Attention Models

Course 4 of the Natural Language Processing Specialization

deeplearning.ai

Outline

- Introduction to Neural Machine Translation
- Seq2Seq model and its shortcomings
- Solution for the information bottleneck

Neural Machine Translation

Neural Machine Translation

How are you today? → Wie geht es Ihnen heute?

Seq2Seq model

Seq2Seq model

- Introduced by Google in 2014

Seq2Seq model

- Introduced by Google in 2014
- Maps variable-length sequences to fixed-length memory

Seq2Seq model

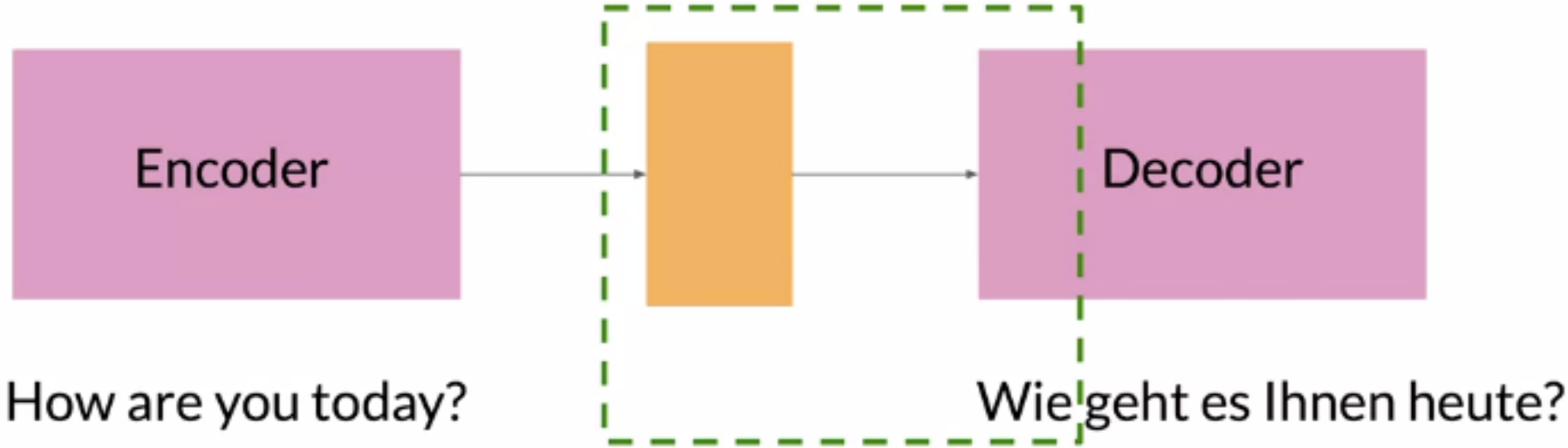
- Introduced by Google in 2014
- Maps variable-length sequences to fixed-length memory
- LSTMs and GRUs are typically used to overcome the vanishing gradient problem

Seq2Seq model

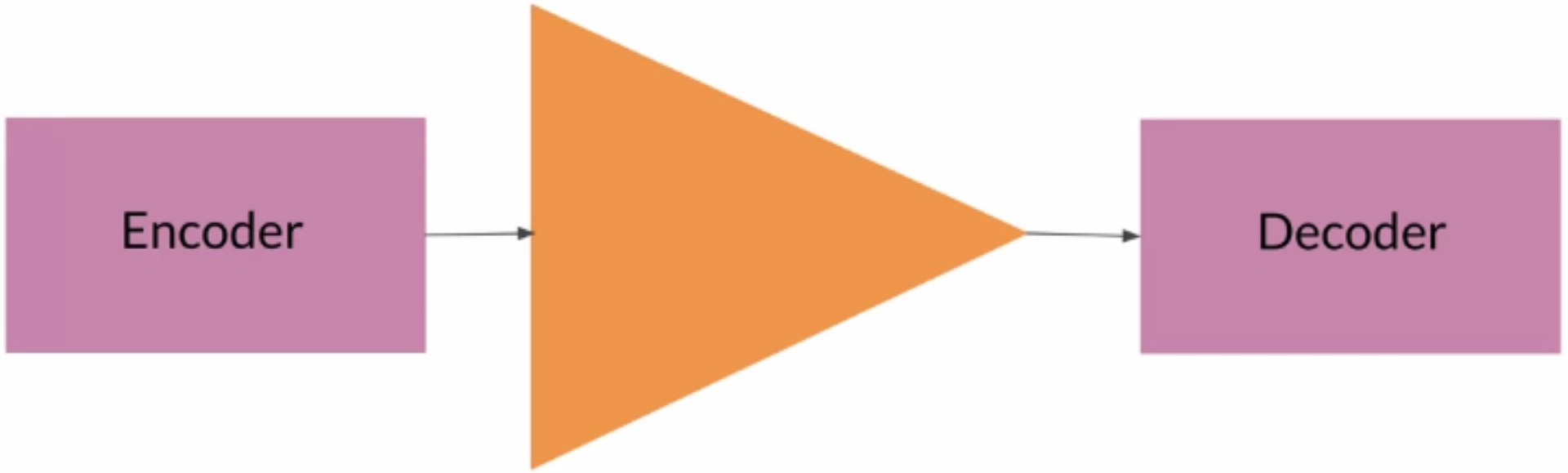
How are you today?

Wie geht es Ihnen heute?

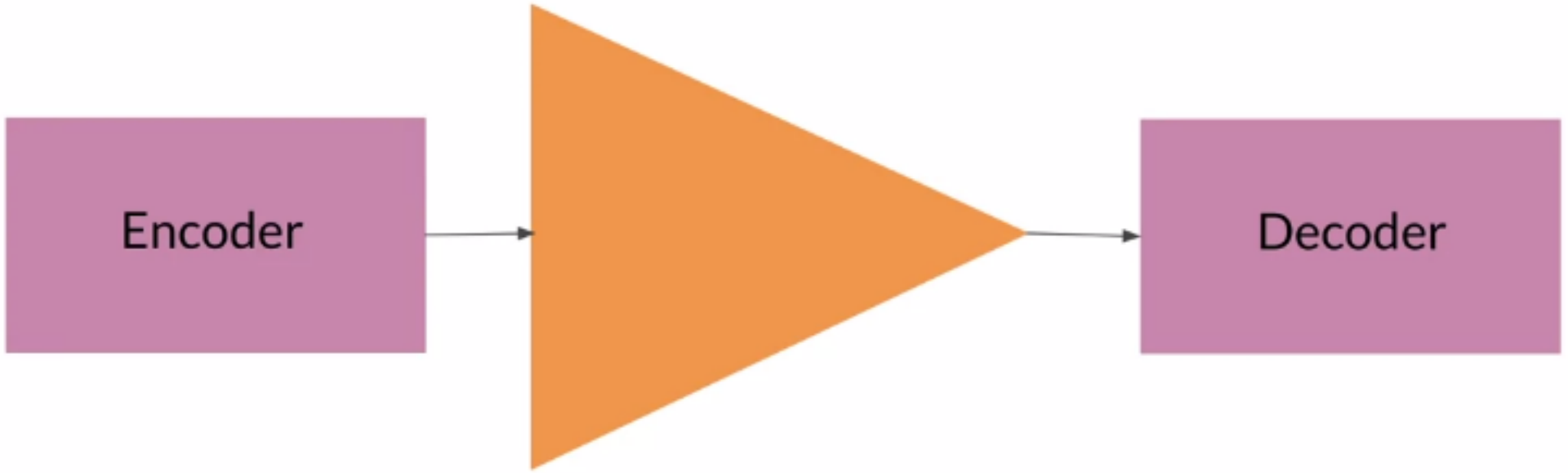
Seq2Seq model



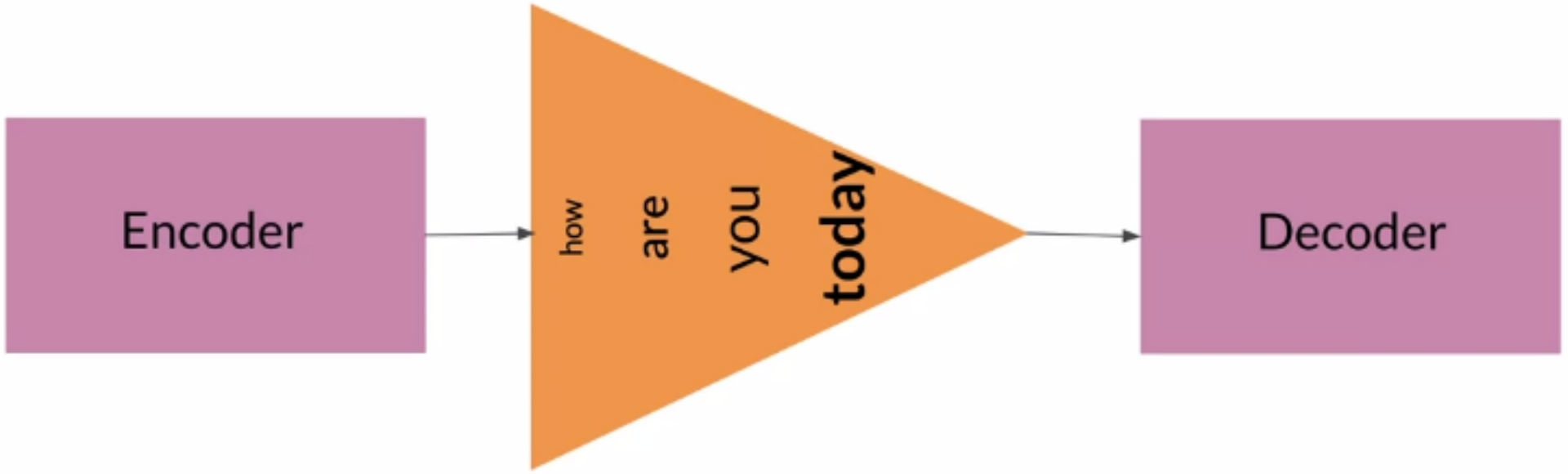
The information bottleneck



The information bottleneck



The information bottleneck



Seq2Seq shortcomings

- Variable-length sentences + fixed-length memory =

Seq2Seq shortcomings

- Variable-length sentences + fixed-length memory =

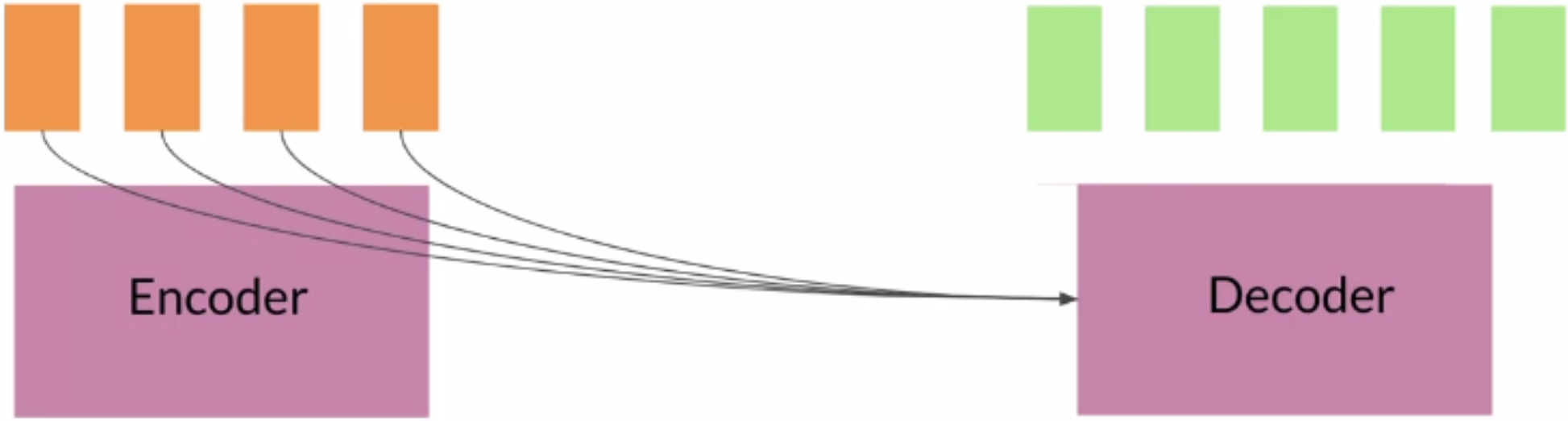
Seq2Seq shortcomings

- Variable-length sentences + fixed-length memory =

- As sequence size increases, model performance decreases

One vector per word

One vector per word



How are you today?

