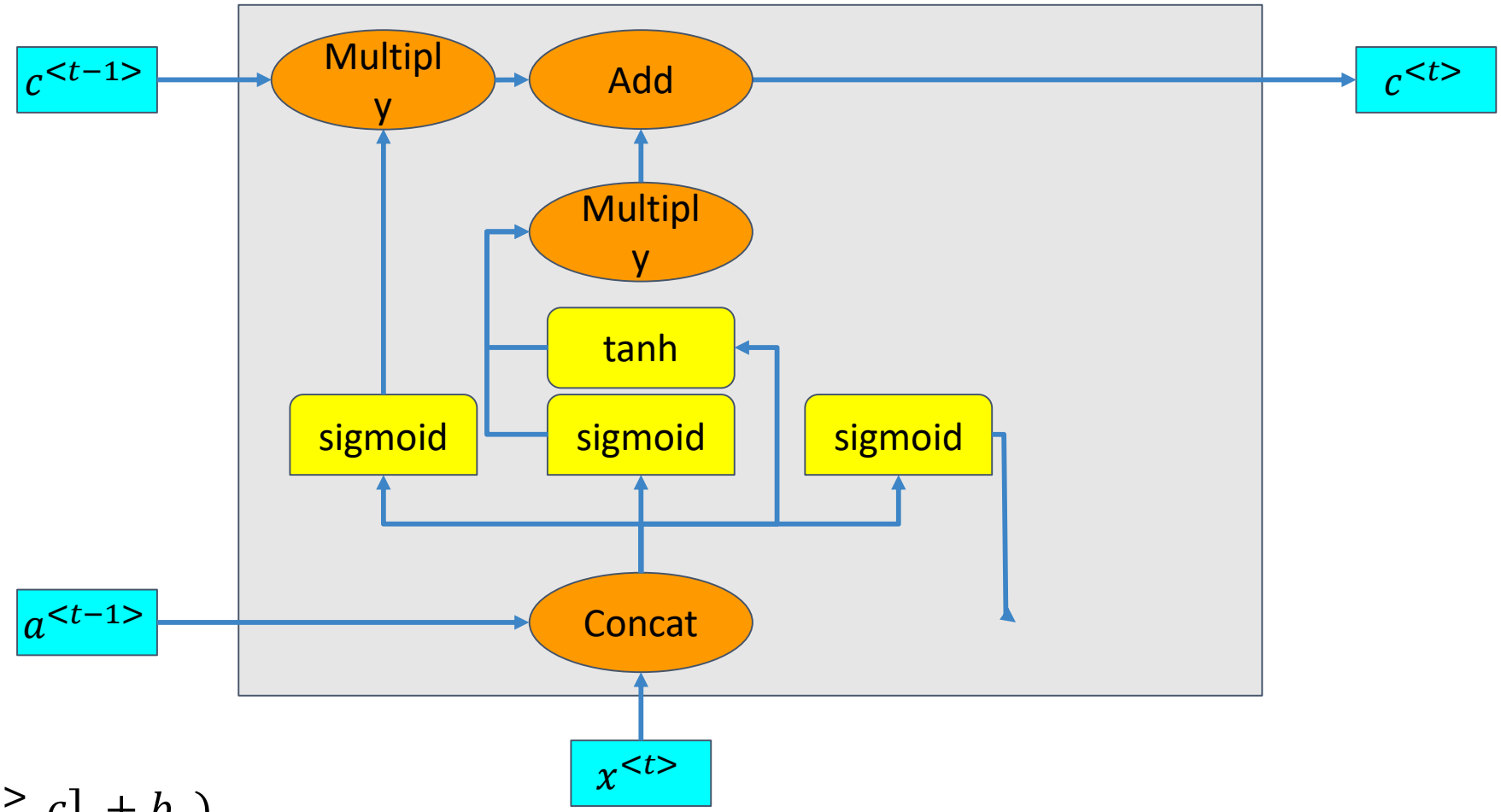
$$\tilde{c}^{<t>} = \tanh(W_c[a^{<t-1>}, x^{<t>}] + b_c)$$

LSTM Unit



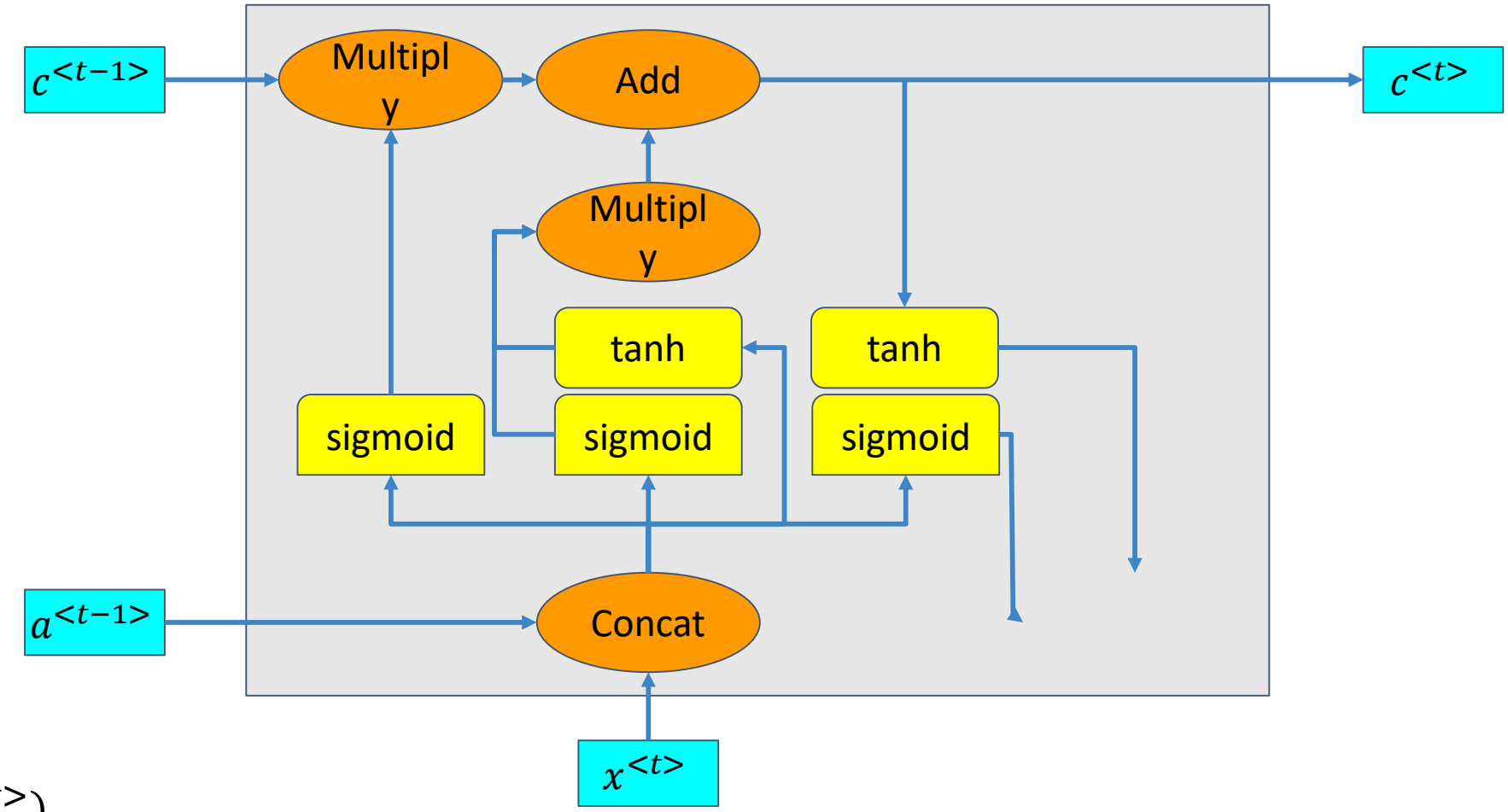
$$c^{<t>} = \Gamma_i * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>}$$

LSTM Unit



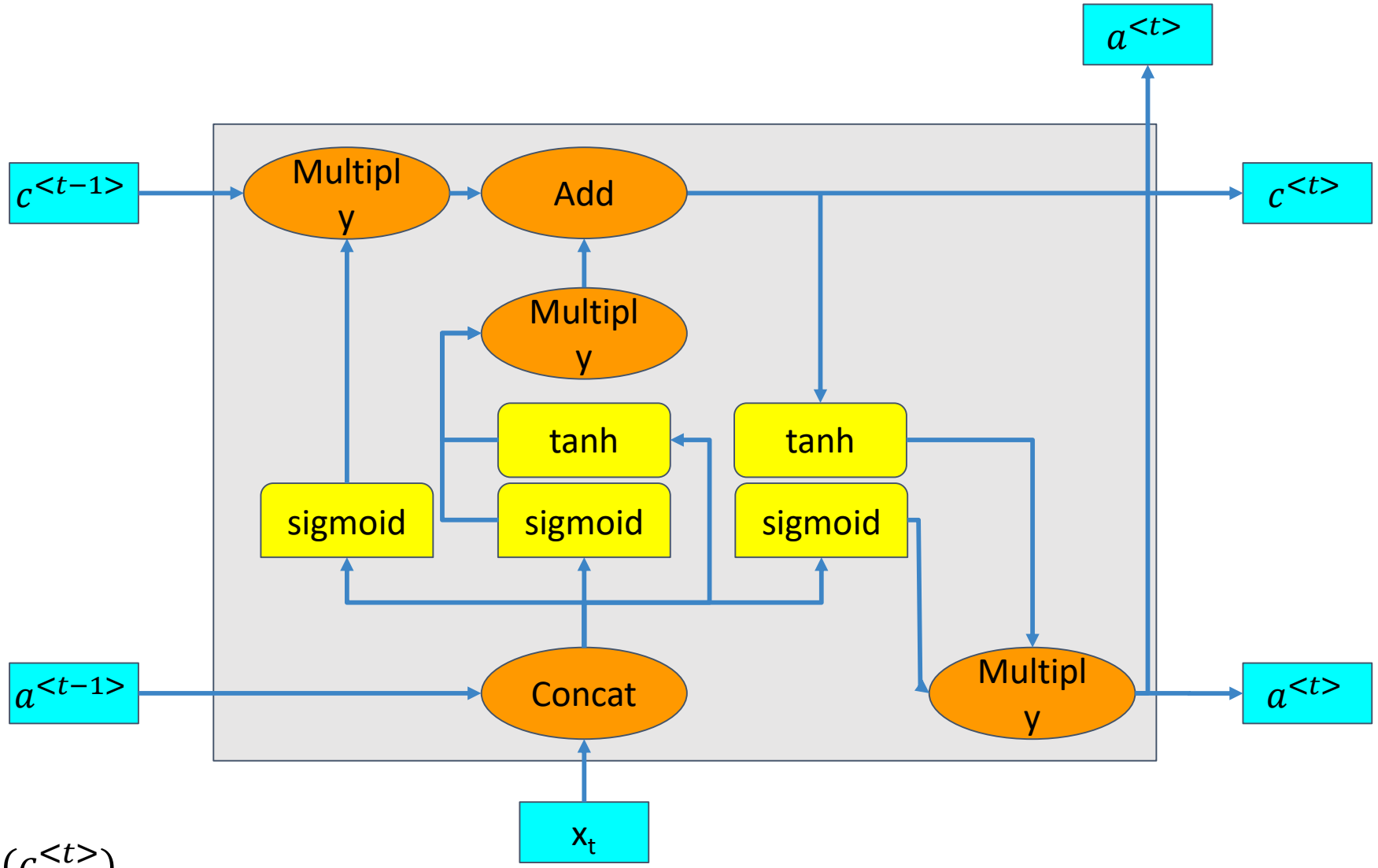
$$\Gamma_o = \sigma(W_o[a^{<t-1>}, c] + b_o)$$