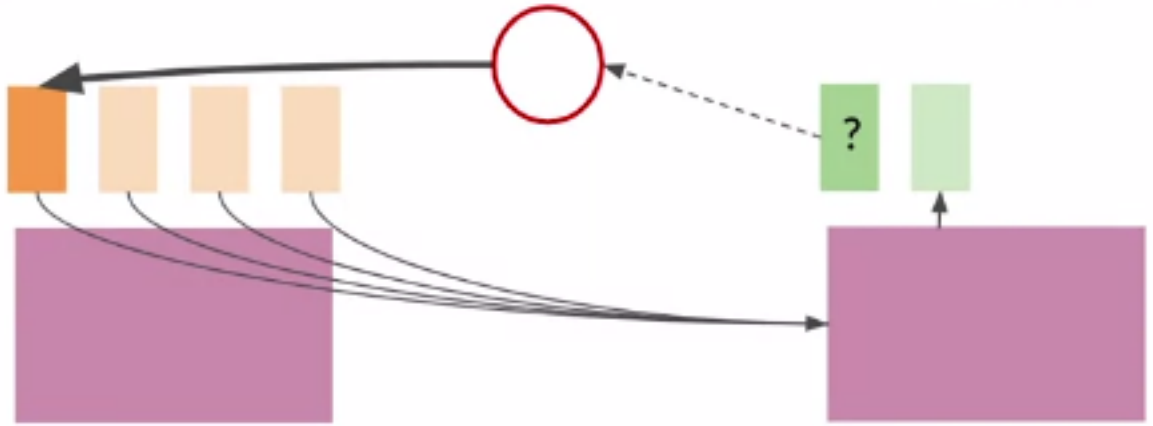
Wie geht es Ihnen heute?

Solution: focus attention in the right place

- Prevent sequence overload by giving the model a way to focus on the **likeliest** words at each step

Solution: focus attention in the right place

- Prevent sequence overload by giving the model a way to focus on the **likeliest** words at each step
- Do this by providing the information specific to each input word



Motivation for alignment

Motivation for alignment

Correctly aligned words are the goal:

- Translating from one language to another
- Word sense discovery and disambiguation

Motivation for alignment

Correctly aligned words are the goal:

- Translating from one language to another
- Word sense discovery and disambiguation

bank

Motivation for alignment

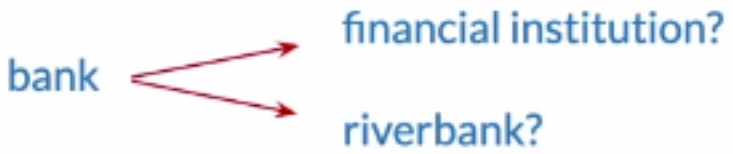
Correctly aligned words are the goal:

- Translating from one language to another
- Word sense discovery and disambiguation

Motivation for alignment

Correctly aligned words are the goal:

- Translating from one language to another
- Word sense discovery and disambiguation



Motivation for alignment

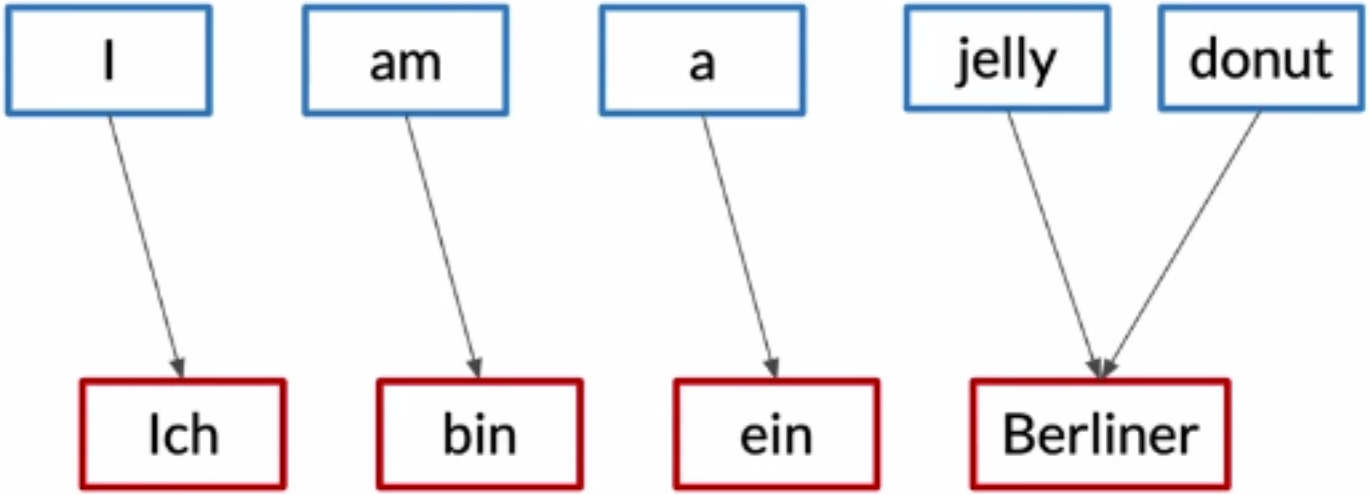
Correctly aligned words are the goal:

- Translating from one language to another
- Word sense discovery and disambiguation

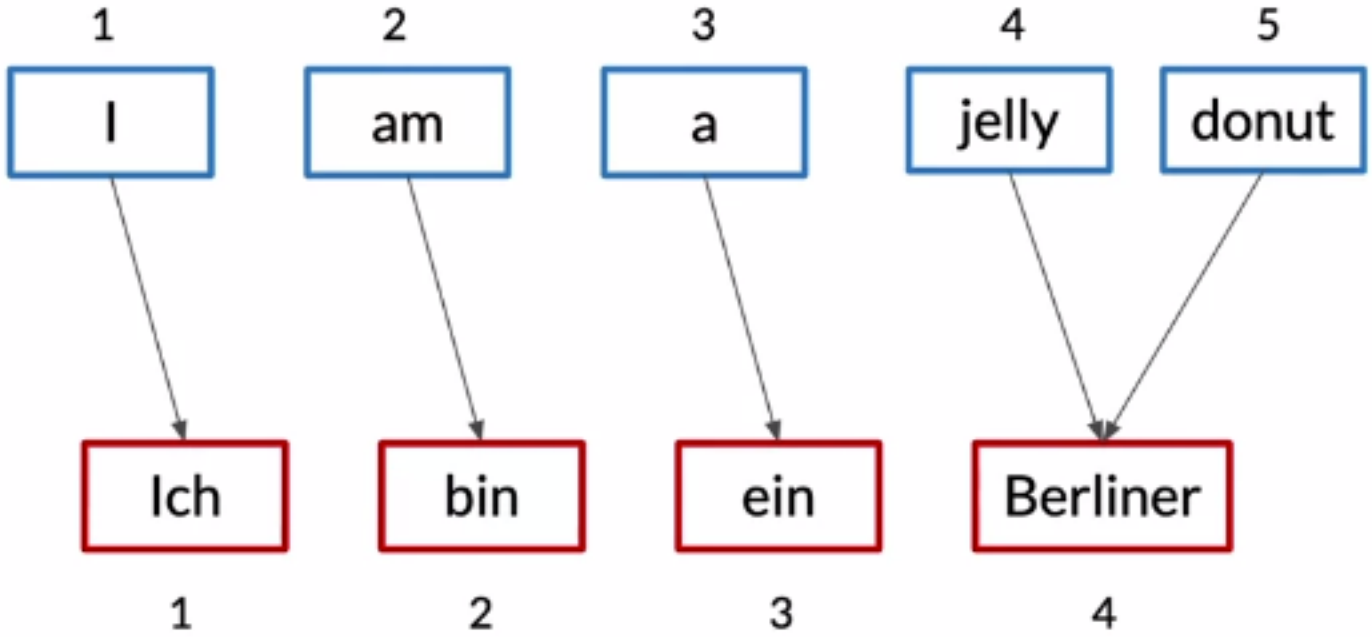
- Achieve alignment with a system for retrieving information step by step and scoring it

Word alignment

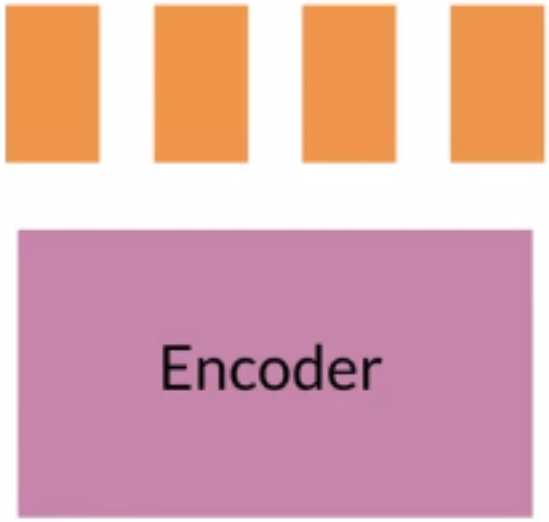
Word alignment

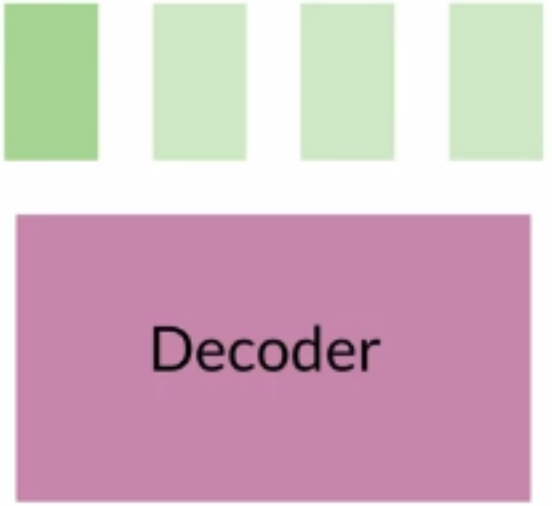


Word alignment



Which word to pay more attention to?

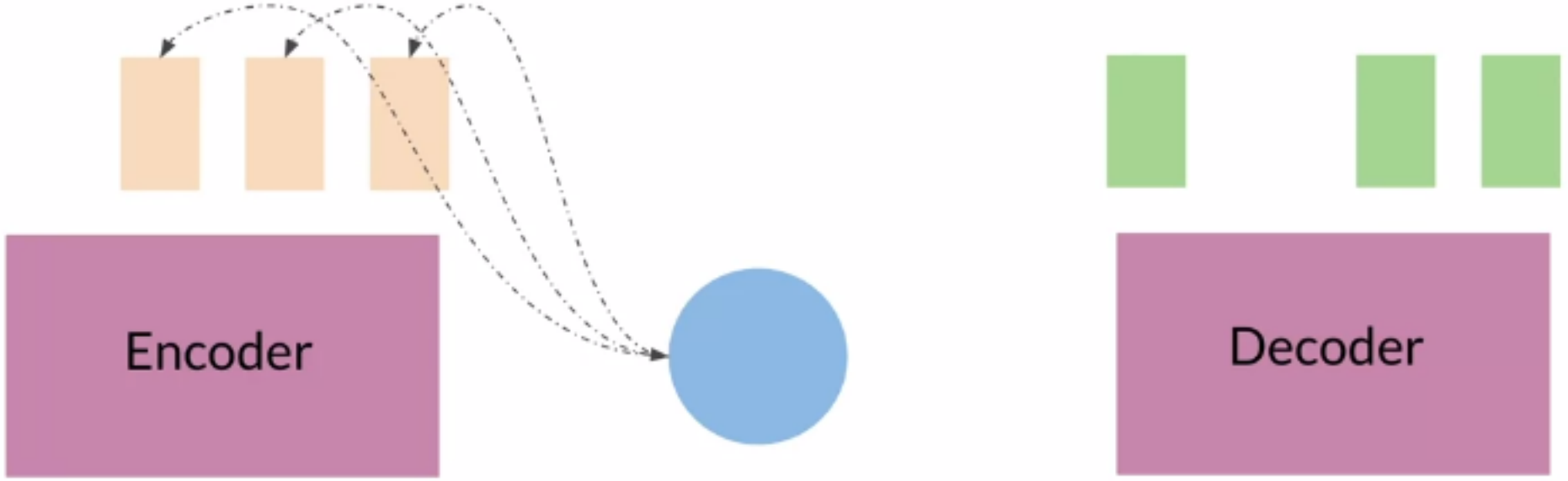




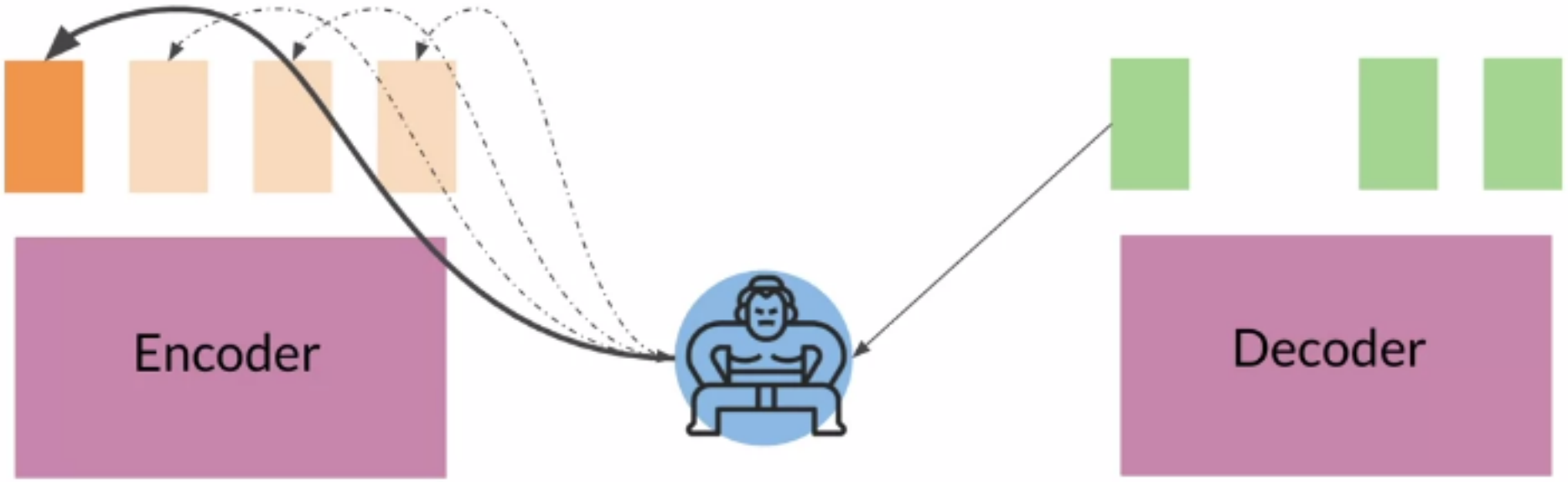
How are you today?

Which word to pay more attention to?

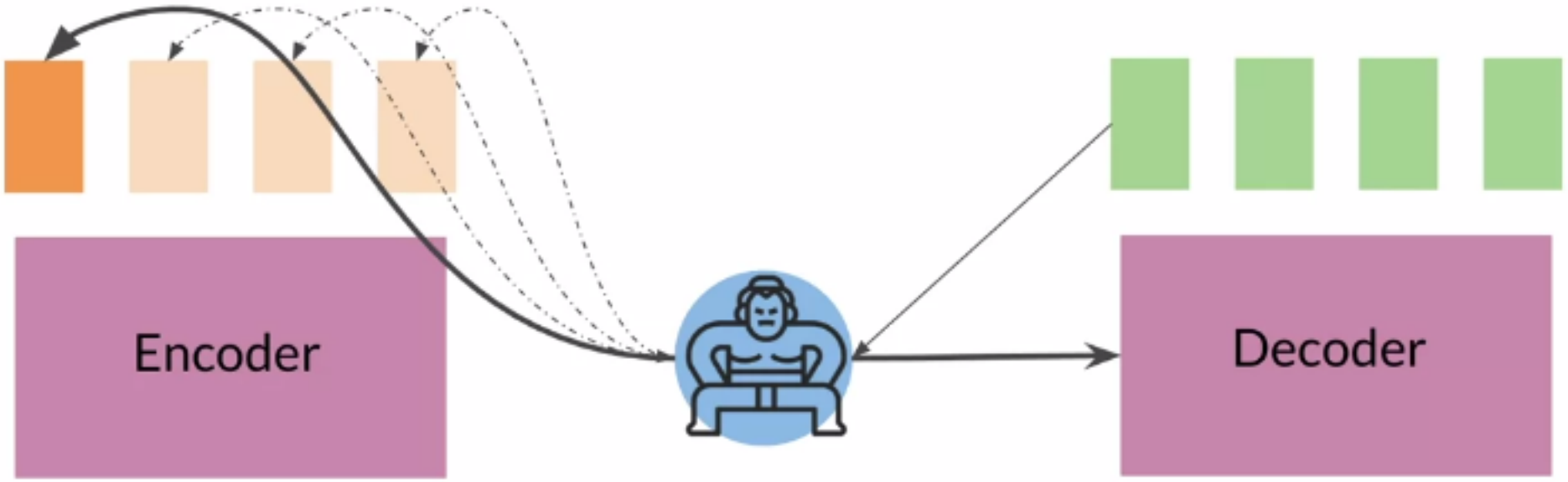
Give some inputs more weight!



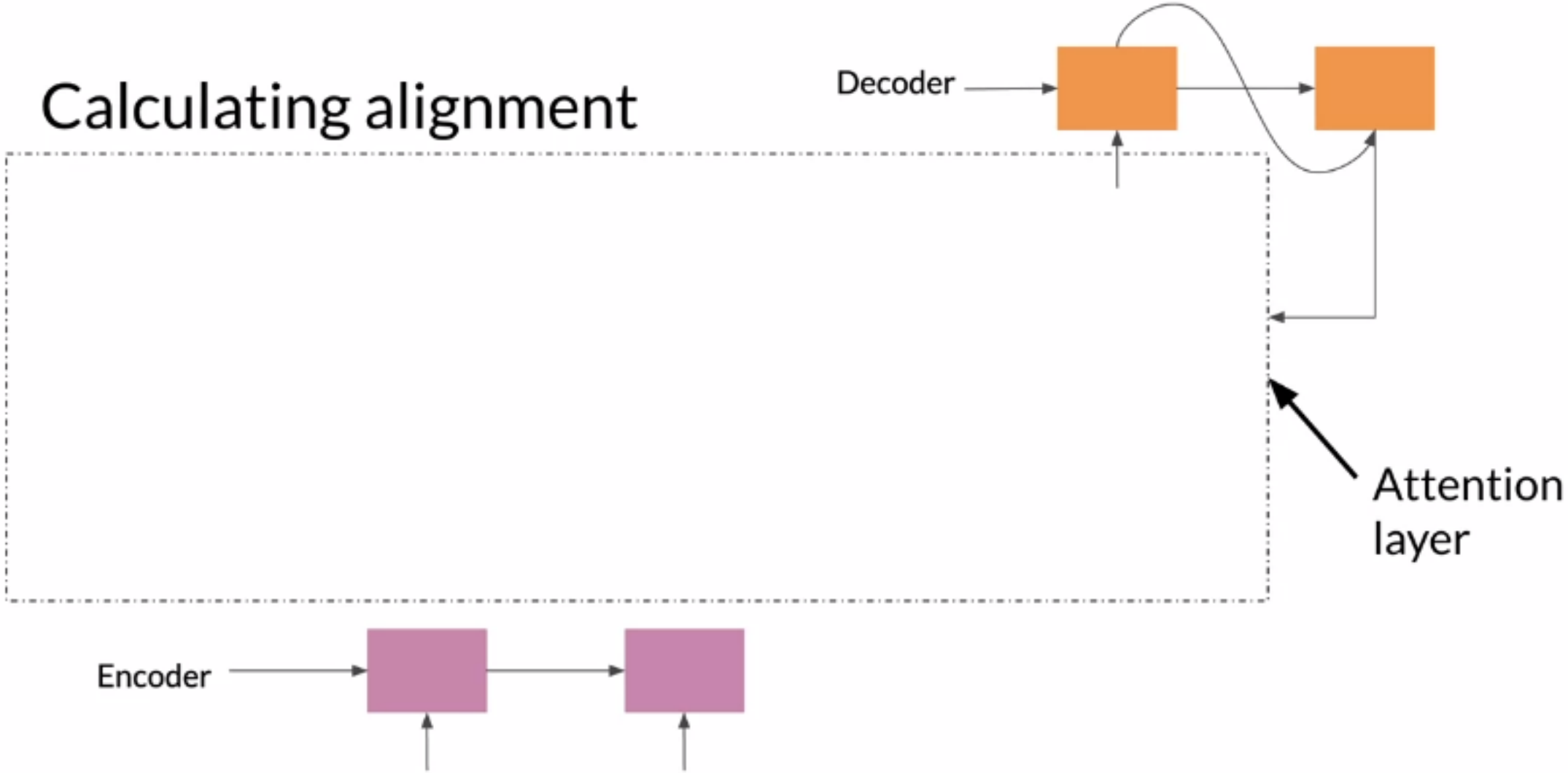
Give some inputs more weight!



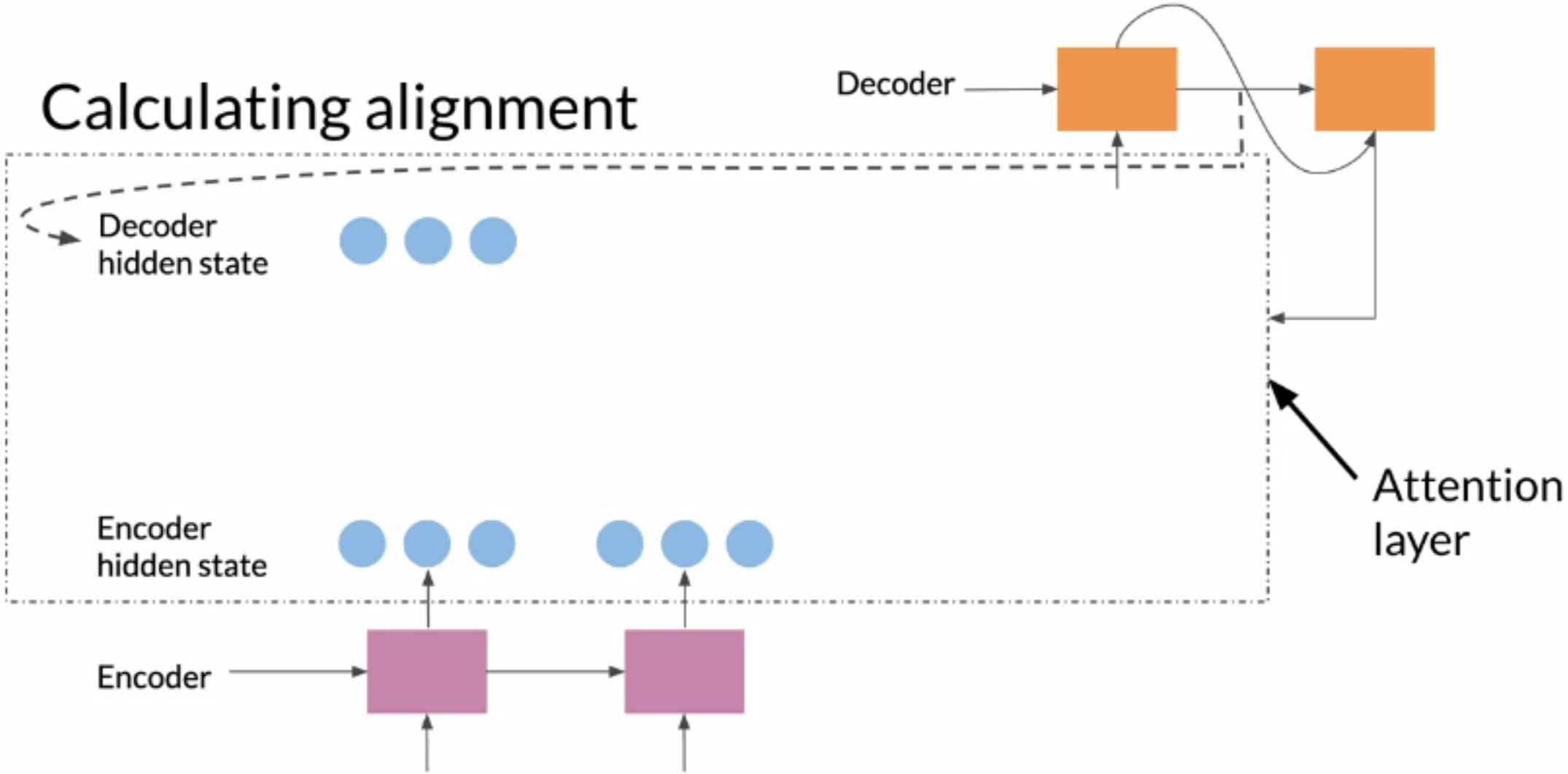
Give some inputs more weight!