LSTM Unit



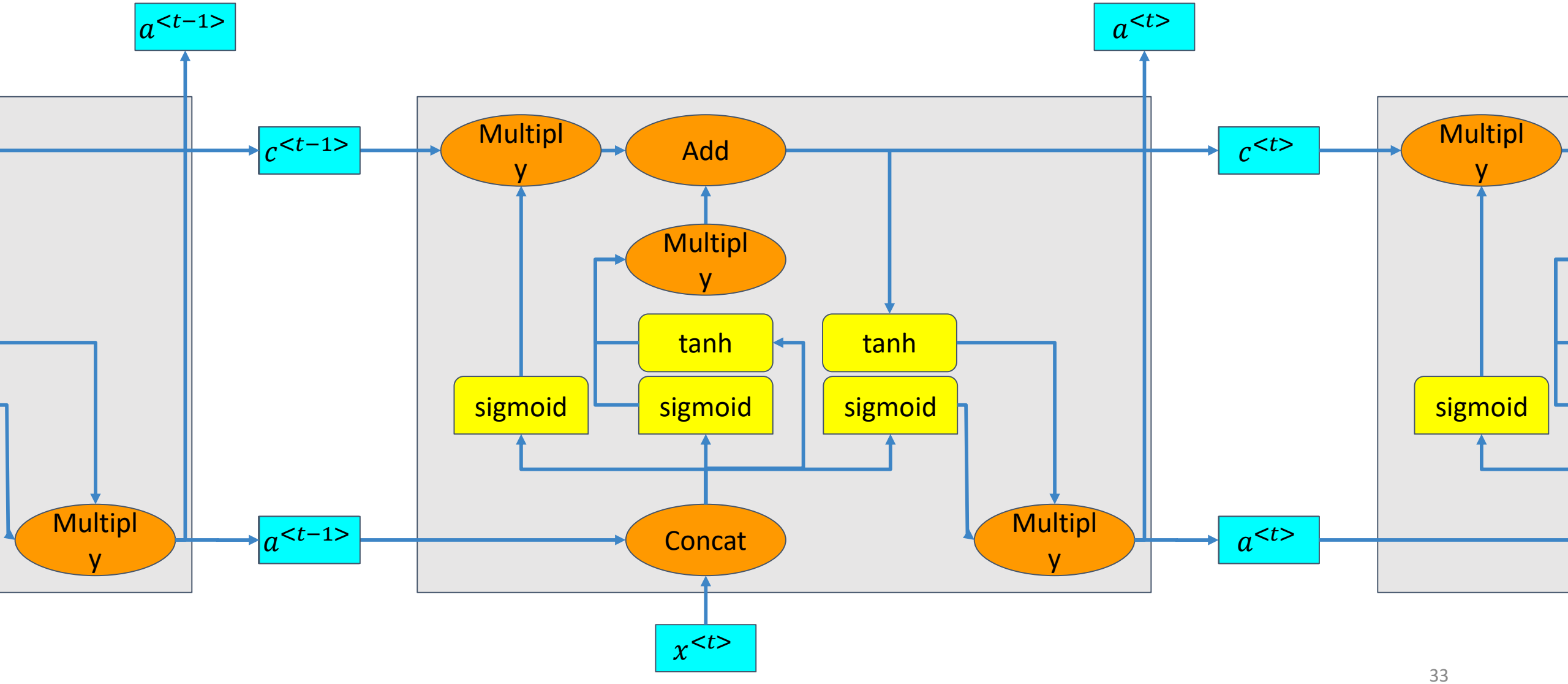
$\tilde{a}^{<t>} = \tanh(c^{<t>})$

LSTM Unit

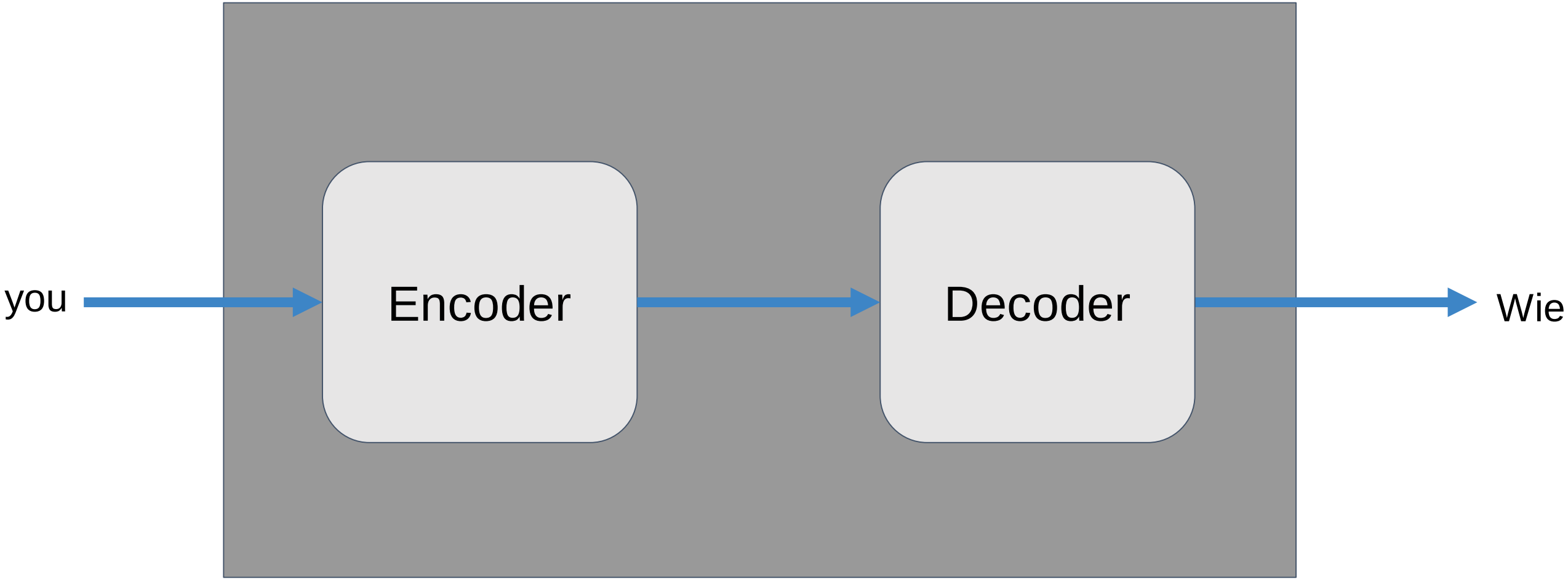


$$a^{<t>} = \Gamma_o \ * \tanh(c^{<t>})$$

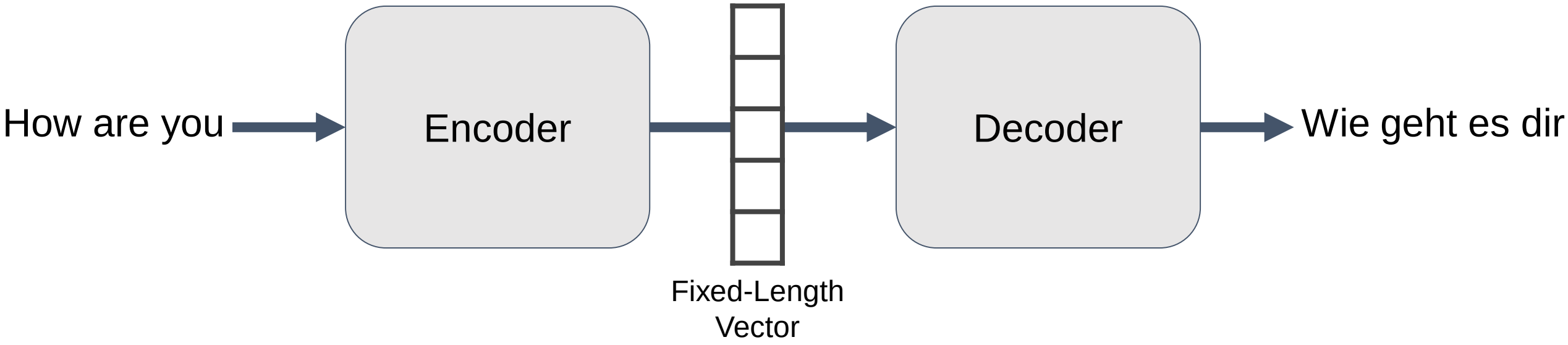
LSTM


$a^{<t-1>}$
$a^{<t>}$
$c^{<t-1>}$
$c^{<t>}$
Multiply
Add
Multiply
tanh
tanh
sigmoid
sigmoid
sigmoid
Multiply
Multiply
$a^{<t-1>}$
Concat
Multiply
$a^{<t>}$
$x^{<t>}$

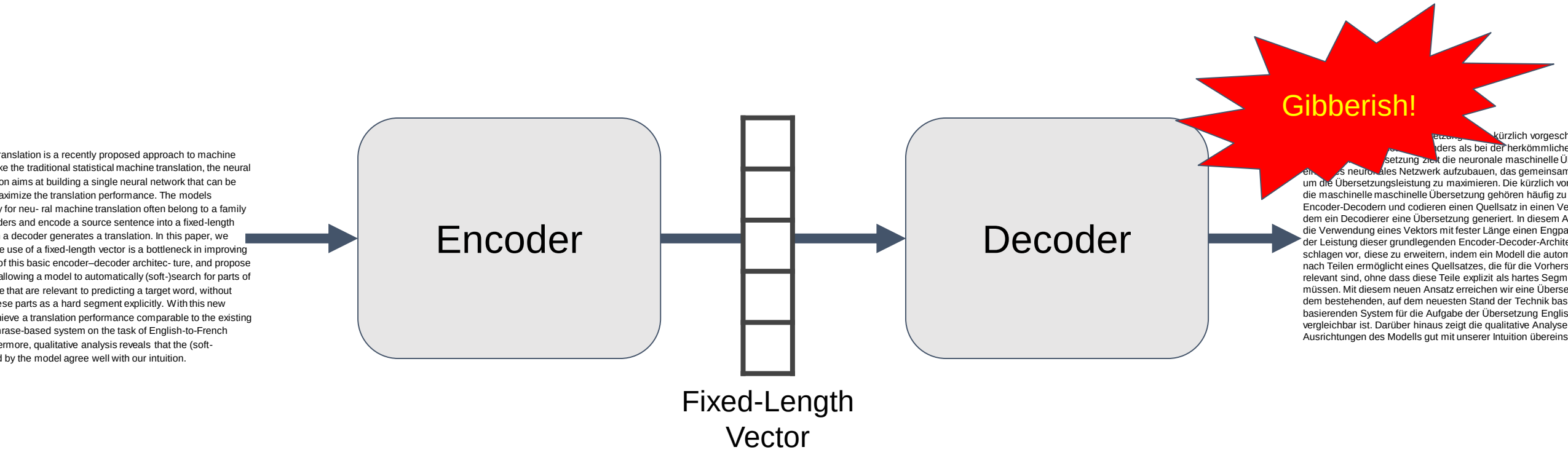
Seq2Seq



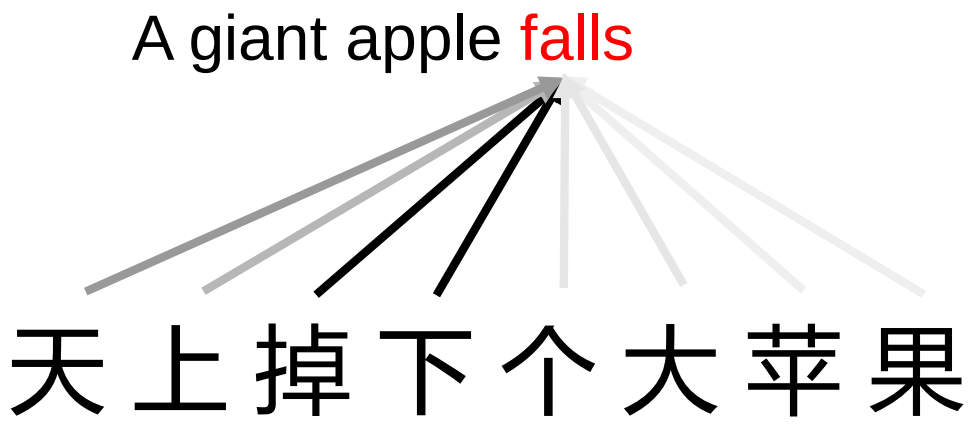
Problem with Vanilla Seq2Seq



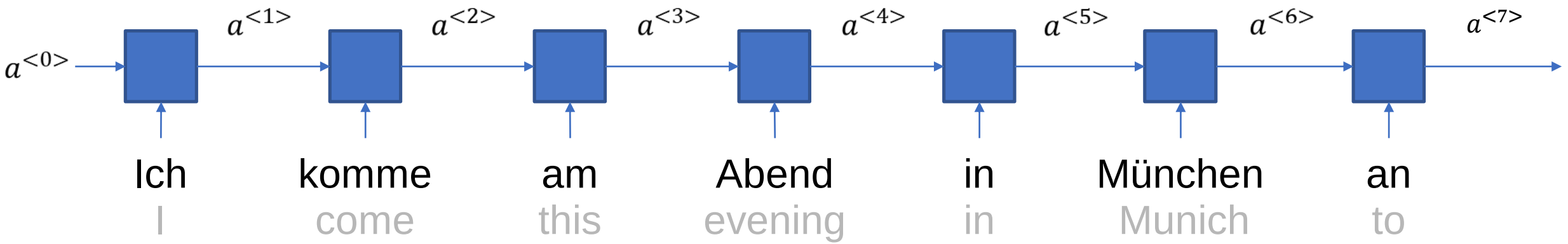
Problem with Vanilla Seq2Seq



Attention (Example)

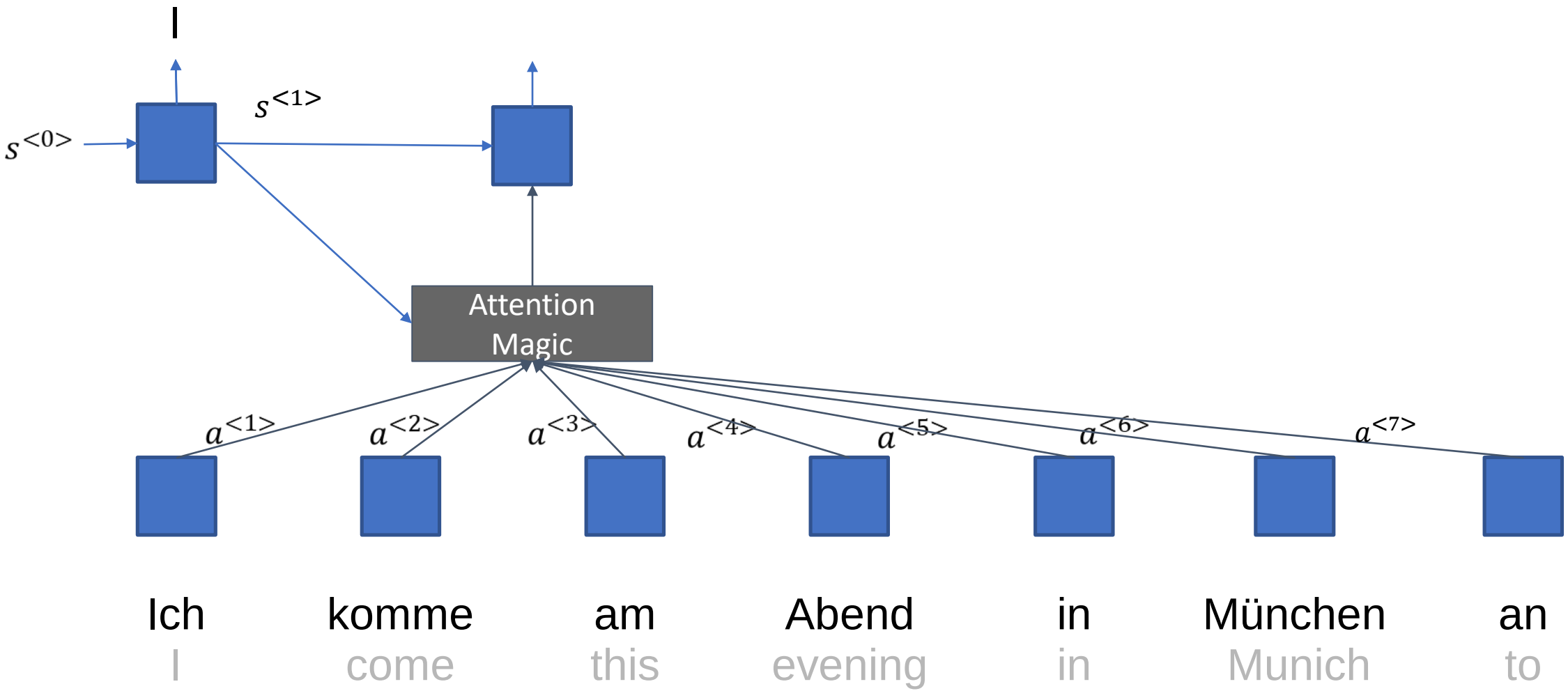


Attention (Intuition)