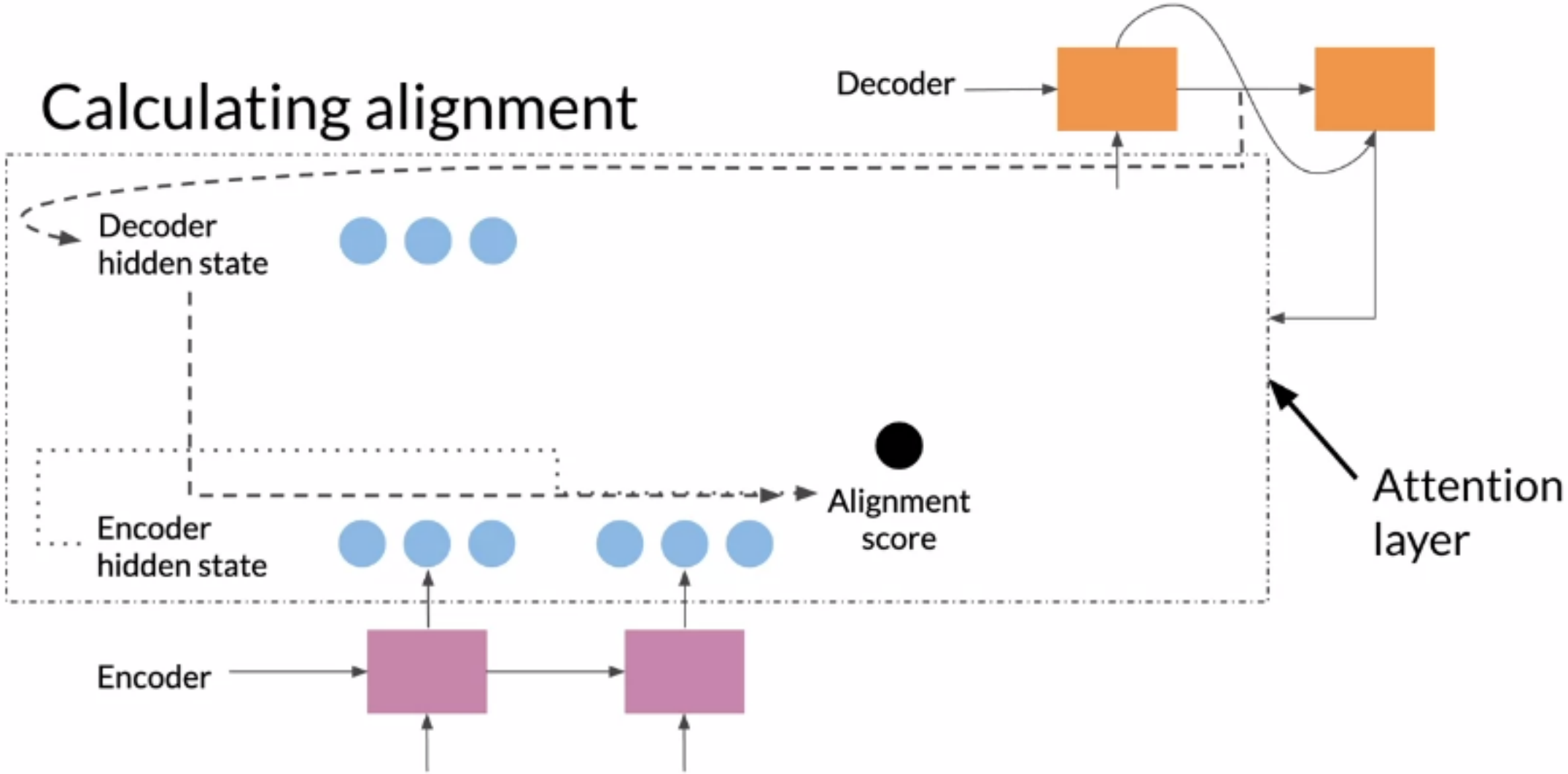
Calculating alignment

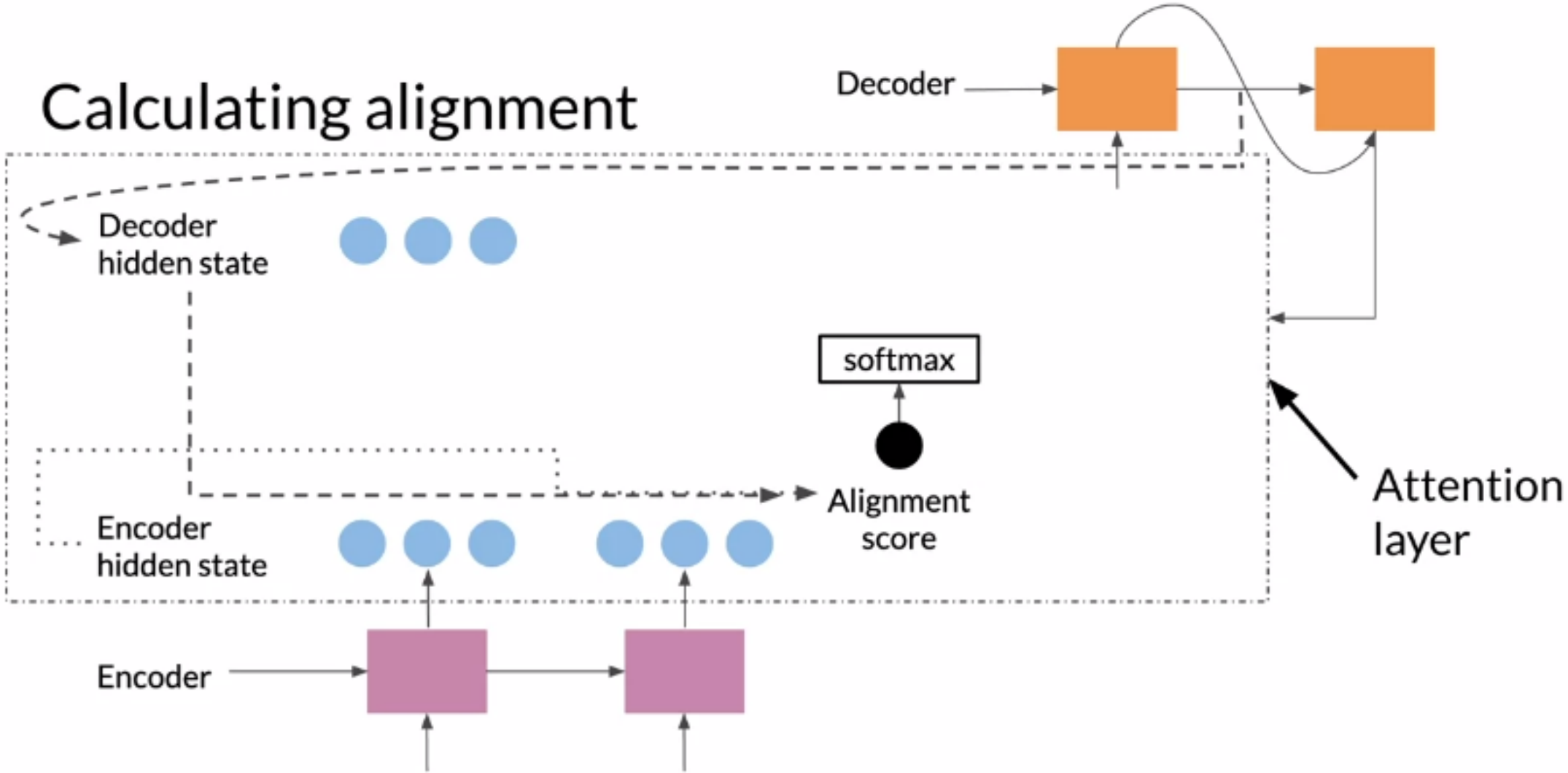


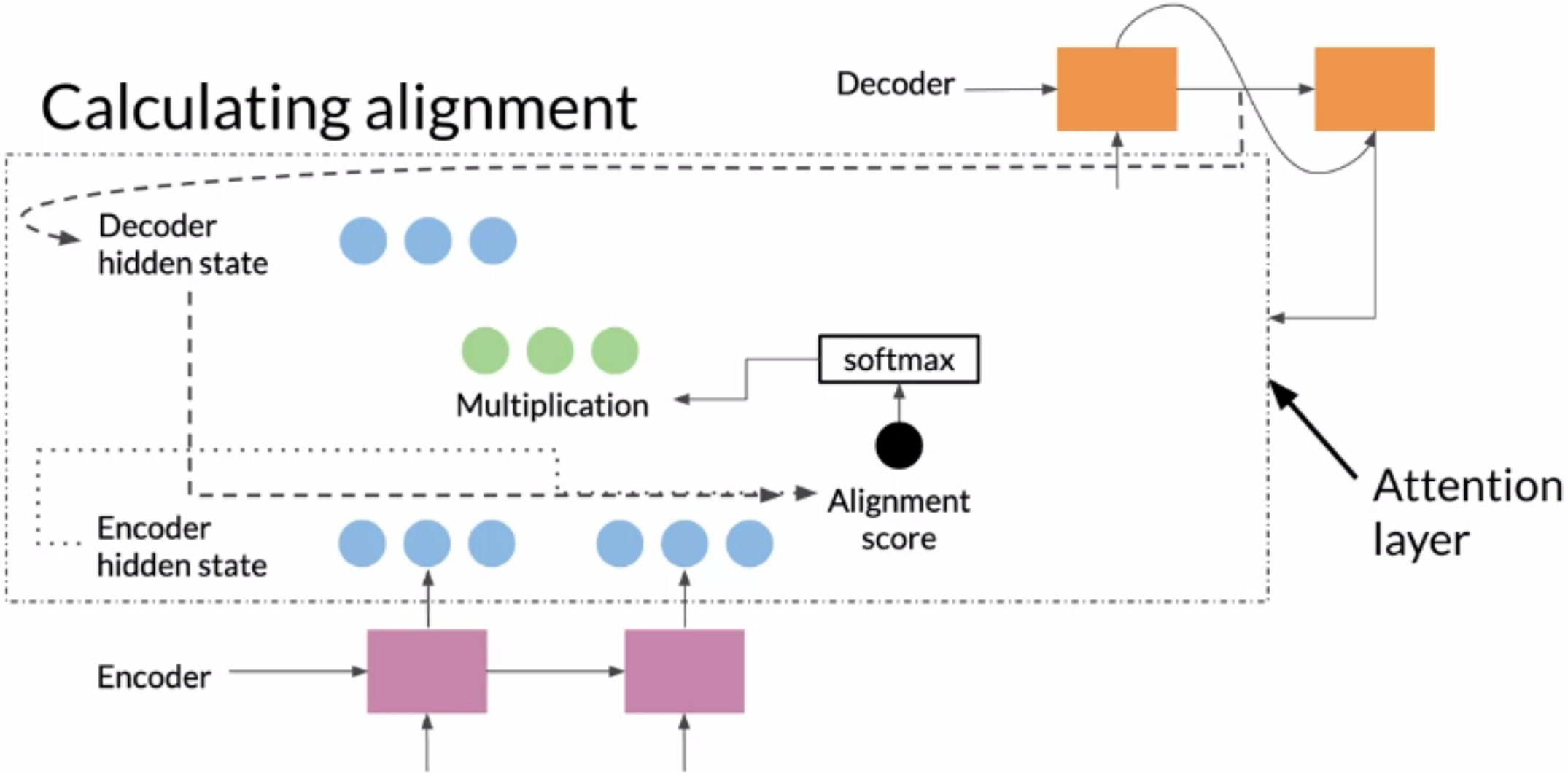
Calculating alignment

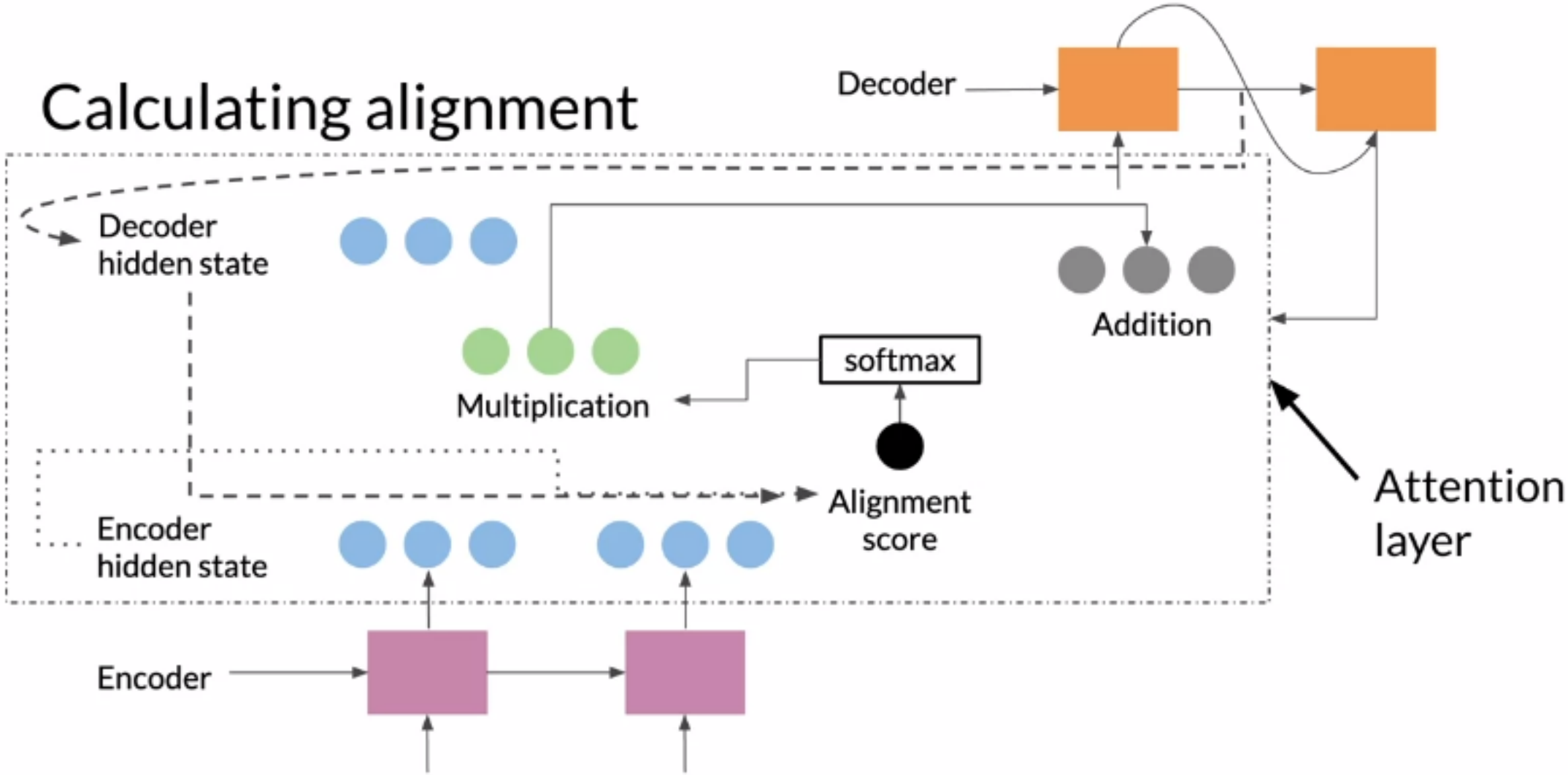


Calculating alignment




Calculating alignment
Decoder
Decoder hidden state
softmax
Alignment score
Attention layer
Encoder hidden state
Encoder


Calculating alignment
Decoder
Decoder hidden state
Multiplication
softmax
Alignment score
Attention layer
Encoder hidden state
Encoder


Calculating alignment
Decoder
Decoder hidden state
Addition
Multiplication
softmax
Alignment score
Attention layer
Encoder hidden state
Encoder

Outline

- Concept of attention for information retrieval
- Keys, Queries, and Values

Information retrieval