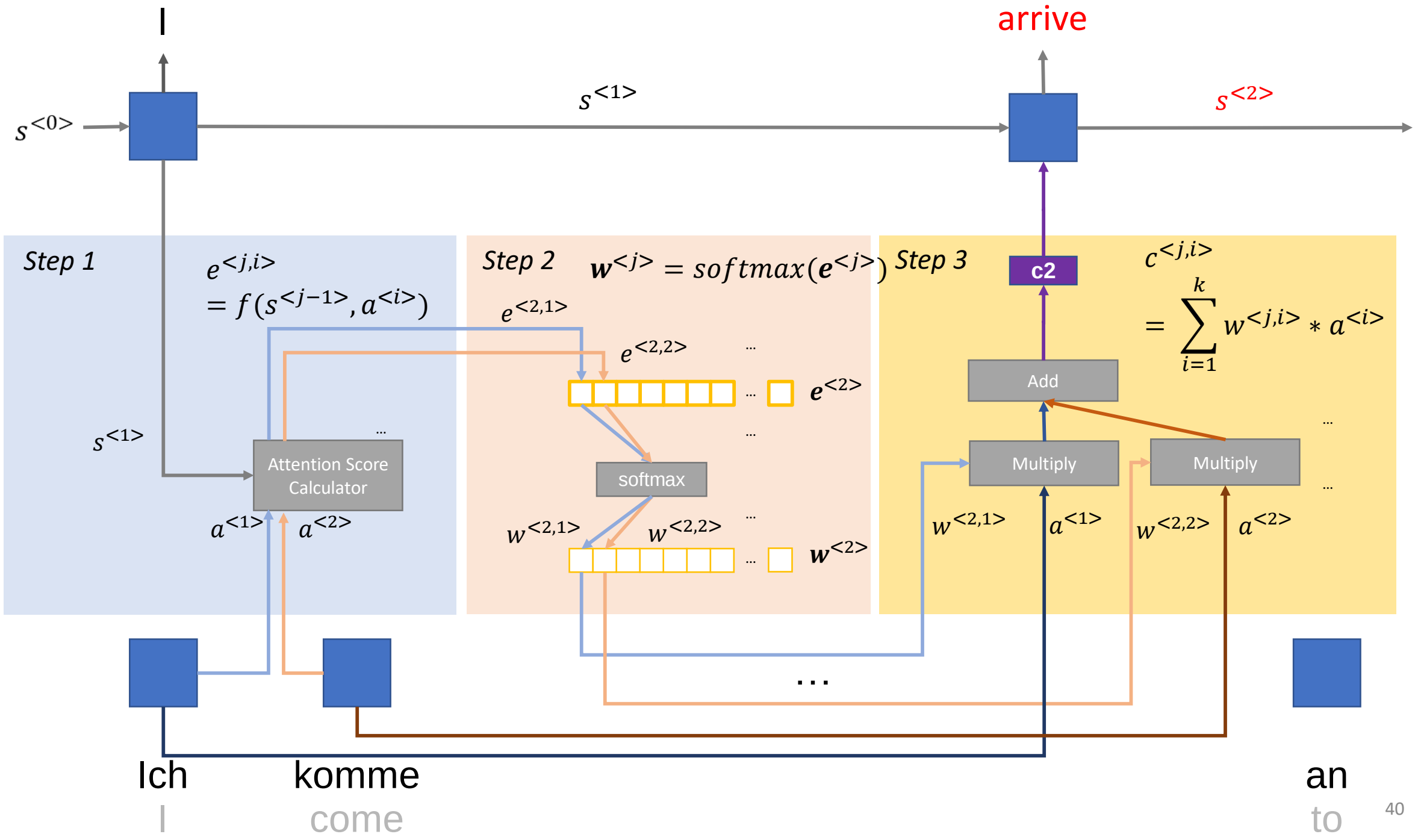


ankommen - arrive

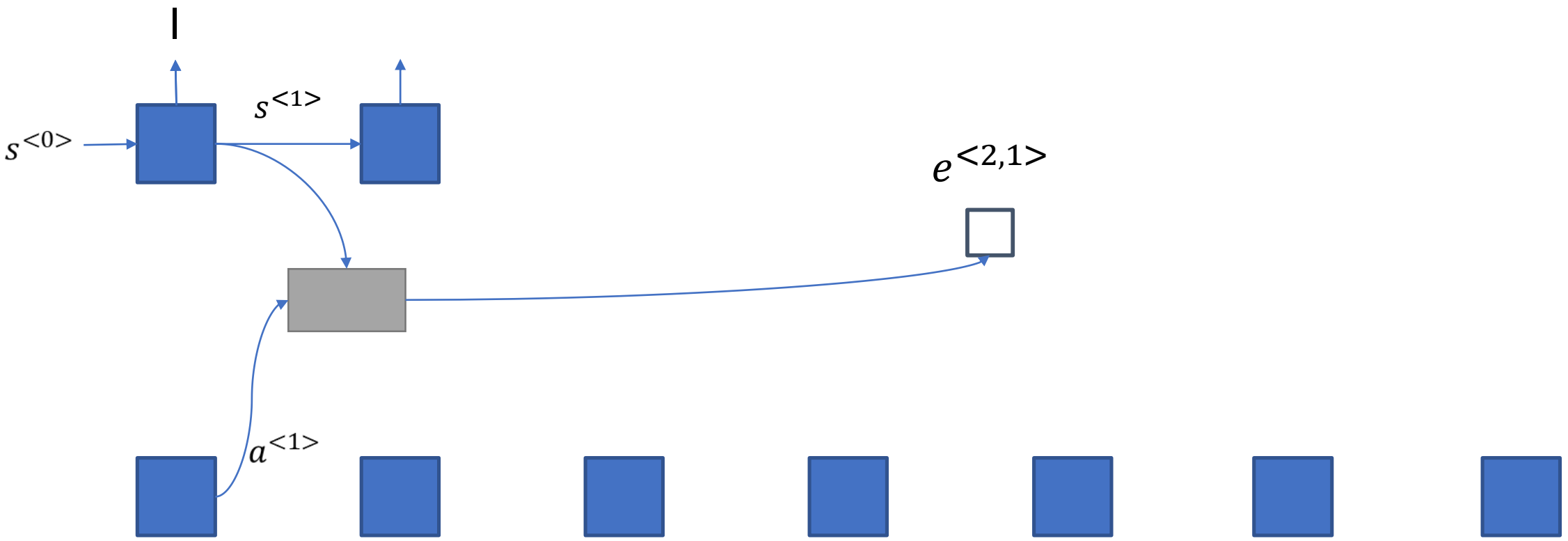
Attention (Intuition)



ankommen - arrive

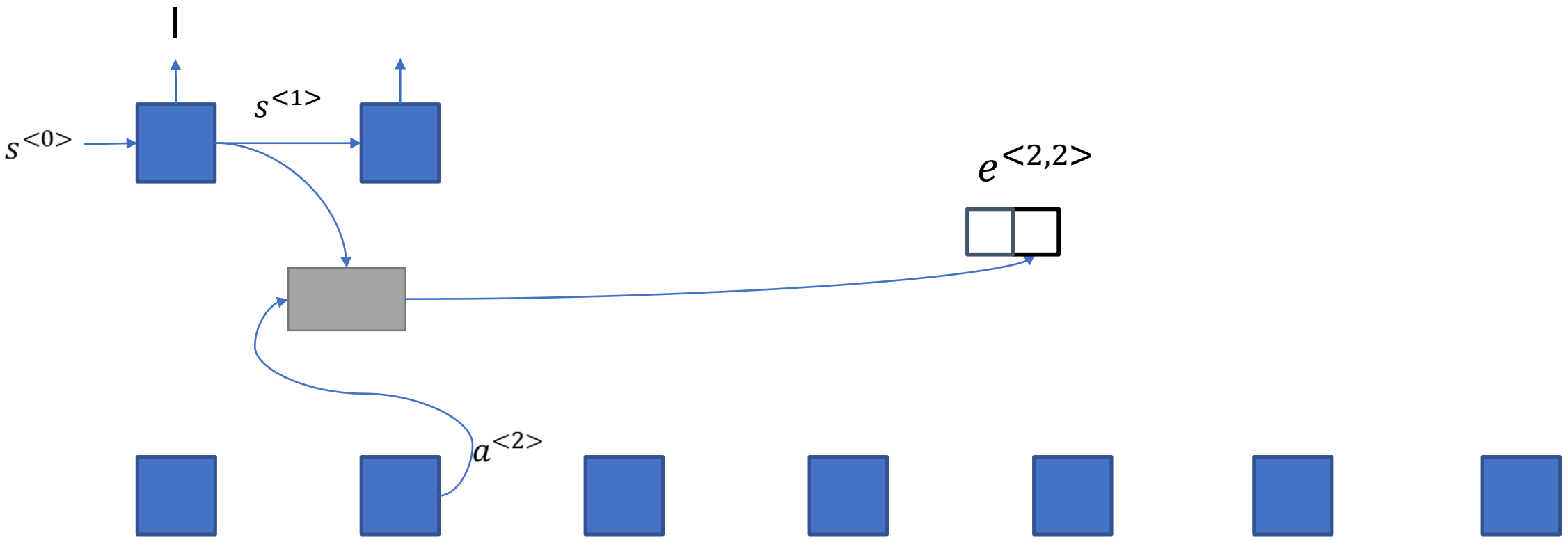

I
arrive
$s^{<0>}$
$s^{<1>}$
$s^{<2>}$
Step 1
$e^{<j,i>} = f(s^{<j-1>}, a^{<i>})$
Step 2
$w^{<j>} = softmax(e^{<j>})$
Step 3
$c^{<j,i>} = \sum_{i=1}^{k} w^{<j,i>} * a^{<i>}$
c2
Add
Multiply
Multiply
Attention Score Calculator
softmax
$e^{<2,1>}$
$e^{<2,2>}$
$e^{<2>}$
$w^{<2,1>}$
$w^{<2,2>}$
$w^{<2>}$
$a^{<1>}$
$a^{<2>}$
...
Ich
komme
an
I
come
to

Step 1



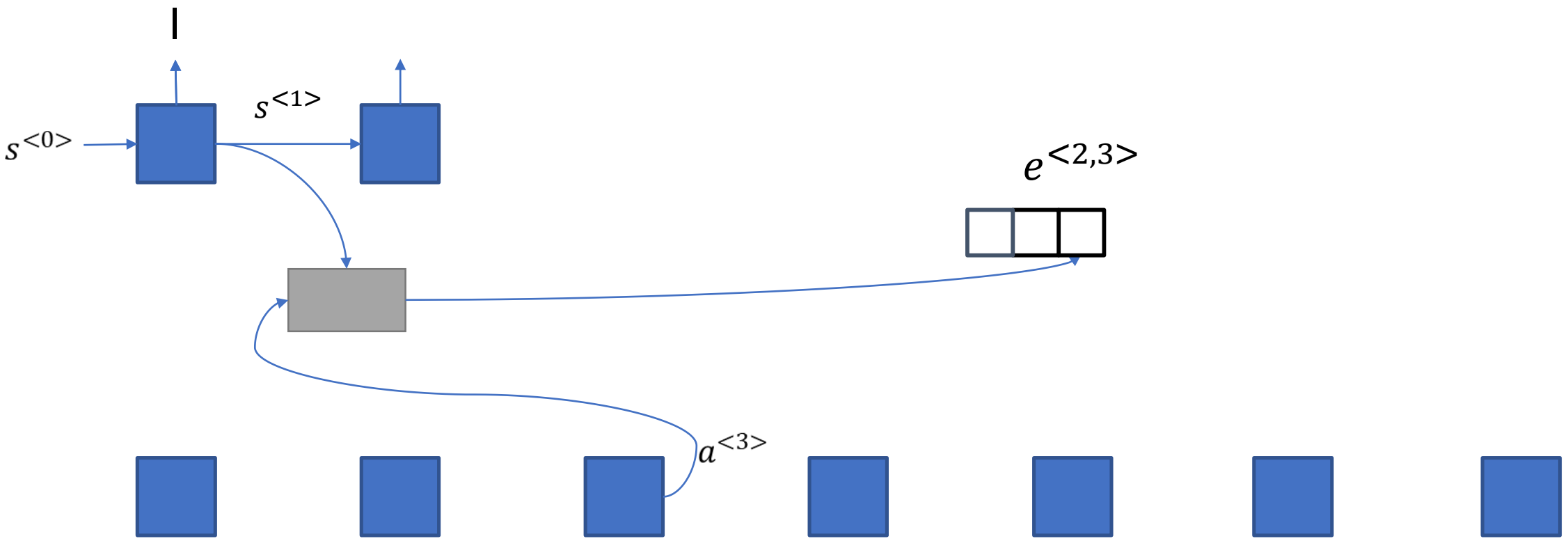
ankommen - arrive

Step 1



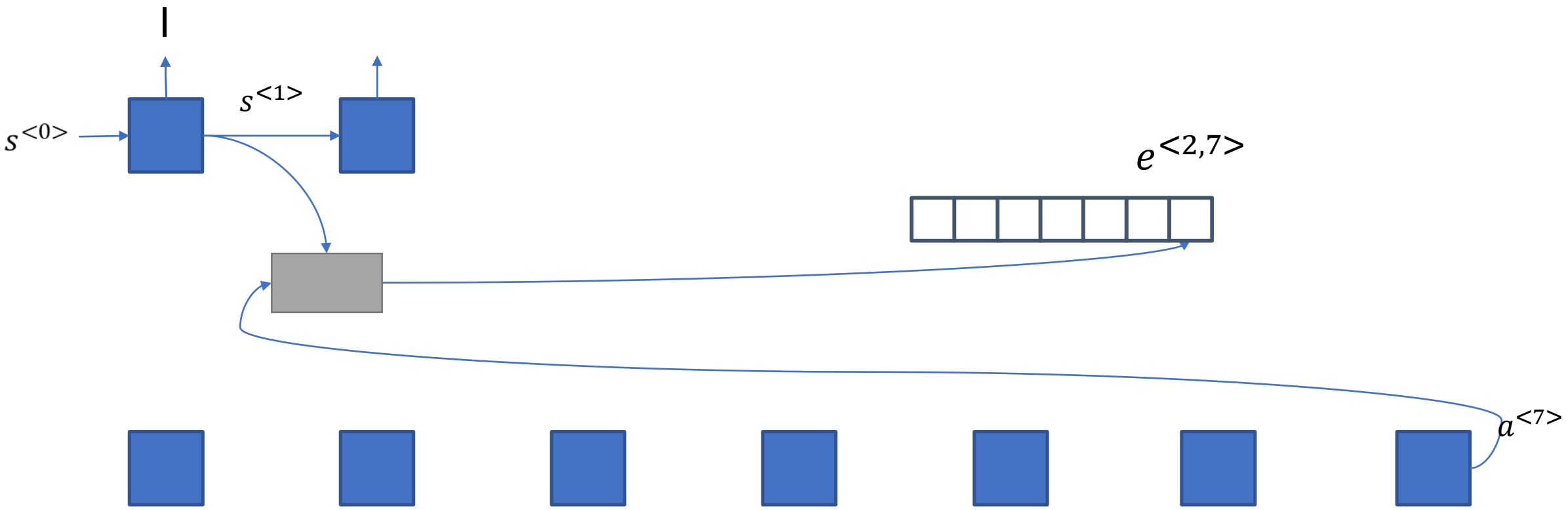
ankommen - arrive

Step 1



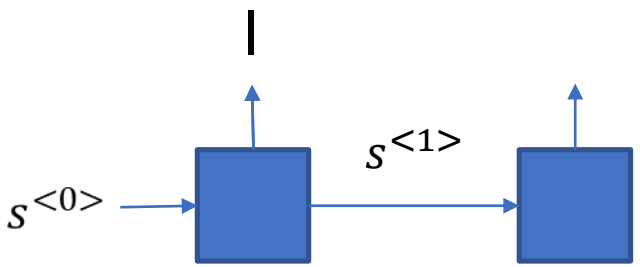
ankommen - arrive

Step 1



ankommen - arrive

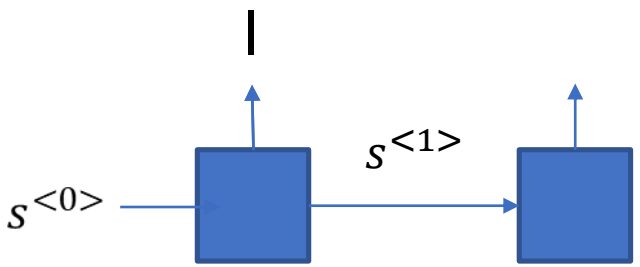
Step 2

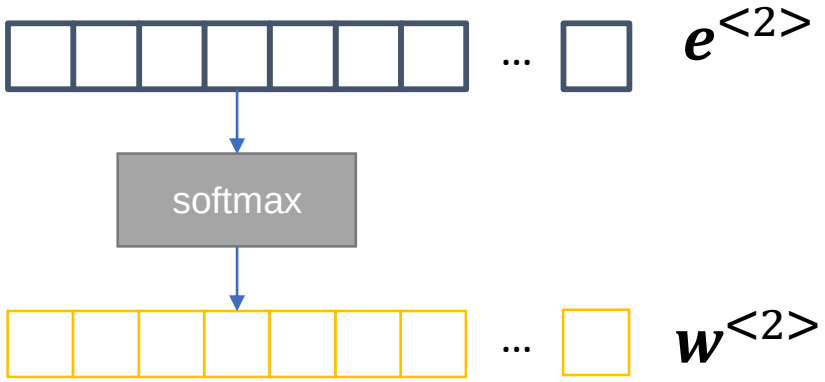


Ich	komme	am	Abend	in	München	an
I	come	this	evening	in	Munich	to

ankommen - arrive

Step 2

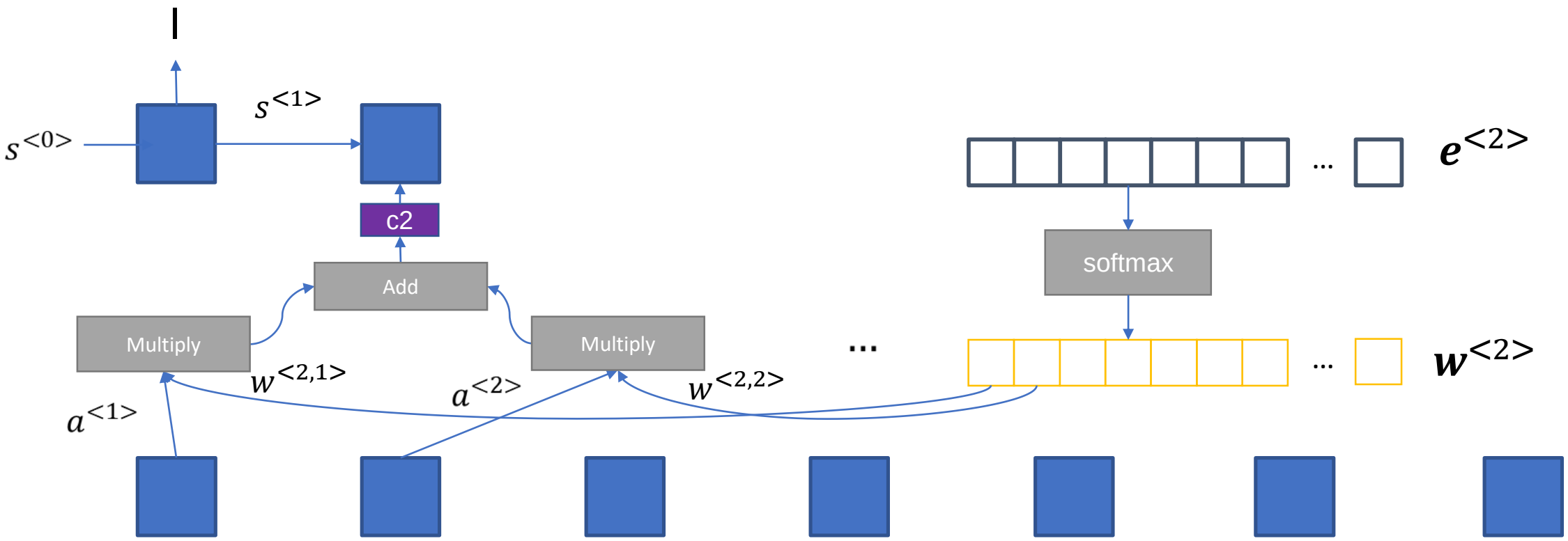




Ich	komme	am	Abend	in	München	an
I	come	this	evening	in	Munich	to

ankommen - arrive

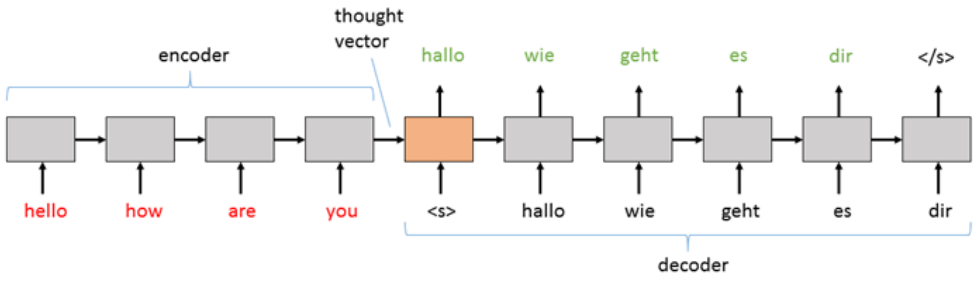
Step 3



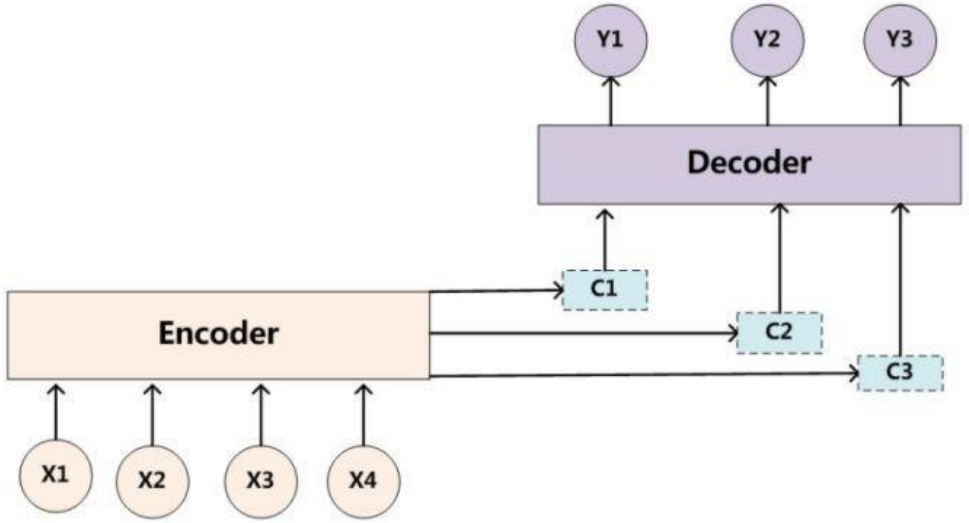
ankommen - arrive

Seq2Seq Practice Revisited

Without Attention



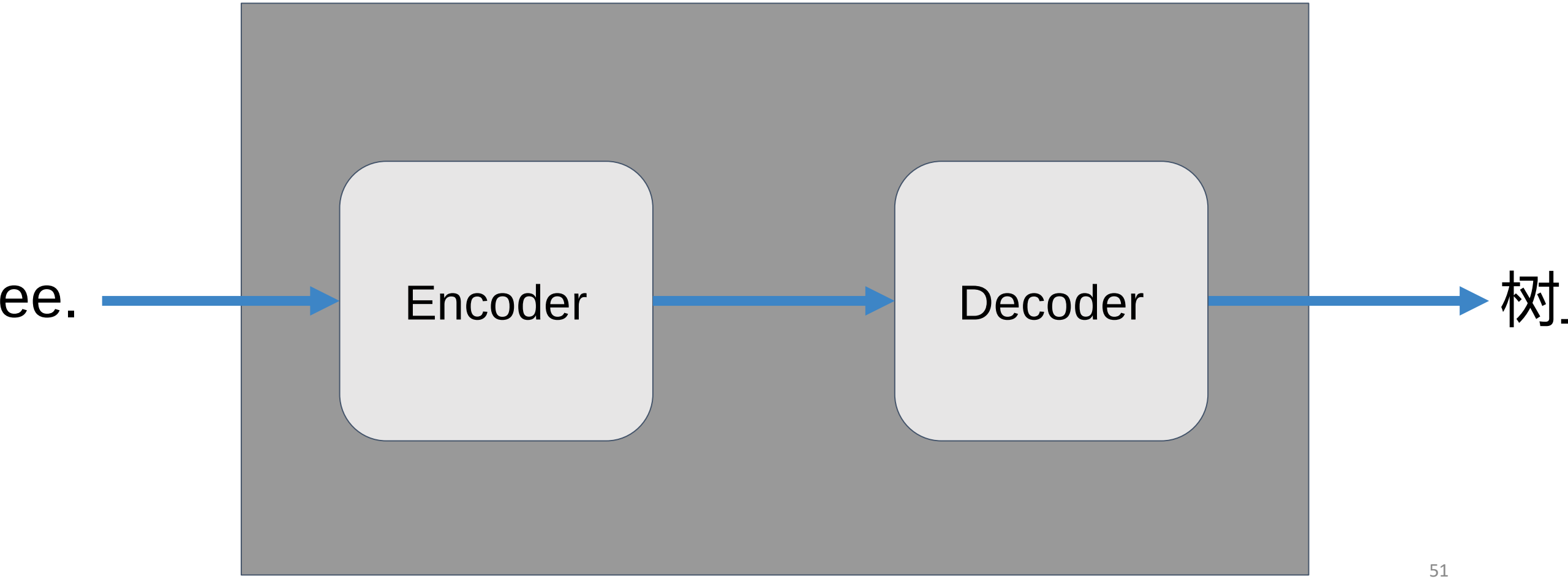
With Attention