Information retrieval

Say you're looking for your keys.

Information retrieval

Say you're looking for your keys.

You ask your mom to help you find them.

Information retrieval

Say you're looking for your keys.

You ask your mom to help you find them.

She weighs the possibilities based on where the keys usually are, then tells you the most likely place.

Information retrieval

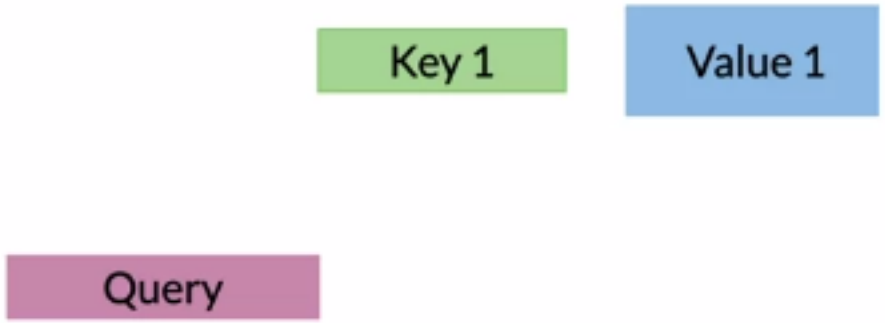
Say you're looking for your keys.

You ask your mom to help you find them.

She weighs the possibilities based on where the keys usually are, then tells you the most likely place.

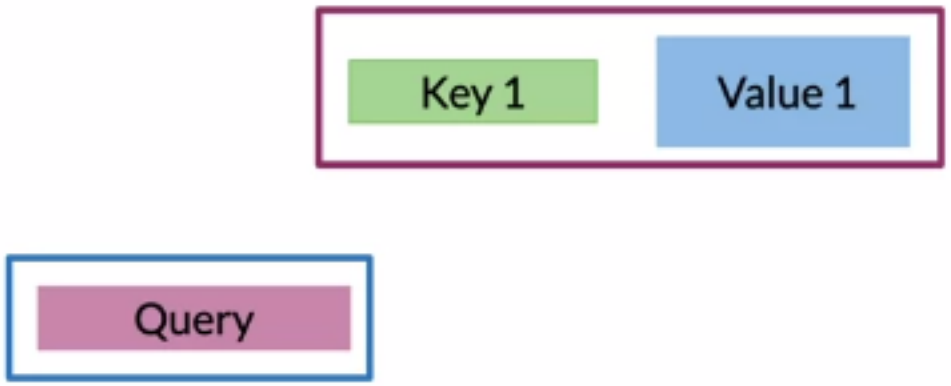
This is what Attention is doing: using your query to look in the right place, and find the key.