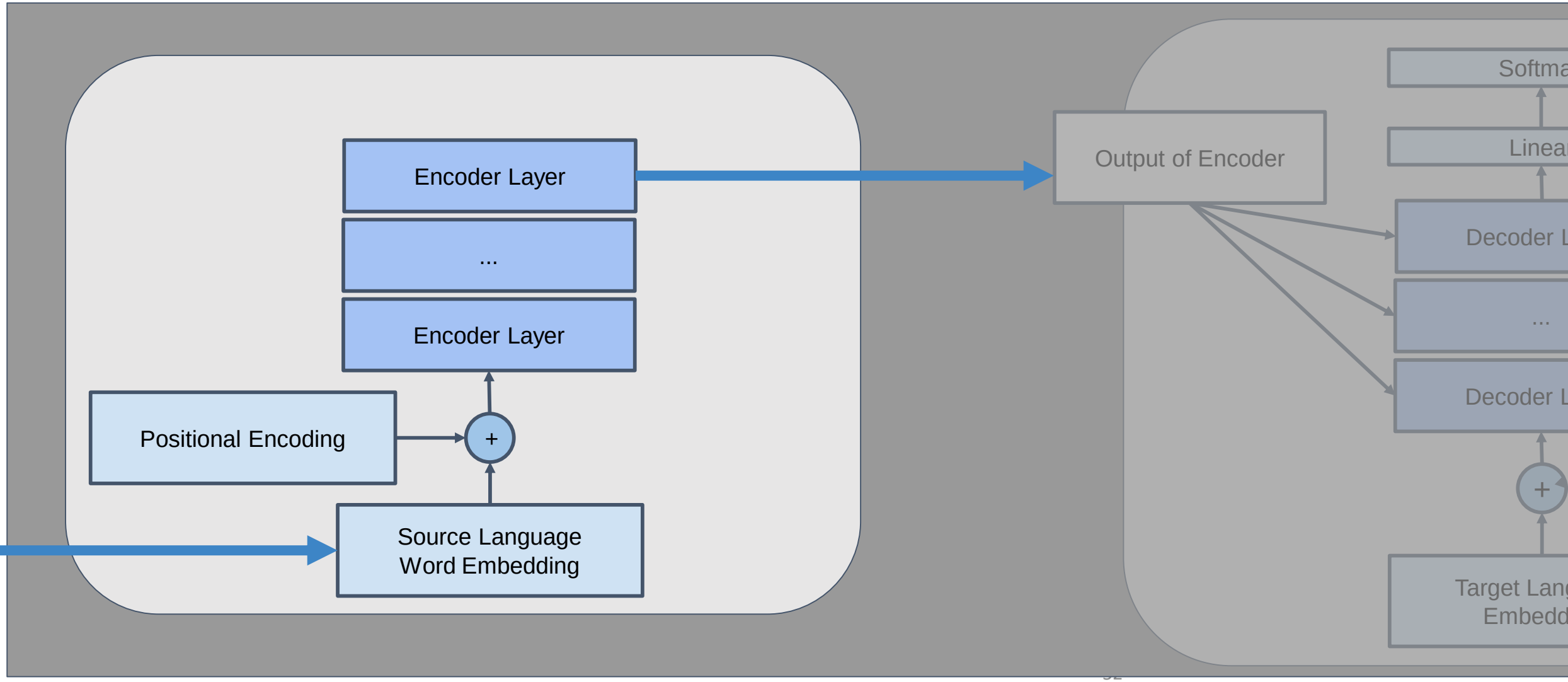
Transformer

Transformer

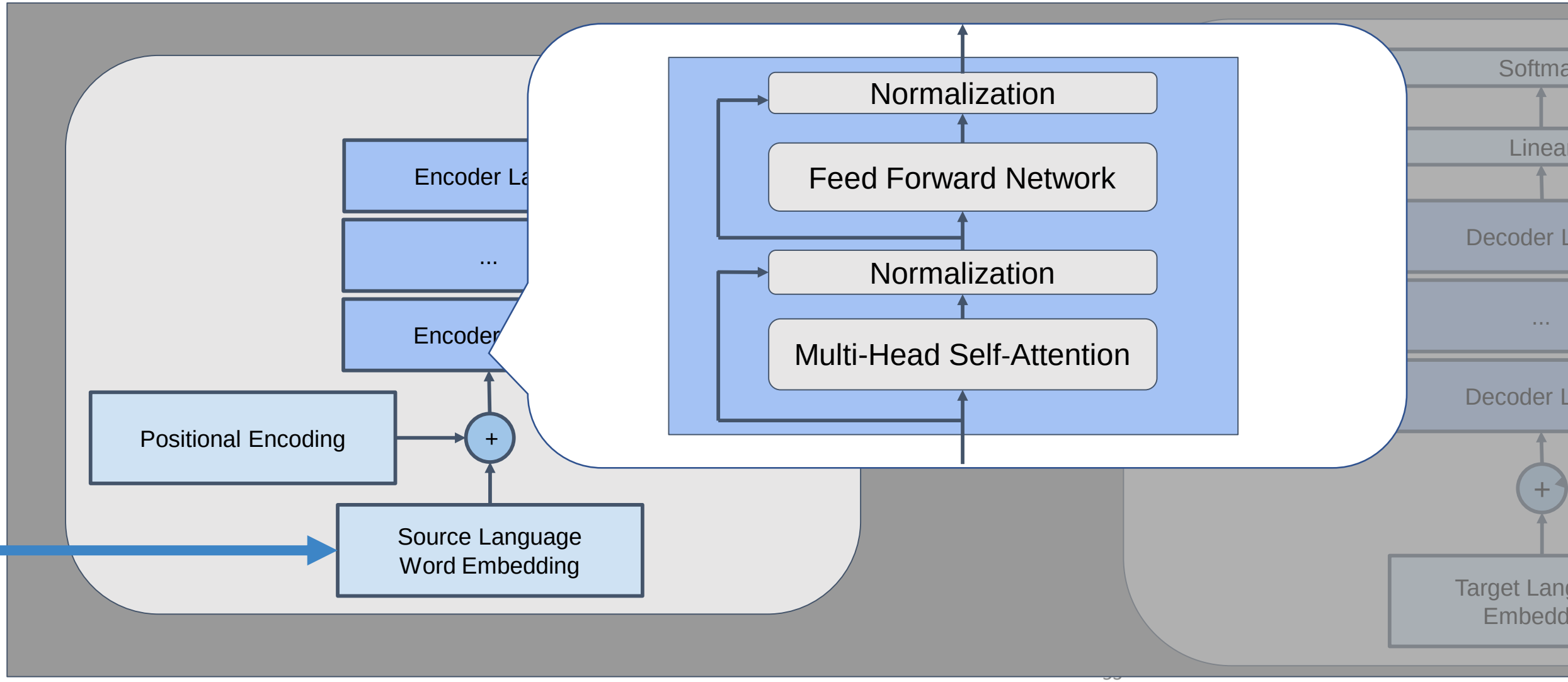
Transformer



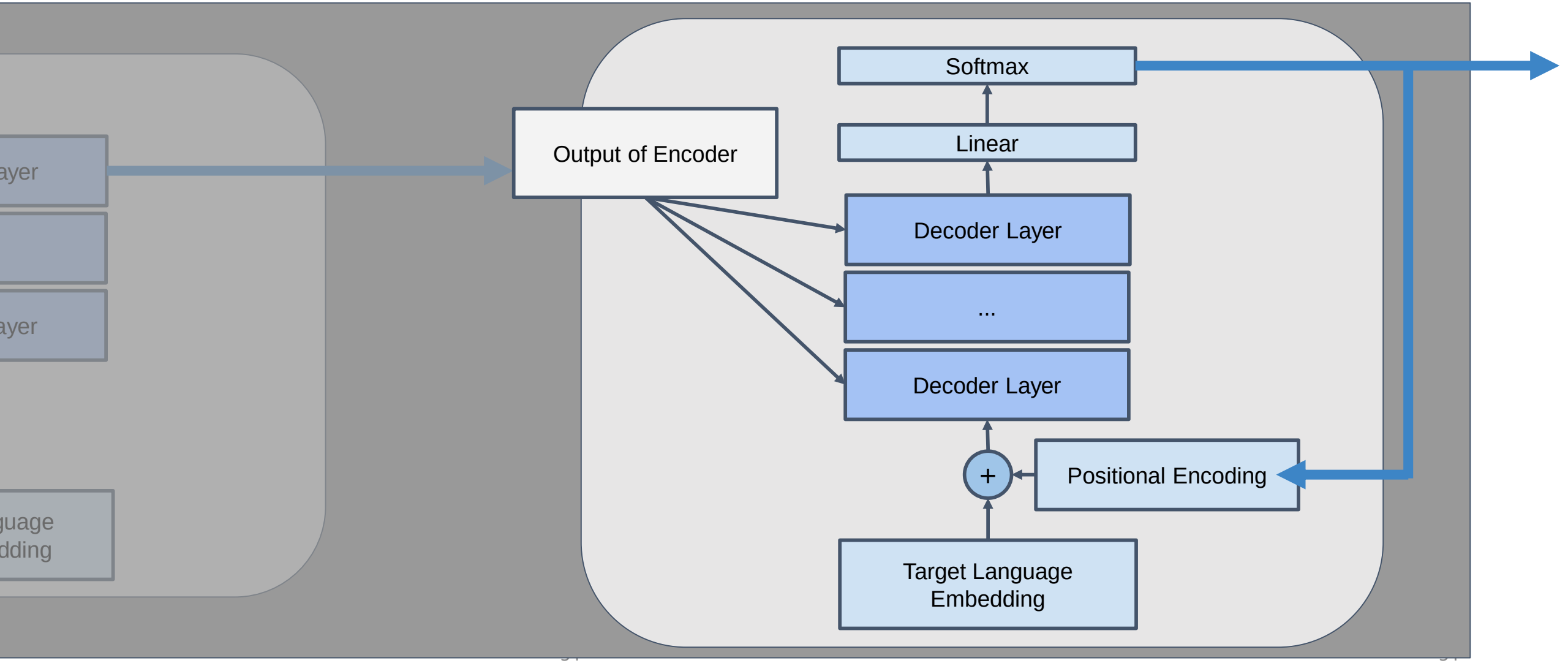
Transformer - Encoder



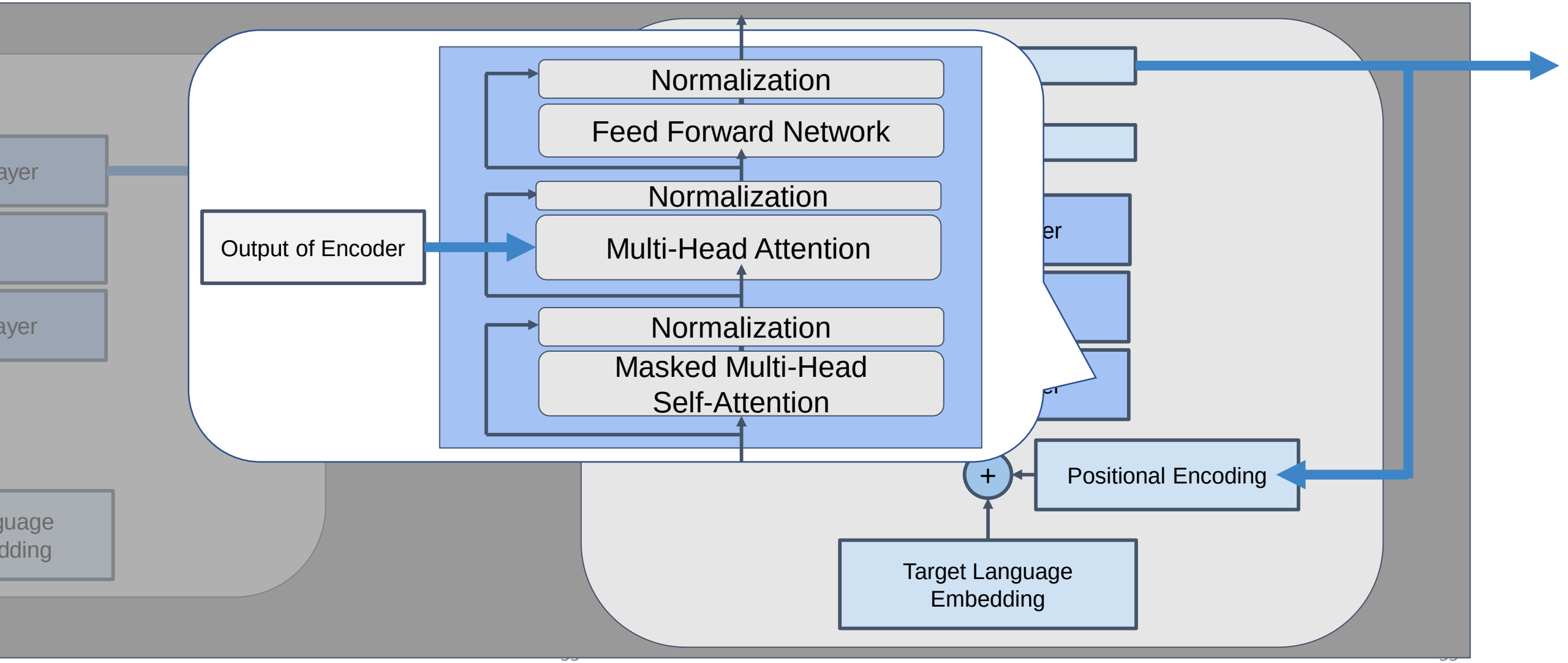
Transformer - Encoder



Transformer - Decoder



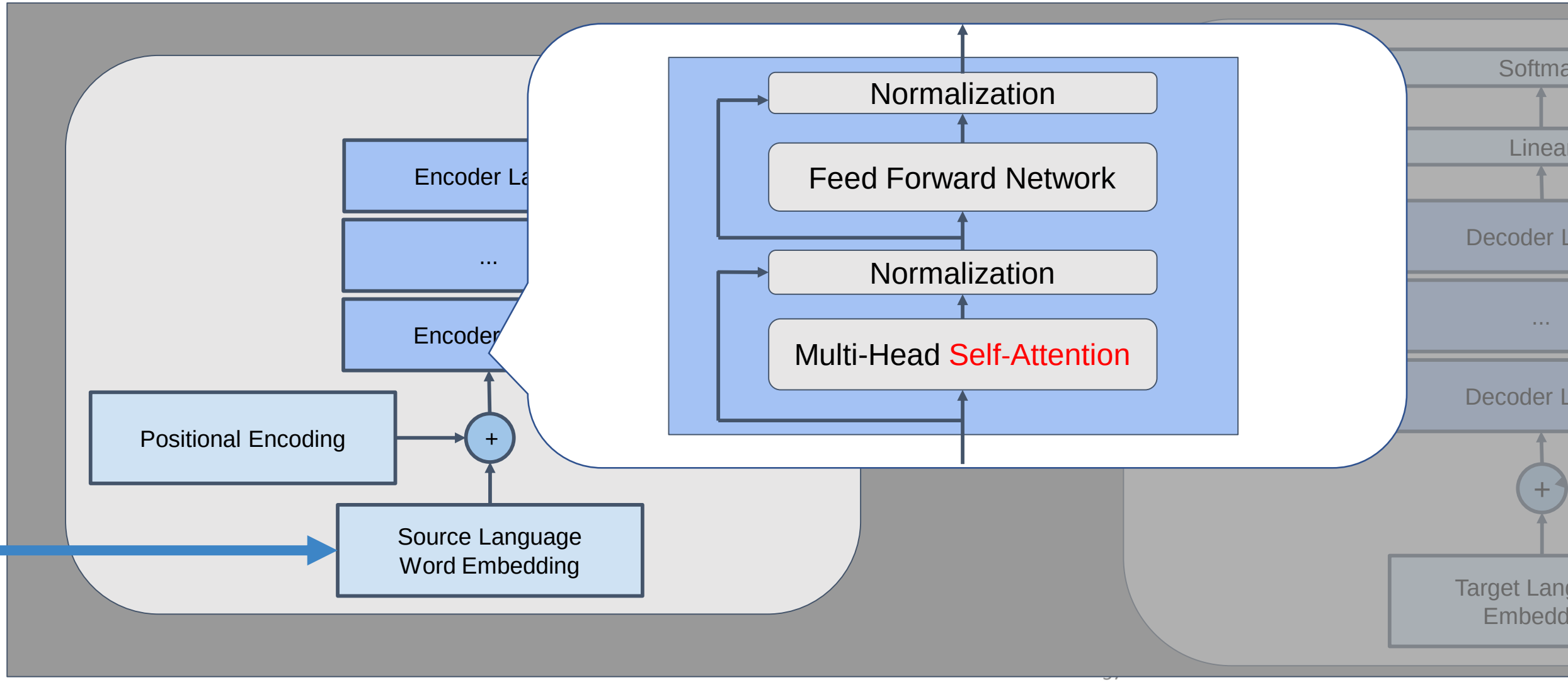
Transformer - Decoder



Transformer - Special Attention Mechanisms

Self-Attention
Multi-Head Attention
Masked Multi-Head Attention

Transformer - Encoder



Transformer - Self-Attention

dawn-to-dusk traffic jams. And some cities have transformed their streets into cycle-path freeways, making giant strides in public health and safety and the sheer livability of their neighborhoods—in the process turning the workaday bicycle into a viable form of mass transit.

If you credit the demographers, this transit trend has legs. The "Millenials," who reached adulthood around the turn of the century and now outnumber baby boomers, tend to favor cities over suburbs, and are far more willing than their parents to ride buses and subways. Part of the reason is their ease with

16

As used in line 58, "credit" most nearly means

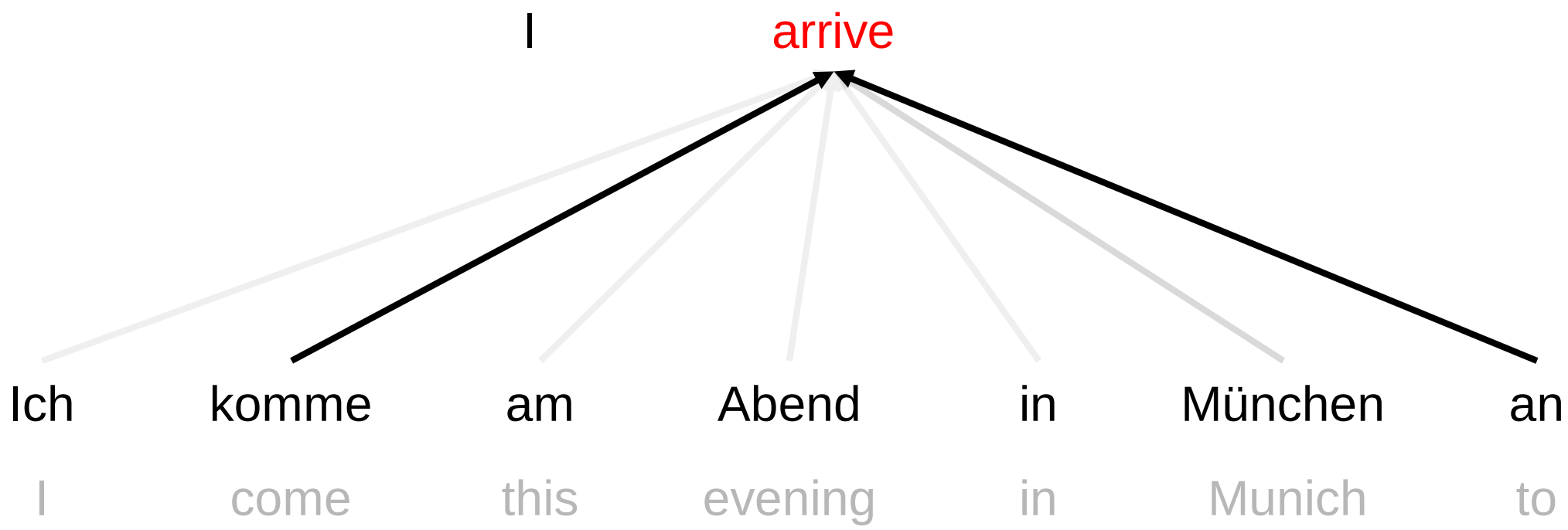
A) endow.

B) attribute.

C) believe.

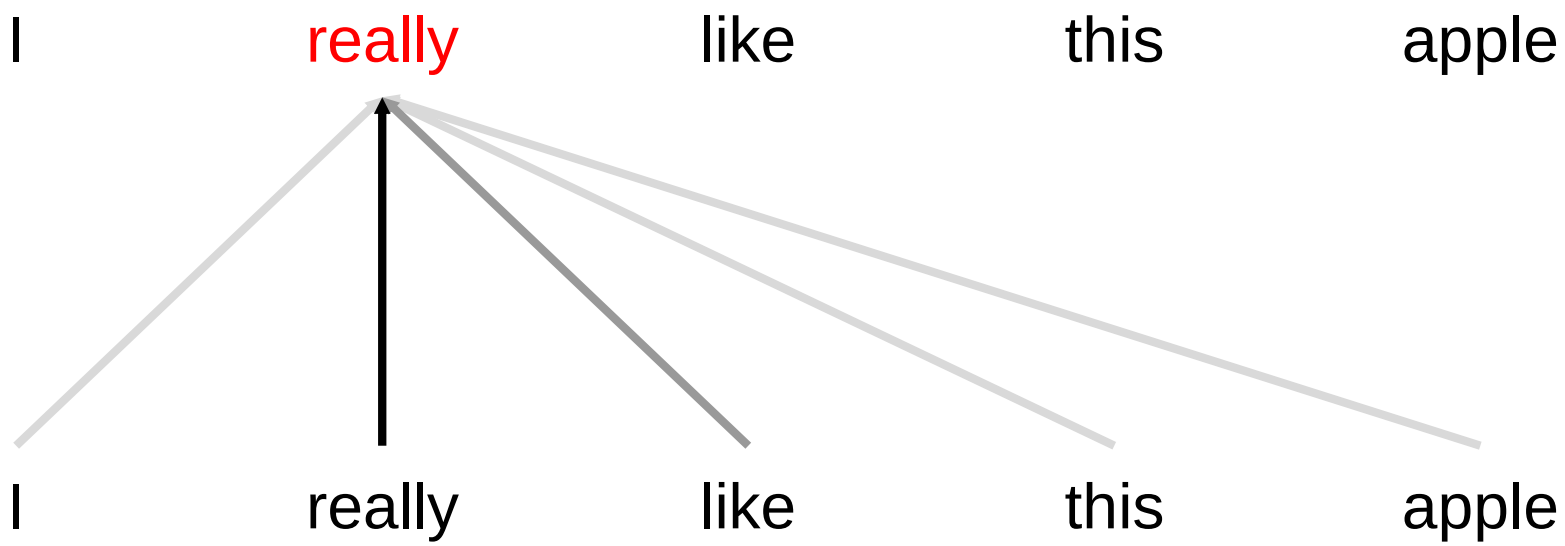
D) honor.