Inside the Attention Layer

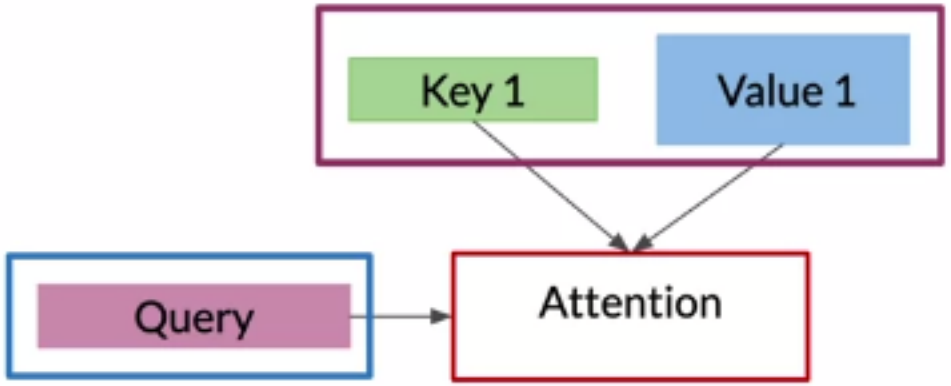


Inside the Attention Layer

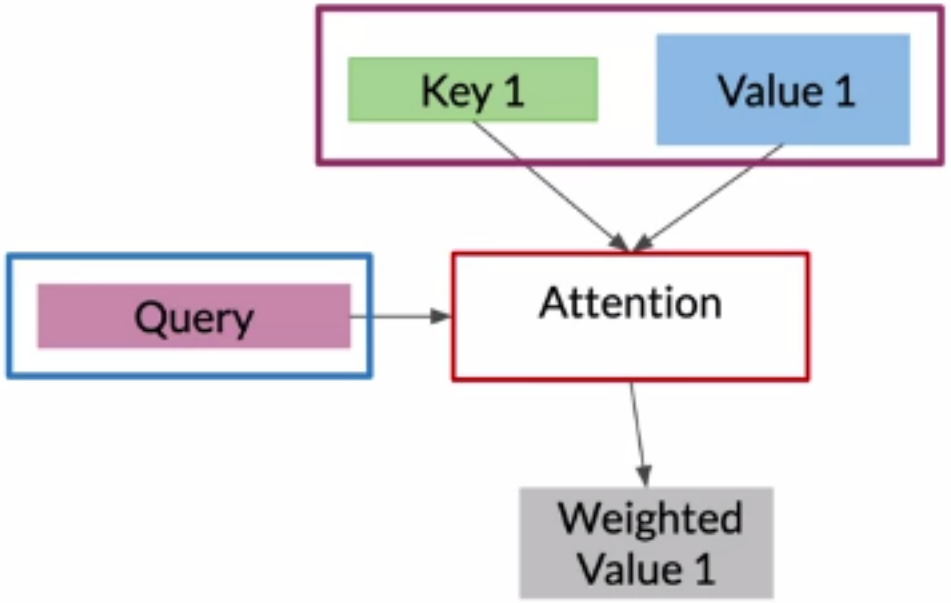
Inside the Attention Layer



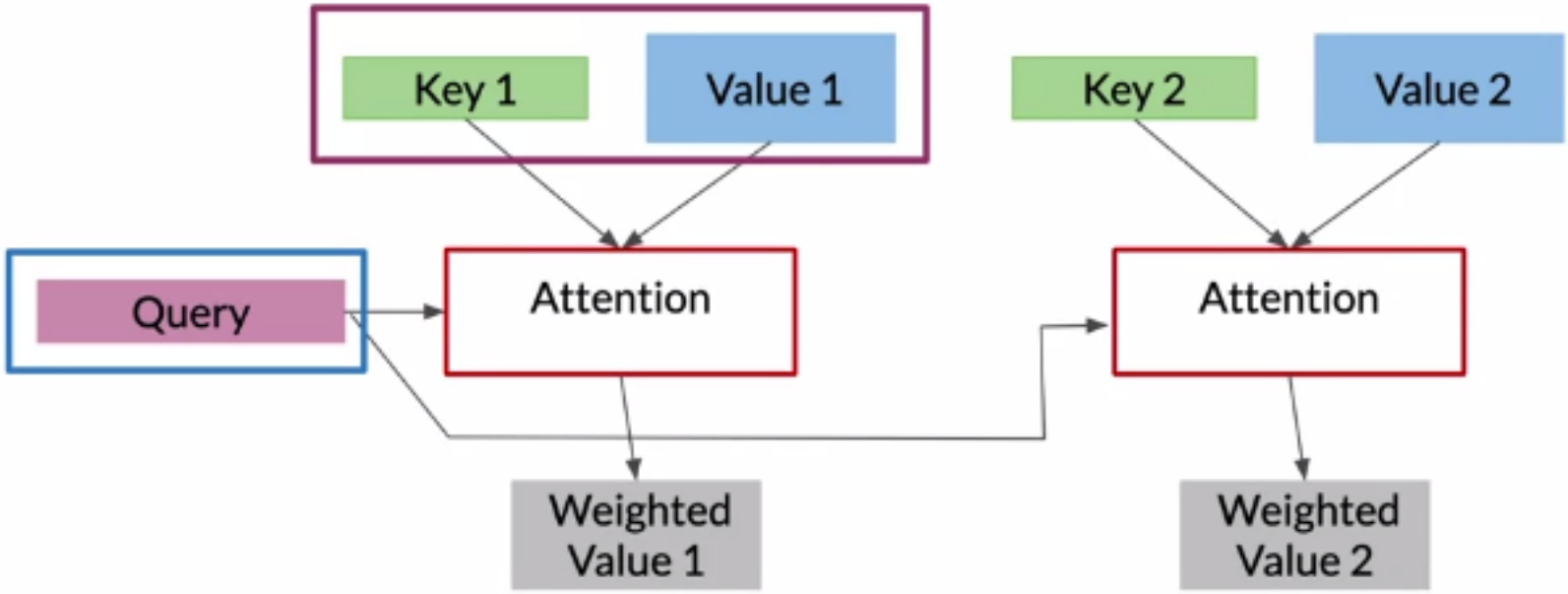
Inside the Attention Layer



Inside the Attention Layer

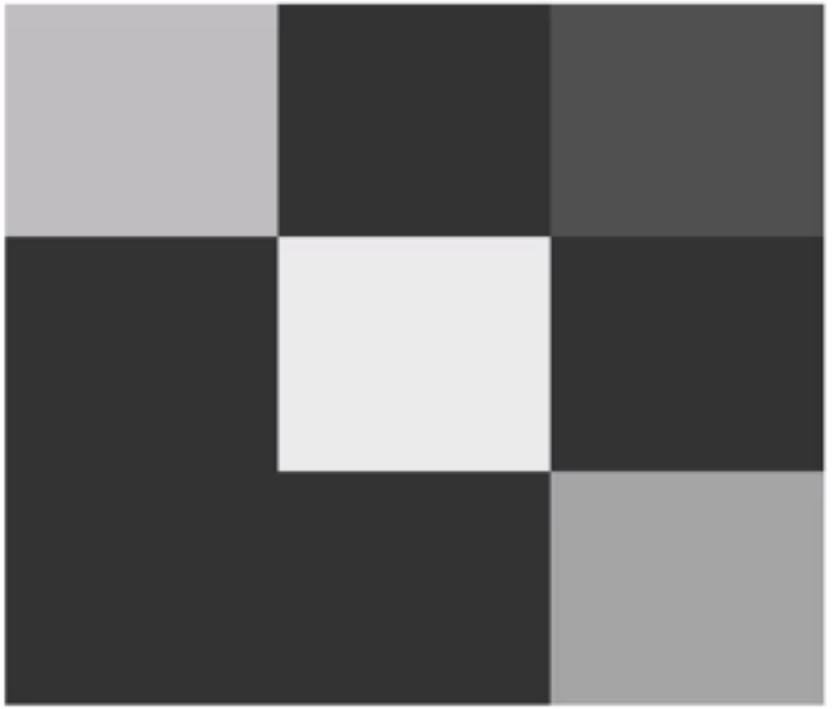


Inside the Attention Layer



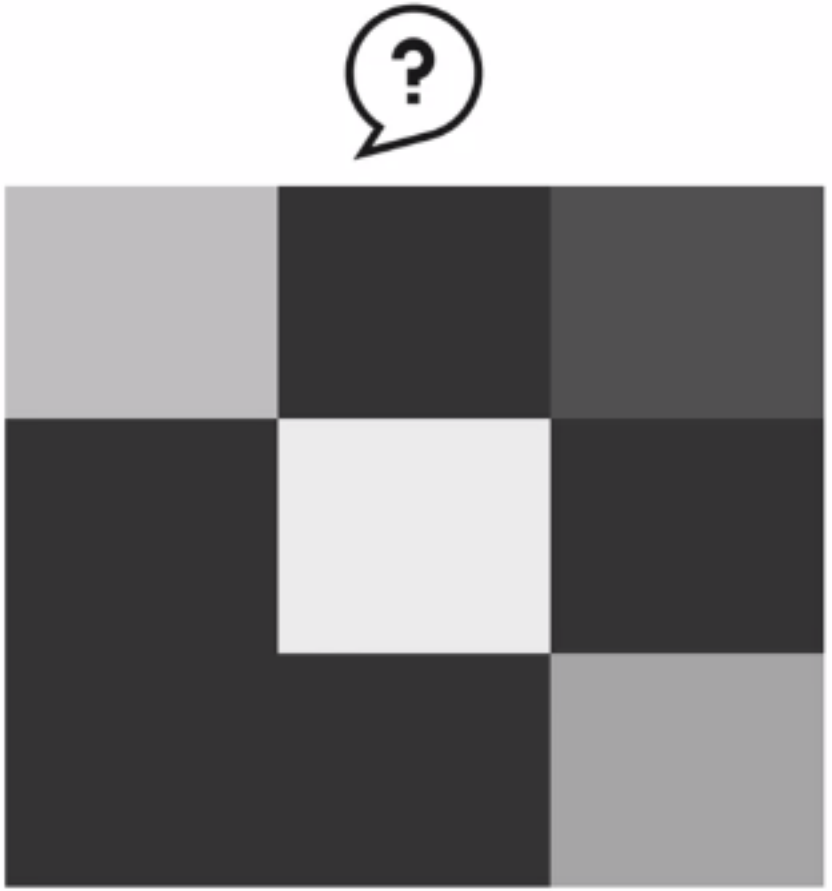
Attention

Keys and queries = 1 matrix
with the words of one query (Q)
as columns and the words of the
keys (K) as the rows



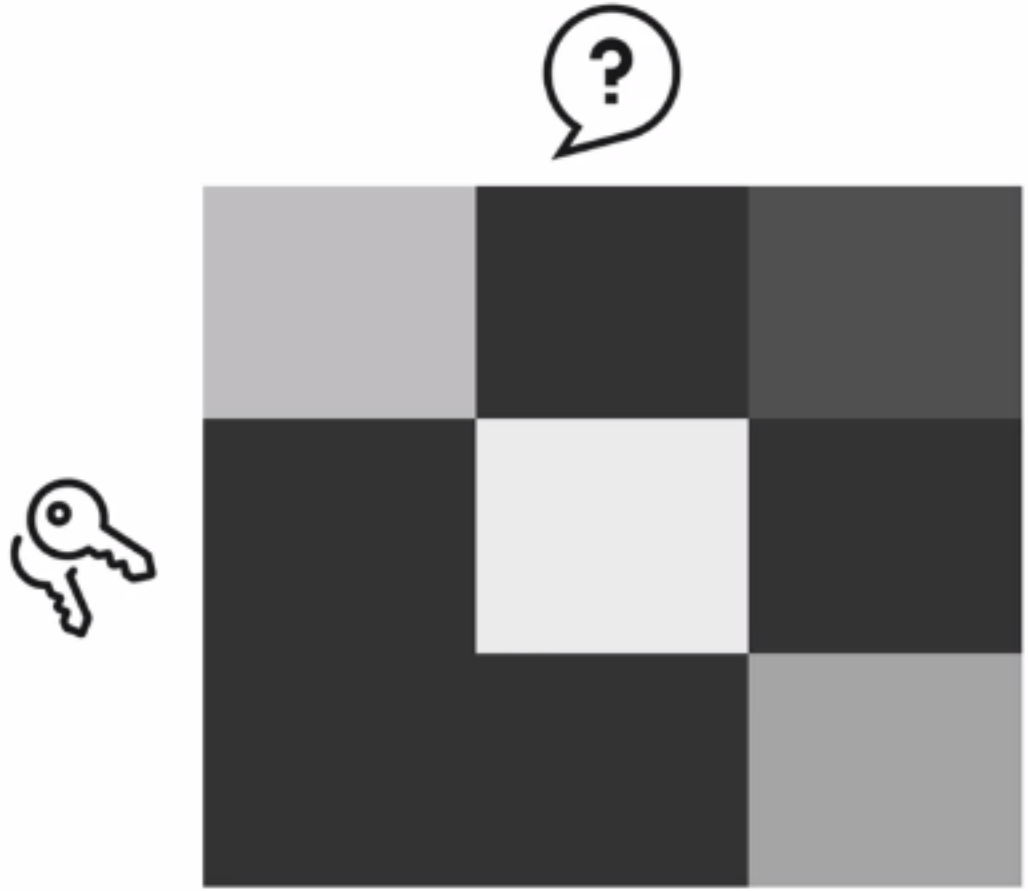
Attention

Keys and queries = 1 matrix
with the words of one query (Q)
as columns and the words of the
keys (K) as the rows



Attention

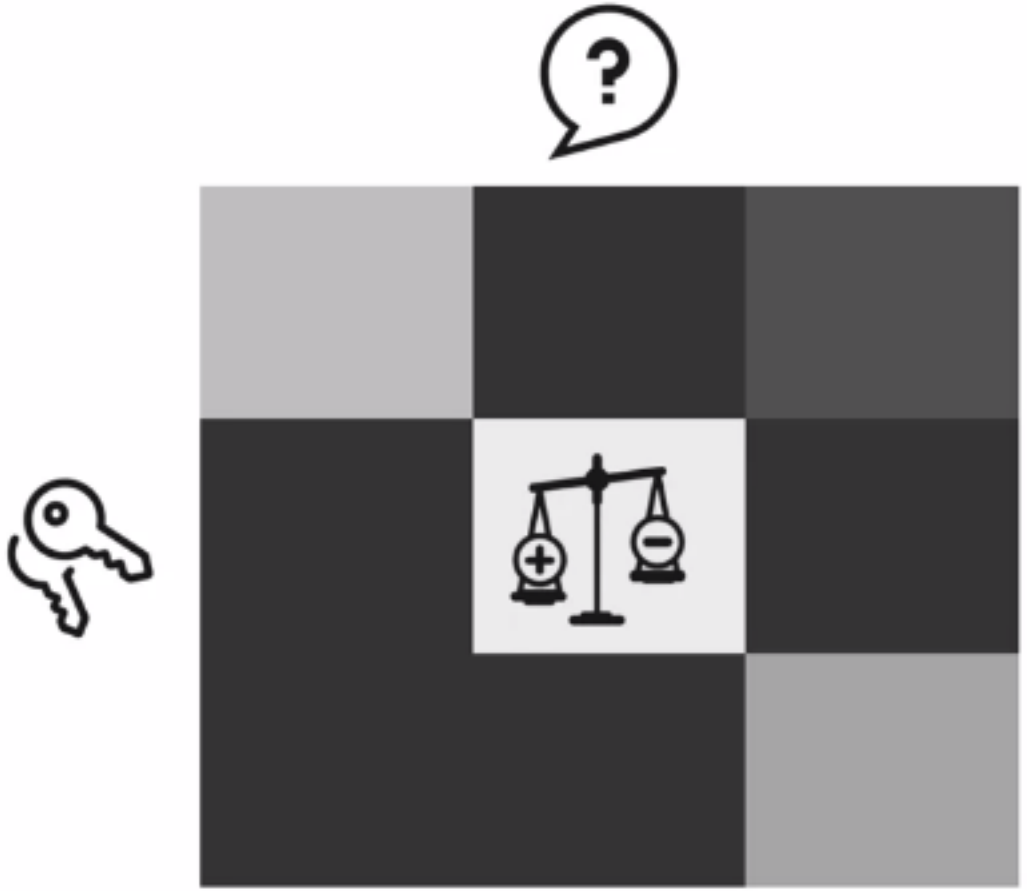
Keys and queries = 1 matrix
with the words of one query (Q)
as columns and the words of the
keys (K) as the rows



Attention

Keys and queries = 1 matrix
with the words of one query (Q)
as columns and the words of the
keys (K) as the rows