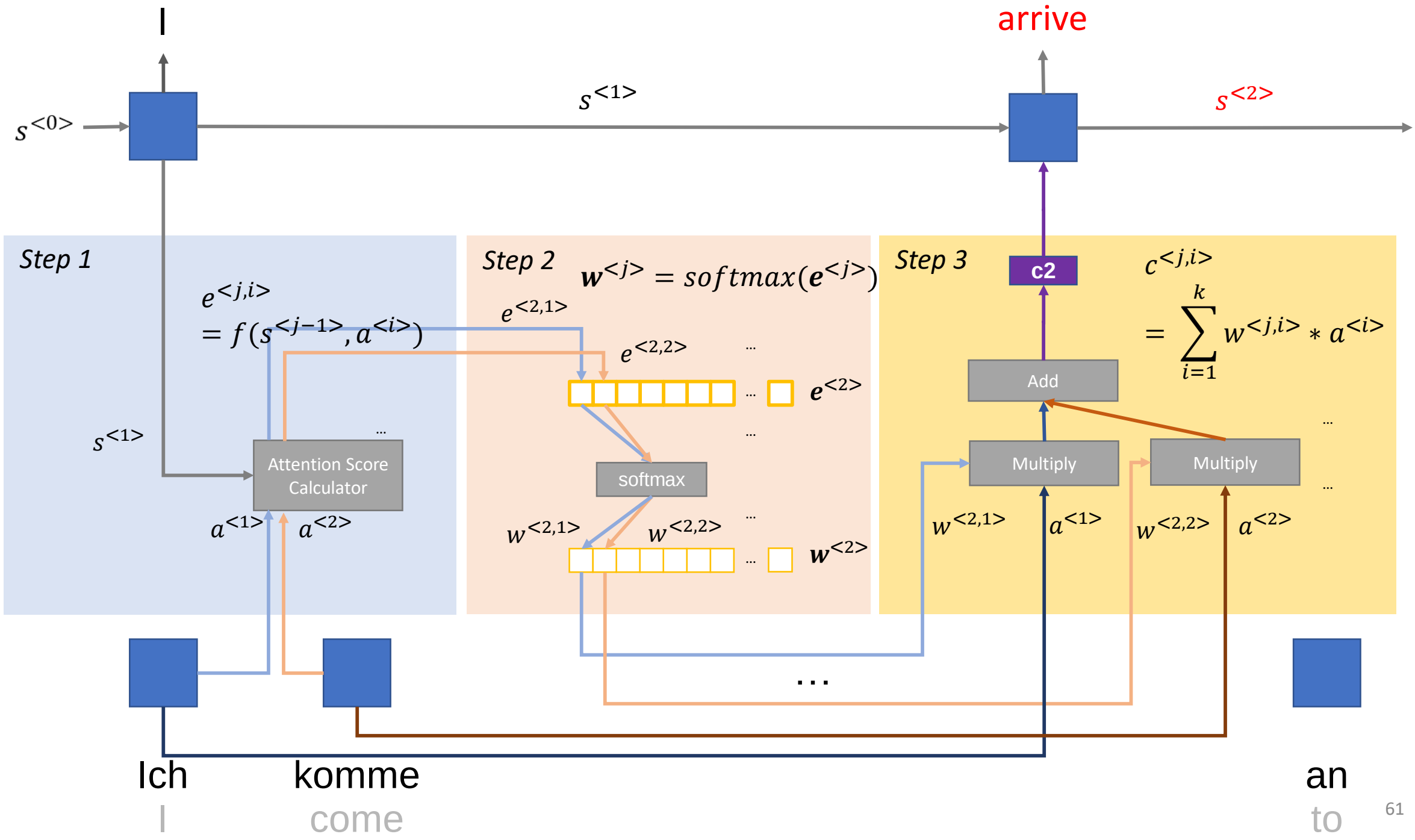
Attention

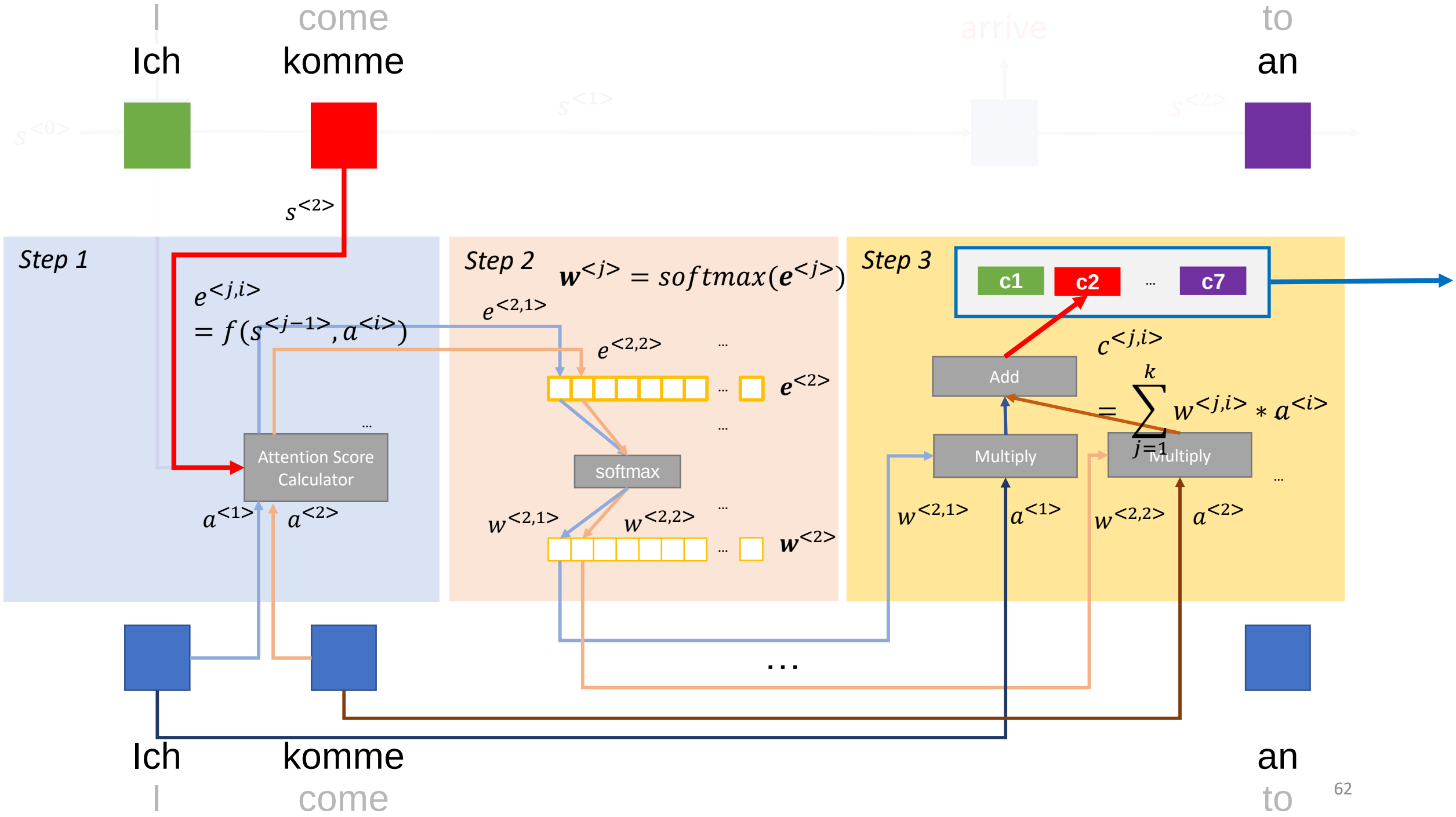


ankommen - arrive

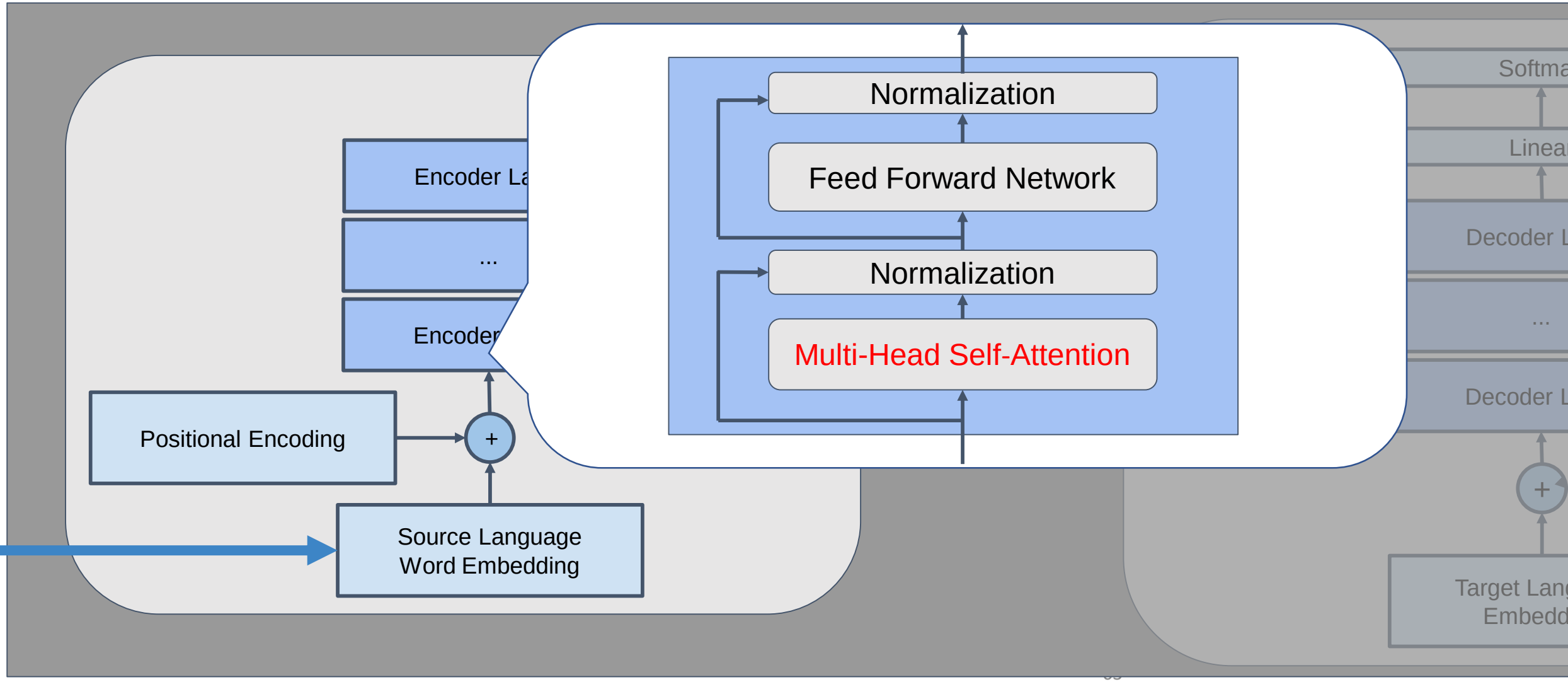
Transformer - Self-Attention



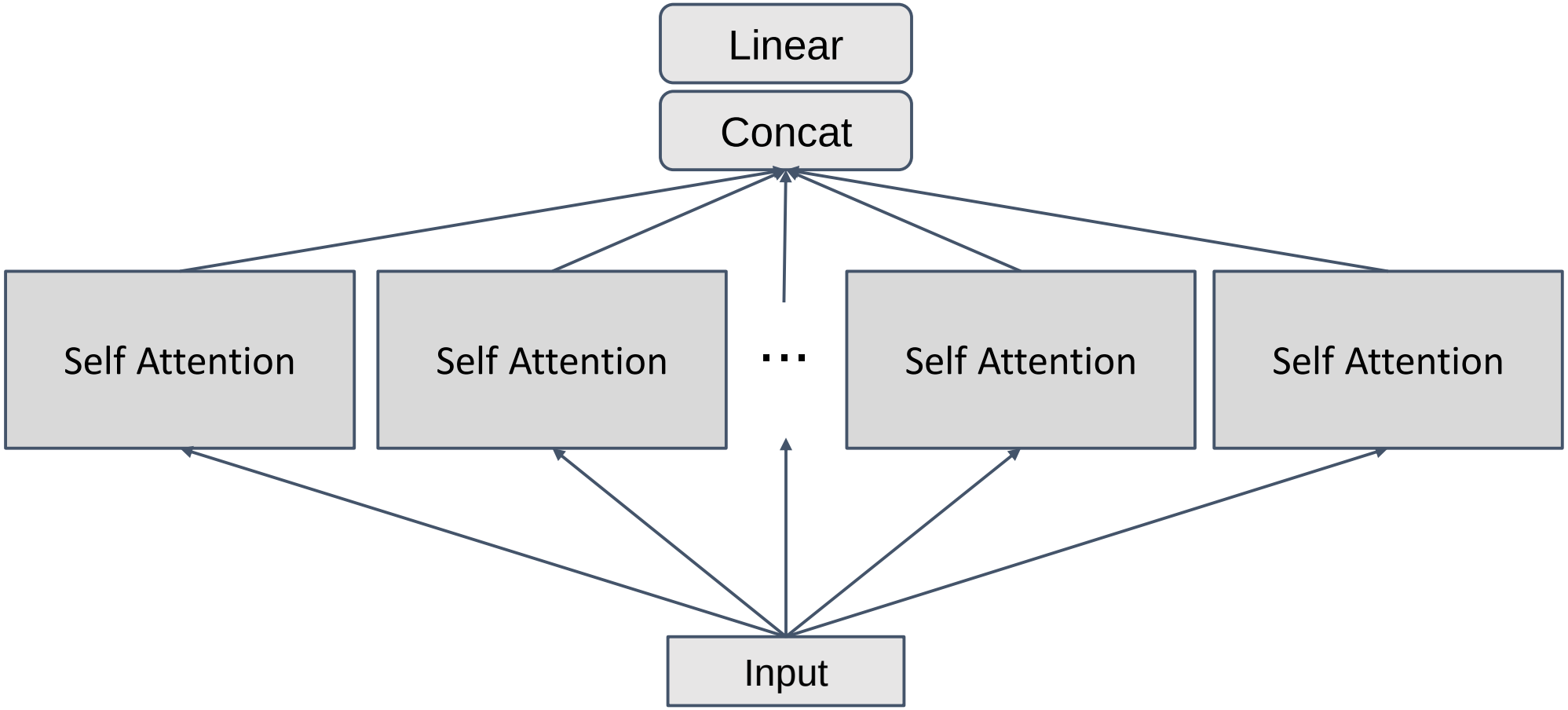

I
arrive
$s^{<0>}$
$s^{<1>}$
$s^{<2>}$
Step 1
$e^{<j,i>} = f(s^{<j-1>}, a^{<i>})$
$s^{<1>}$
Attention Score Calculator
$a^{<1>}$
$a^{<2>}$
Step 2
$w^{<j>} = softmax(e^{<j>})$
$e^{<2,1>}$
$e^{<2,2>}$
$e^{<2>}$
softmax
$w^{<2,1>}$
$w^{<2,2>}$
$w^{<2>}$
Step 3
c2
$c^{<j,i>} = \sum_{i=1}^{k} w^{<j,i>} * a^{<i>}$
Add
Multiply
Multiply
$w^{<2,1>}$
$a^{<1>}$
$w^{<2,2>}$
$a^{<2>}$
Ich
komme
an
I
come
to


Ich
komme
an
Step 1
Step 2
Step 3
$s^{<2>}$
$e^{<j,i>} = f(s^{<j-1>}, a^{<i>})$
Attention Score Calculator
$a^{<1>}$
$a^{<2>}$
$w^{<j>} = softmax(e^{<j>})$
$e^{<2,1>}$
$e^{<2,2>}$
$e^{<2>}$
softmax
$w^{<2,1>}$
$w^{<2,2>}$
$w^{<2>}$
c1
c2
c7
$c^{<j,i>}$
Add
$= \sum_{j=1}^{k} w^{<j,i>} * a^{<i>}$
Multiply
Multiply
$w^{<2,1>}$
$a^{<1>}$
$w^{<2,2>}$
$a^{<2>}$
Ich
komme
an