Value score (V) assigned based on
the closeness of the match

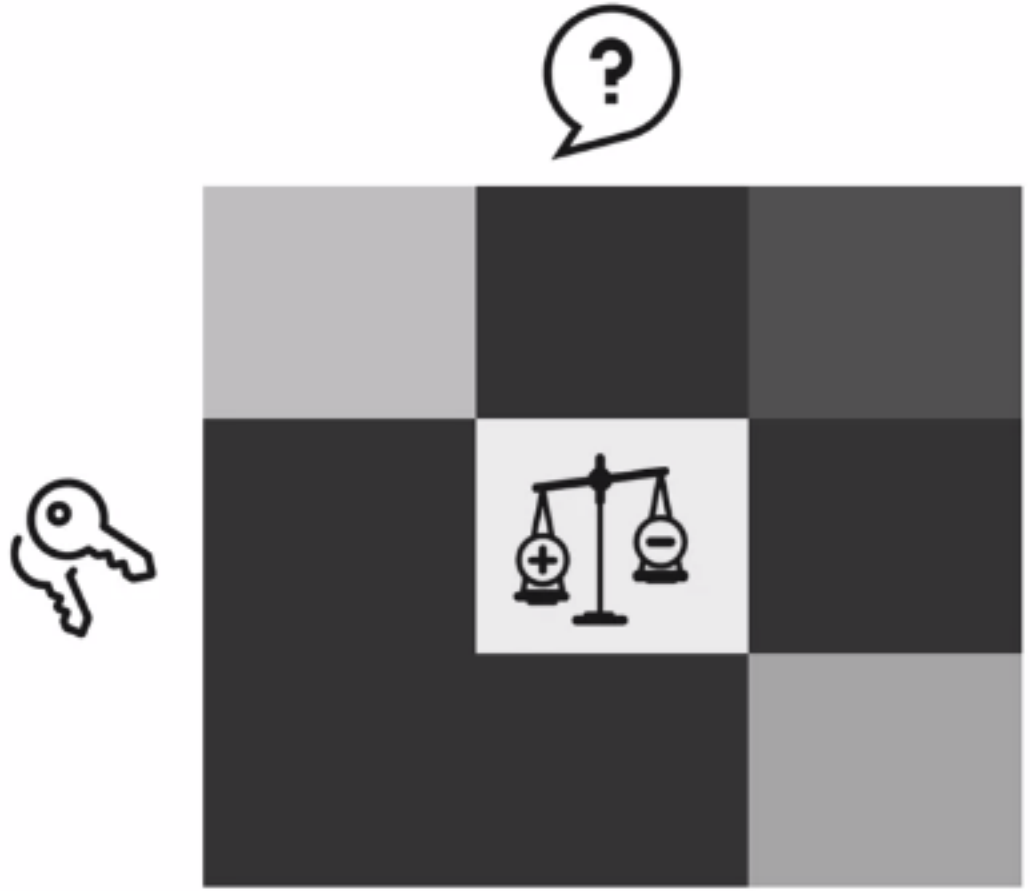


Attention

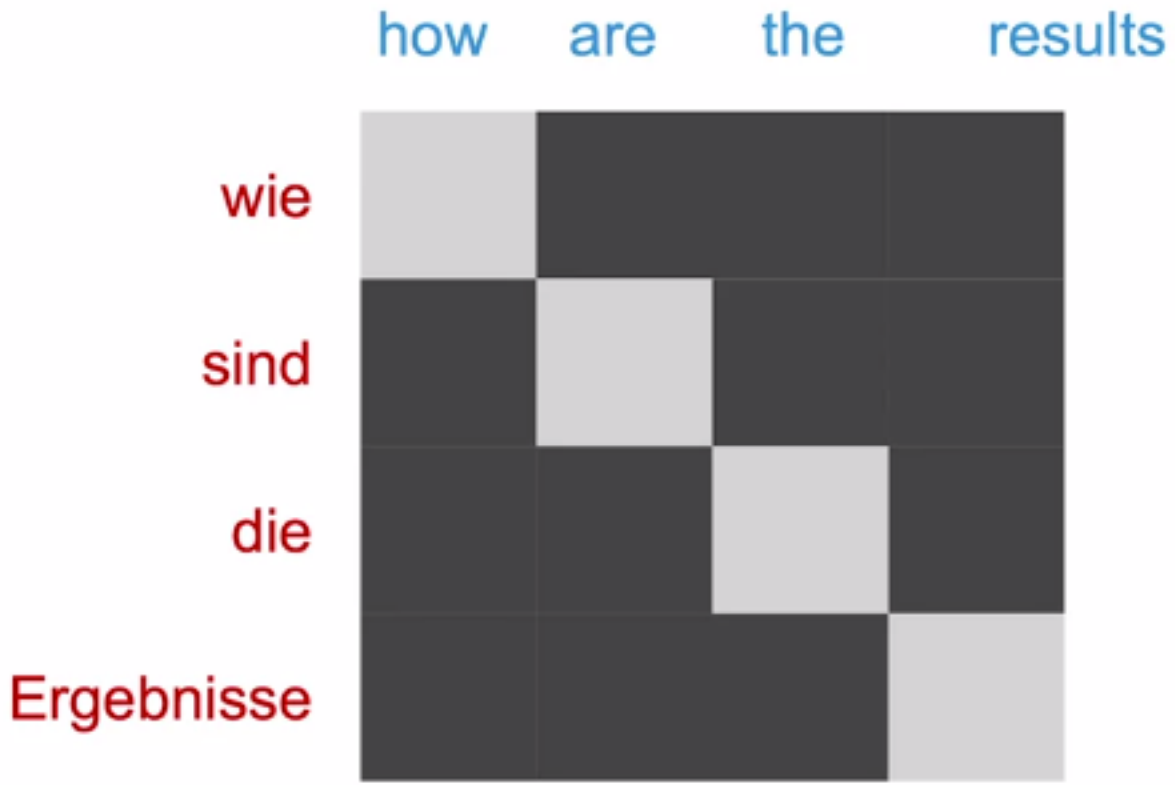
Keys and queries = 1 matrix with the words of one query (Q) as columns and the words of the keys (K) as the rows

Value score (V) assigned based on the closeness of the match

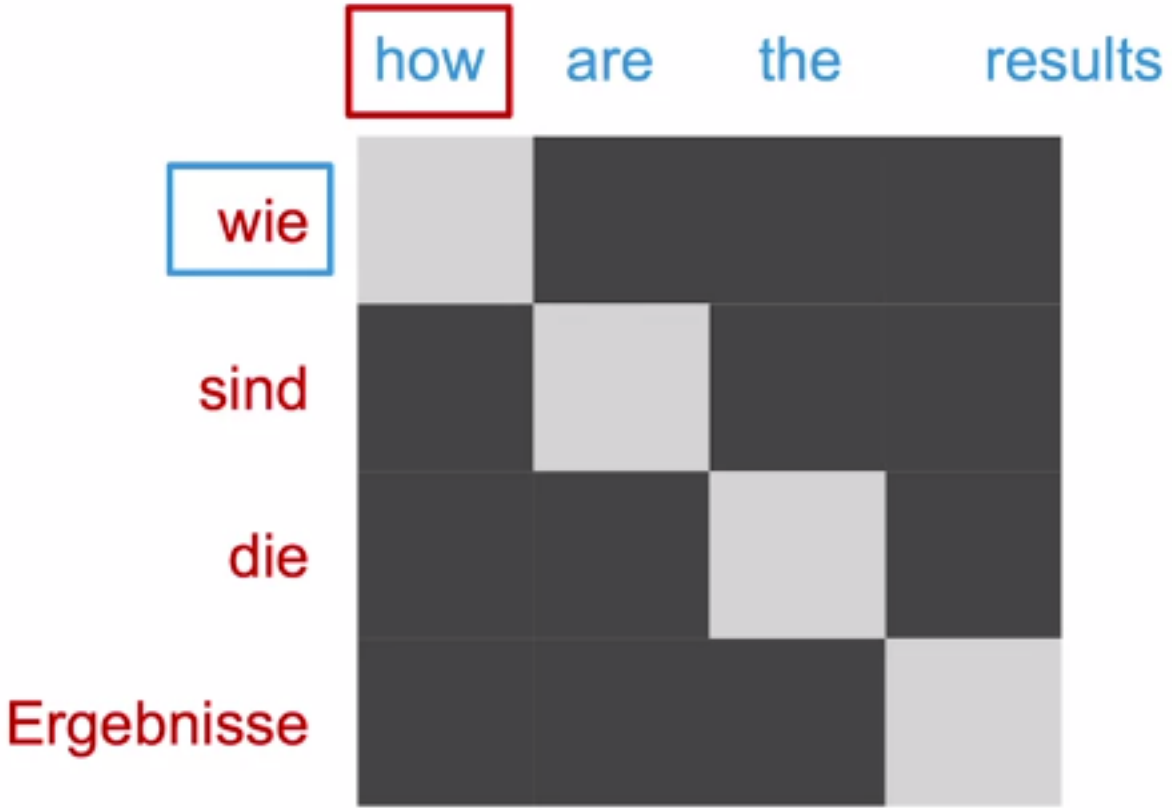
Attention = *Softmax*(QK^T)V



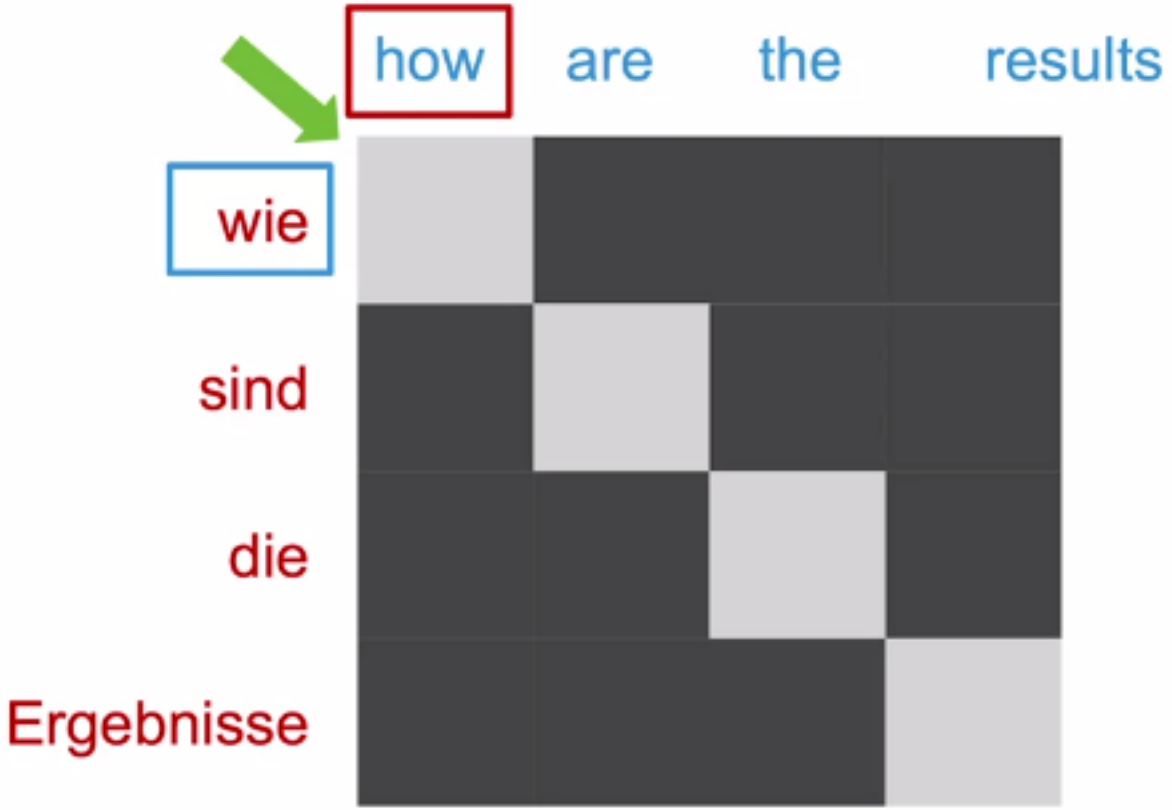
Neural machine translation with attention



Neural machine translation with attention



Neural machine translation with attention



Flexible attention

For languages with different grammar structures, attention still looks at the correct token between them

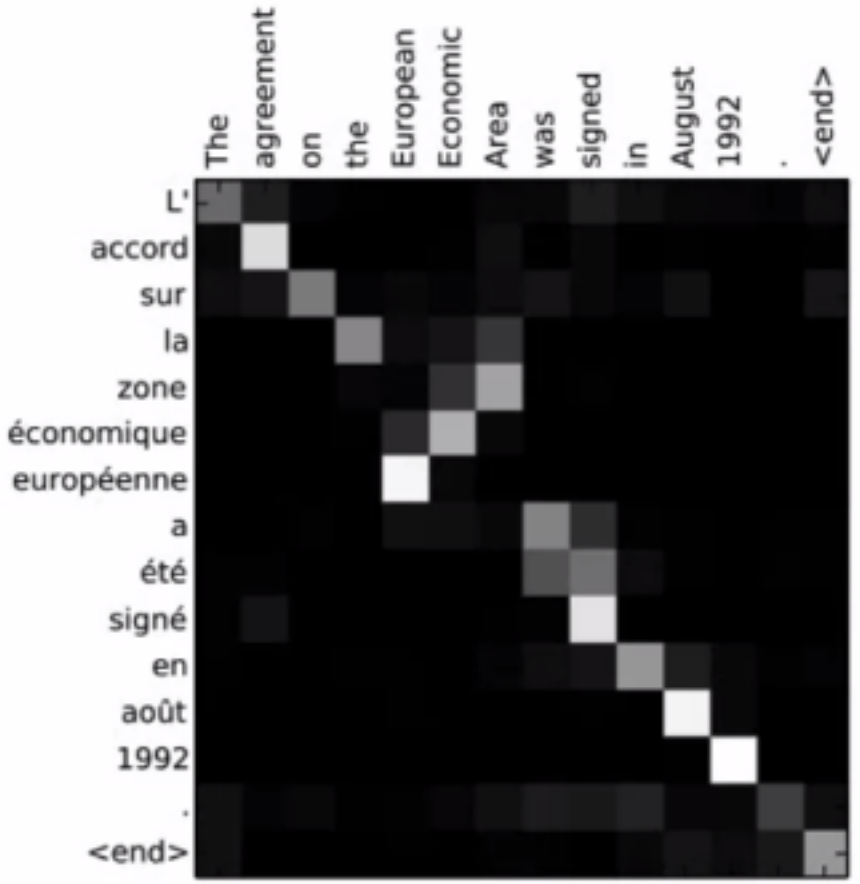


image © (Bahdanau et al., 2015)

Flexible attention

For languages with different grammar structures, attention still looks at the correct token between them