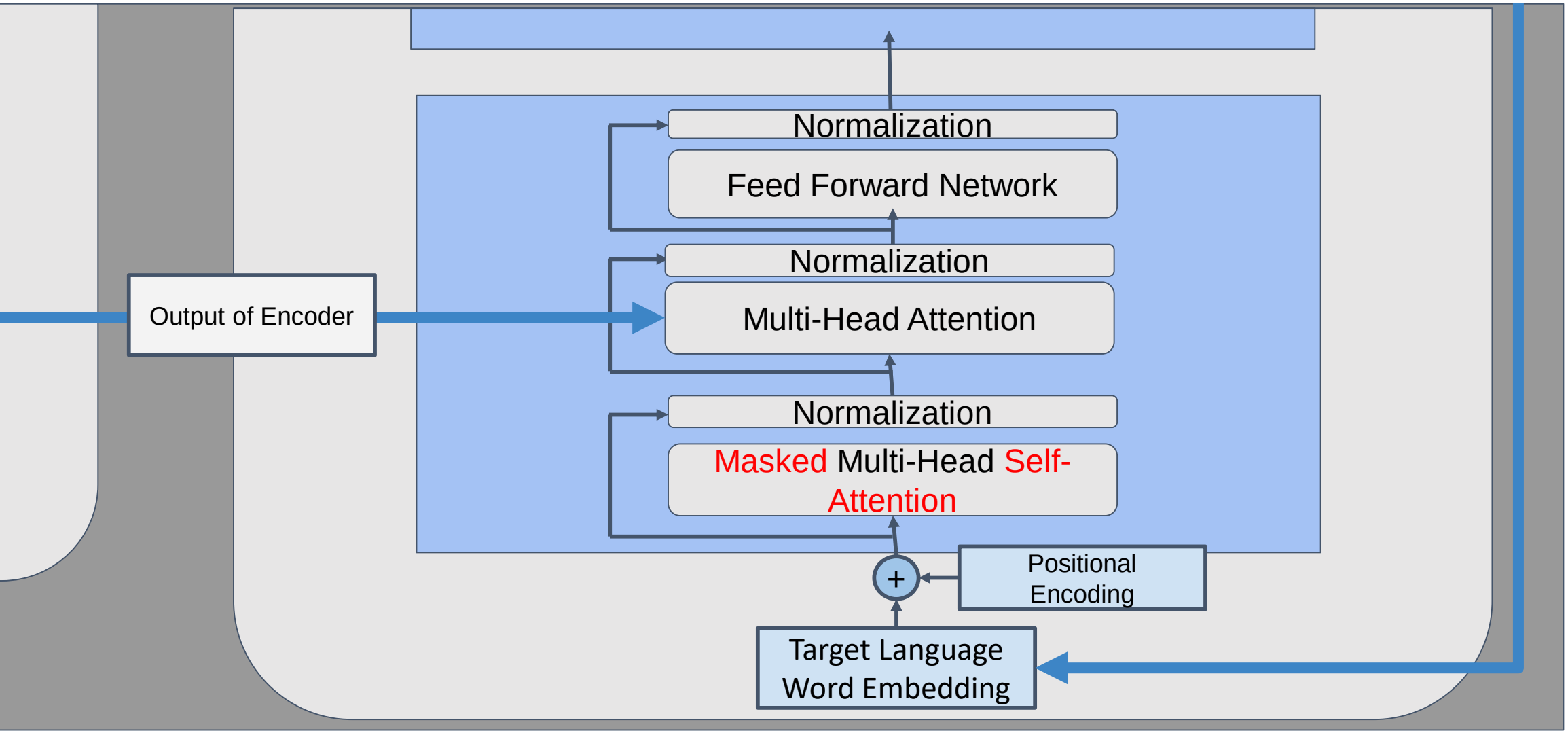
Transformer - Encoder



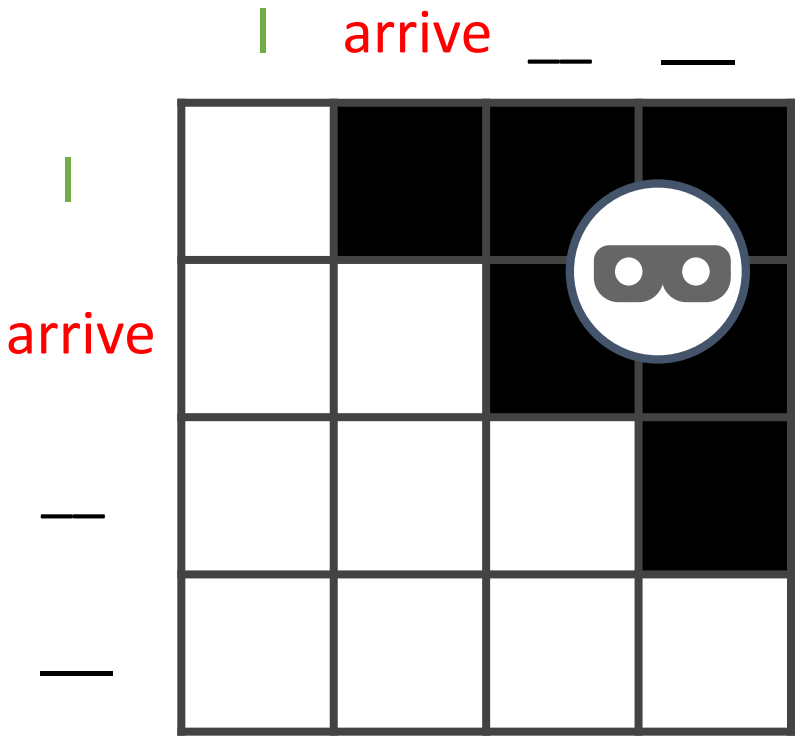
Transformer - Multi-Head Attention

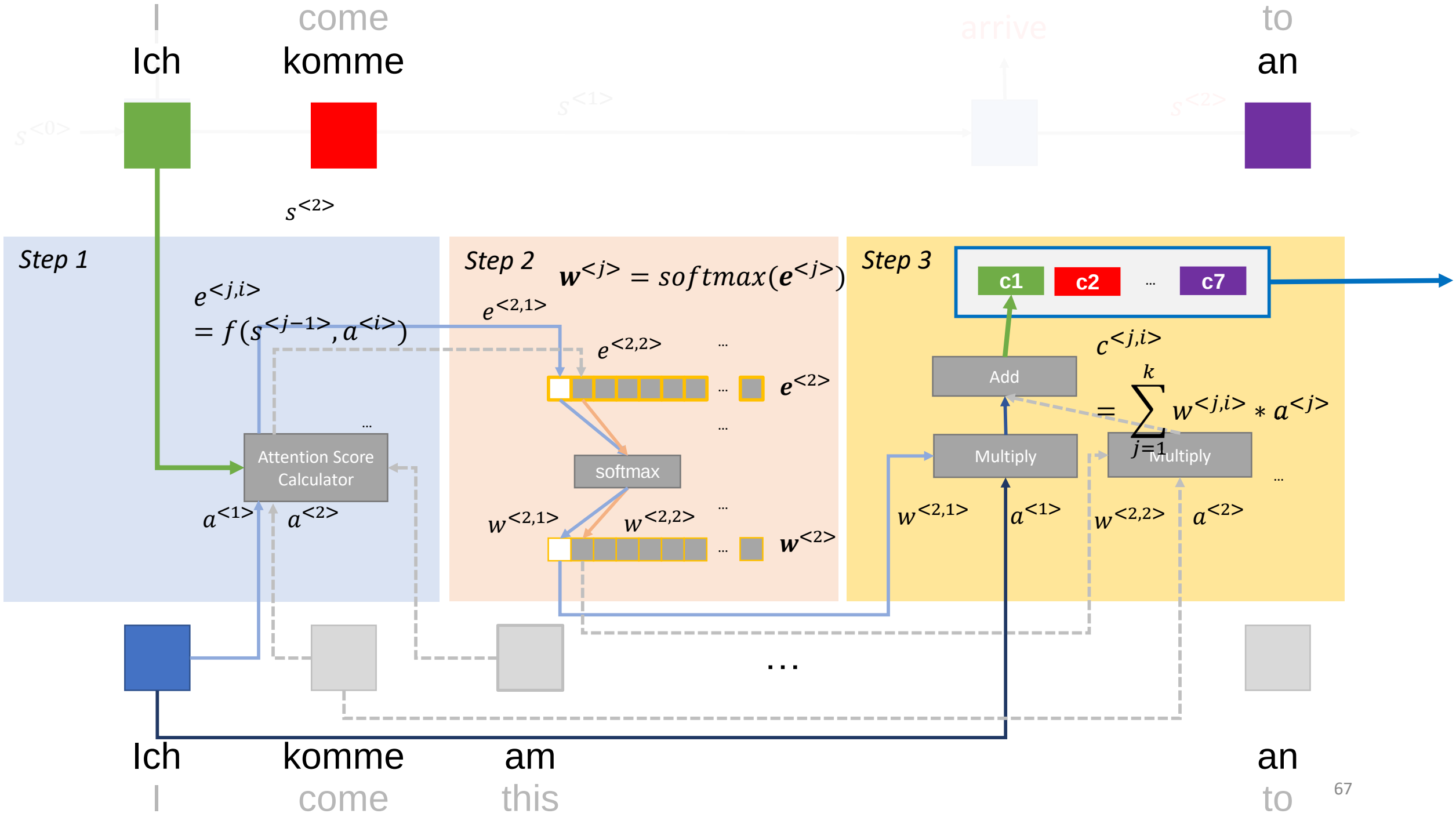


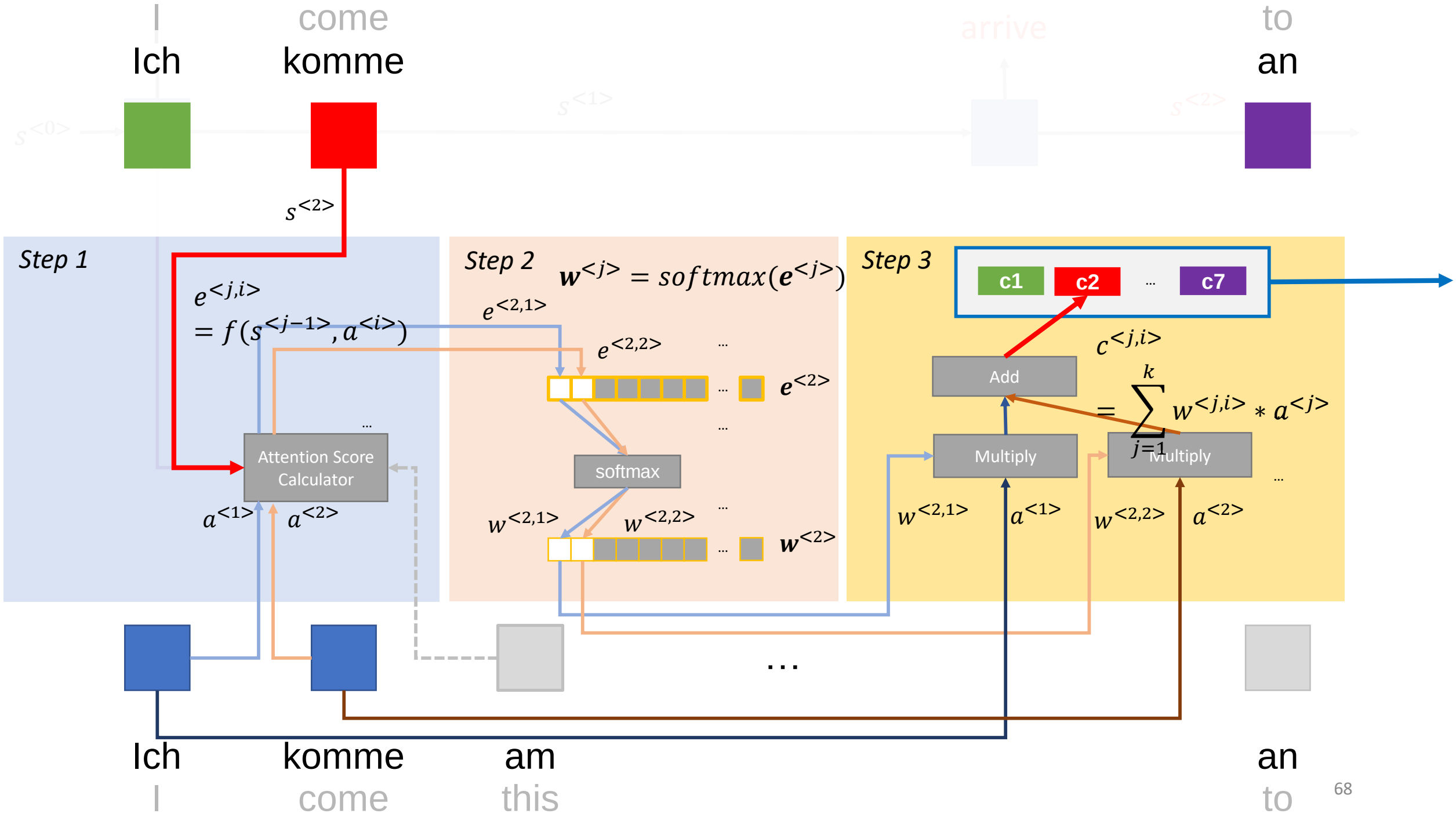
Transformer - Decoder - Decoder Layer



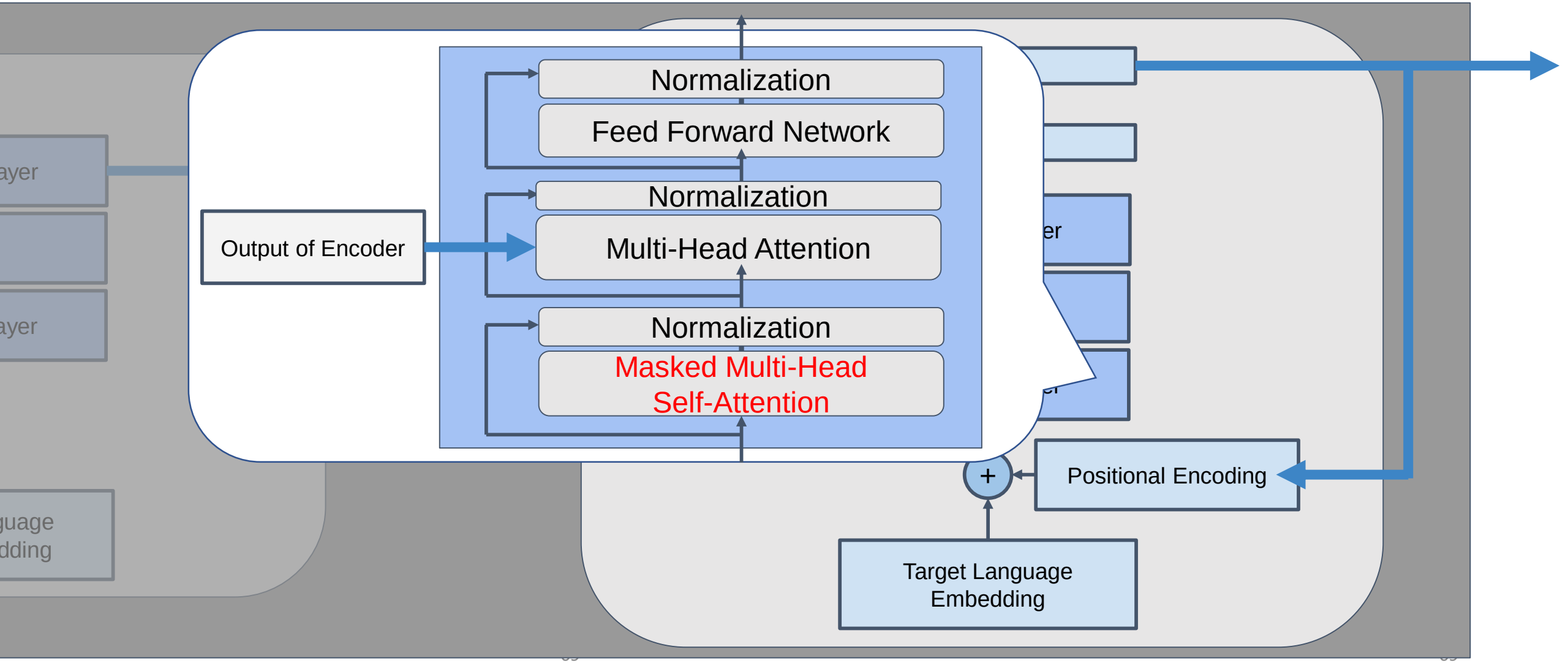
Masked Self-Attention



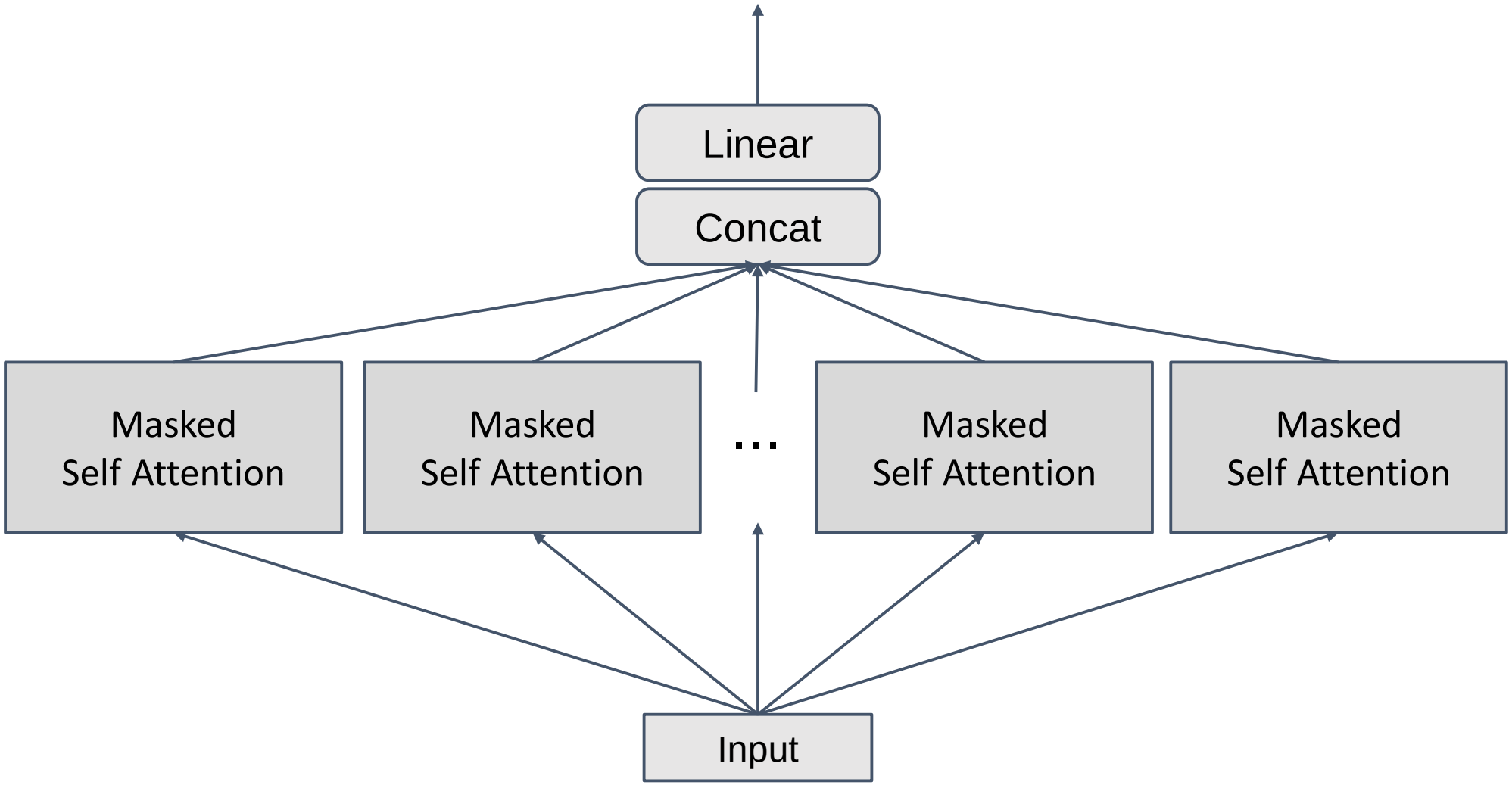

Ich
komme
an
$s^{<2>}$
Step 1
$e^{<j,i>} = f(s^{<j-1>}, a^{<i>})$
Attention Score Calculator
$a^{<1>}$
$a^{<2>}$
Step 2
$w^{<j>} = softmax(e^{<j>})$
$e^{<2,1>}$
$e^{<2,2>}$
$e^{<2>}$
softmax
$w^{<2,1>}$
$w^{<2,2>}$
$w^{<2>}$
Step 3
c1
c2
c7
$c^{<j,i>} = \sum_{j=1}^{k} w^{<j,i>} * a^{<j>}$
Add
Multiply
Multiply
$w^{<2,1>}$
$a^{<1>}$
$w^{<2,2>}$
$a^{<2>}$
Ich
komme
am
an


Ich
komme
an
$s^{<2>}$
Step 1
$e^{<j,i>} = f(s^{<j-1>}, a^{<i>})$
Attention Score Calculator
$a^{<1>}$
$a^{<2>}$
Step 2
$w^{<j>} = softmax(e^{<j>})$
$e^{<2,1>}$
$e^{<2,2>}$
$e^{<2>}$
softmax
$w^{<2,1>}$
$w^{<2,2>}$
$w^{<2>}$
Step 3
c1
c2
c7
$c^{<j,i>}$
Add
$= \sum_{j=1}^{k} w^{<j,i>} * a^{<j>}$
Multiply
Multiply
$w^{<2,1>}$
$a^{<1>}$
$w^{<2,2>}$
$a^{<2>}$
Ich
komme
am
an

Transformer - Decoder



Transformer - Masked M.H. Self Attention



Seq2Seq

How are you → Translator → Wie geht’s

 → Captioner → A cat laughts

Experiment Analysis

Experiment Analysis

Qualitative

- Visualization:
  - Hypertools
  - MatplotLib
  - BertViz
- Explain how your model differs from others
- Explain if your experiments verify your hypothesis
- Explain the numbers such as analyze why some parameters are chosen and why the outputs may have certain values

Quantitative

- Compute a score with a set of suitable metrics:
  - BLEU for machine translation
  - ROUGE for summarization
- Compare your results with SOTAs (using a table)
- Present experiment-specific statistics using figures such as charts
- Be as specific with numbers as possible and use them to back up your verification of hypothesis