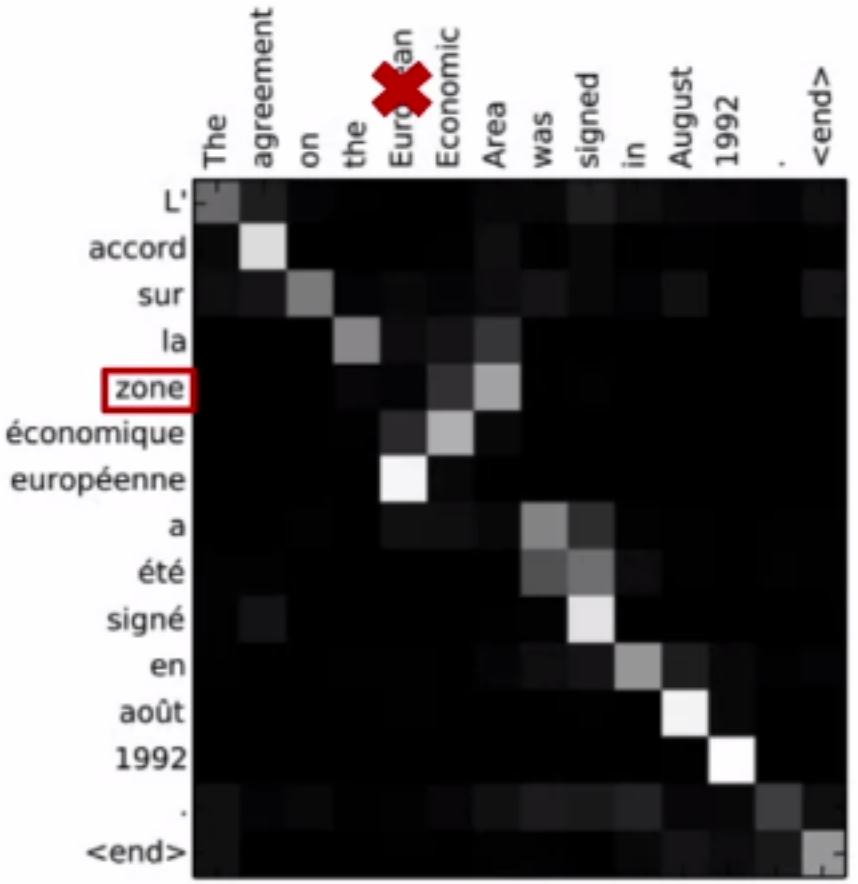


image © (Bahdanau et al., 2015)

Flexible attention

For languages with different grammar structures, attention still looks at the correct token between them

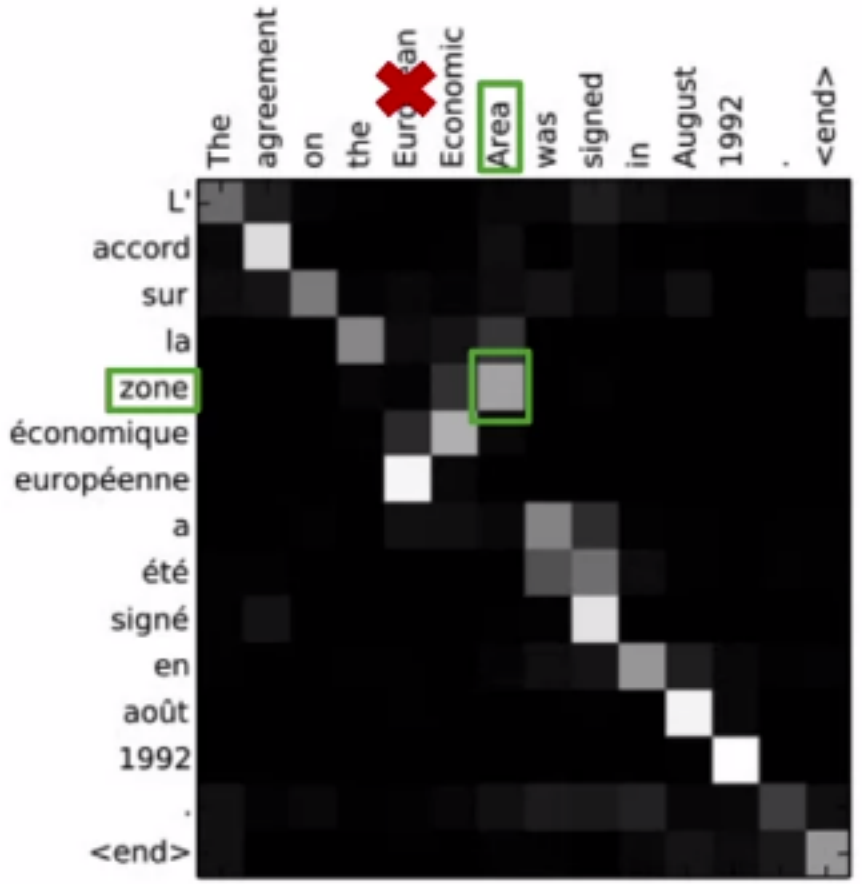


image © (Bahdanau et al., 2015)

Summary

- Attention is an added layer that lets a model focus on what's important

Summary

- Attention is an added layer that lets a model focus on what's important
- Queries, Values, and Keys are used for information retrieval inside the Attention layer

Summary

- Attention is an added layer that lets a model focus on what's important
- Queries, Values, and Keys are used for information retrieval inside the Attention layer
- This flexible system finds matches even between languages with very different grammatical structures

Data in machine translation

Data in machine translation

English	German
I am hungry!	Ich habe Hunger.
...	...
I watched the soccer game.	Ich habe das Fußballspiel gesehen.

Data in machine translation

English	German
I am hungry!	Ich habe Hunger.
...	...
I watched the soccer game.	Ich habe das Fußballspiel gesehen.

Attention! (no pun intended) Assignment dataset is not as squeaky-clean as this example and contains some Spanish translations.

Machine translation setup

State-of-the-art uses pre-trained vectors

Machine translation setup

State-of-the-art uses pre-trained vectors

Otherwise, represent words with a one-hot vector to create the input

Machine translation setup

State-of-the-art uses pre-trained vectors

Otherwise, represent words with a one-hot vector to create the input

Keep track of index mappings with word2ind and ind2word dictionaries

Machine translation setup

State-of-the-art uses pre-trained vectors

Otherwise, represent words with a one-hot vector to create the input

Keep track of index mappings with word2ind and ind2word dictionaries

Use start-of and end-of sequence tokens:

<SOS>  <EOS>

Preparing to Translate to German

ENGLISH SENTENCE:

Both the ballpoint and the mechanical pencil in the series are equipped with a special mechanism: when the twist mechanism is activated, the lead is pushed forward.

TOKENIZED VERSION OF THE ENGLISH SENTENCE:

[4546 4 11358 362 8 4 23326 20104 1745 8210 9641 5 6
4 3103 31 2767 30 13 914 4797 64 196 4 22474 5 4797 16
24864 86 2 4 1060 16 6413 1138 3 1 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]

English to German

GERMAN TRANSLATION:

Der Kugelschreiber und der Drehbleistift der Serie sind mit einem besonderen Mechanismus ausgestattet: Bei Betätigung der Drehmechanik wird die Schreibmine nach vorne geschoben.

TOKENIZED VERSION OF THE GERMAN TRANSLATION:

[149 3892 5280 14774 2418 12 11 9883 6959 7298 15157 5 11 8453 75
39 114 5324 10565 2520 64 752 12954 26538 147 11 9883 23326 20104
300 78 10 21150 10166 126 14566 5 23850 1171 3 1 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]

English to German

GERMAN TRANSLATION:

Der Kugelschreiber und der Drehbleistift der Serie sind mit einem besonderen Mechanismus ausgestattet: Bei Betätigung der Drehmechanik wird die Schreibmine nach vorne geschoben.

TOKENIZED VERSION OF THE GERMAN TRANSLATION:

[149 3892 5280 14774 2418 12 11 9883 6959 7298 15157 5 11 8453 75
39 114 5324 10565 2520 64 752 12954 26538 147 11 9883 23326 20104
300 78 10 21150 10166 126 14566 5 23850 1171 3 1 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]

Outline

- Teacher forcing
- Model for NMT with attention

How to know predictions are correct?

How to know predictions are correct?

Teacher forcing allows the model to “check its work” at each step

Or, compare its prediction against the real output during training

How to know predictions are correct?

Teacher forcing allows the model to “check its work” at each step

Or, compare its prediction against the real output during training

Result: Faster, more accurate training

Teacher forcing: motivation

How are the results?

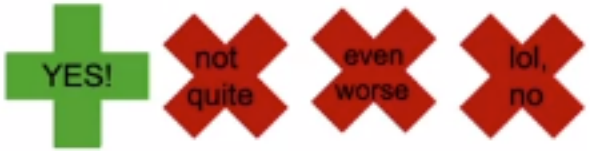
Encoder

Wie sind die Ergebnisse?

Actual target:

Wie geht zu Hause?

Prediction:



Decoder

Teacher forcing: motivation

How are the results?

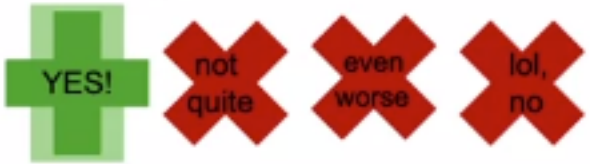
Encoder

Wie sind die Ergebnisse?

Actual target:

Wie geht zu Hause?

Prediction:



Decoder

Teacher forcing: motivation

How are the results?

Encoder

Decoder

Teacher forcing: motivation

How are the results?

Encoder

Decoder

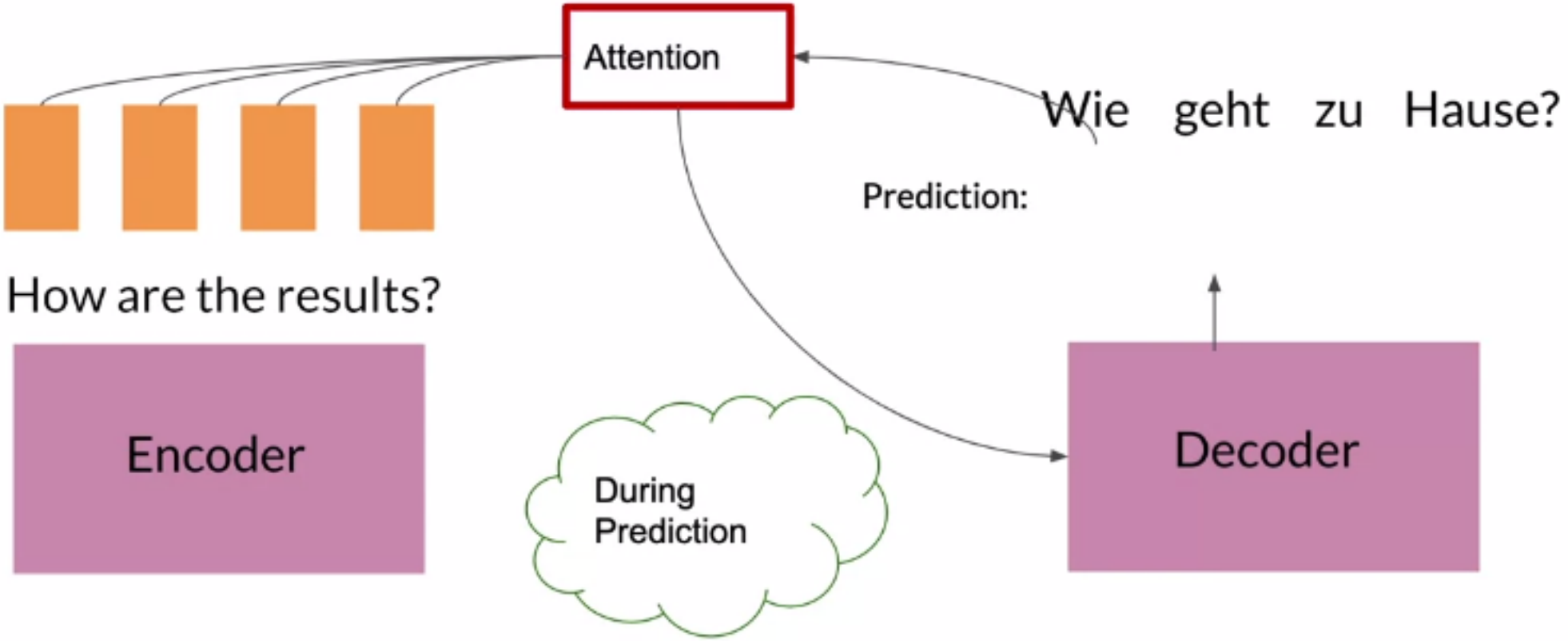
Teacher forcing: motivation

How are the results?

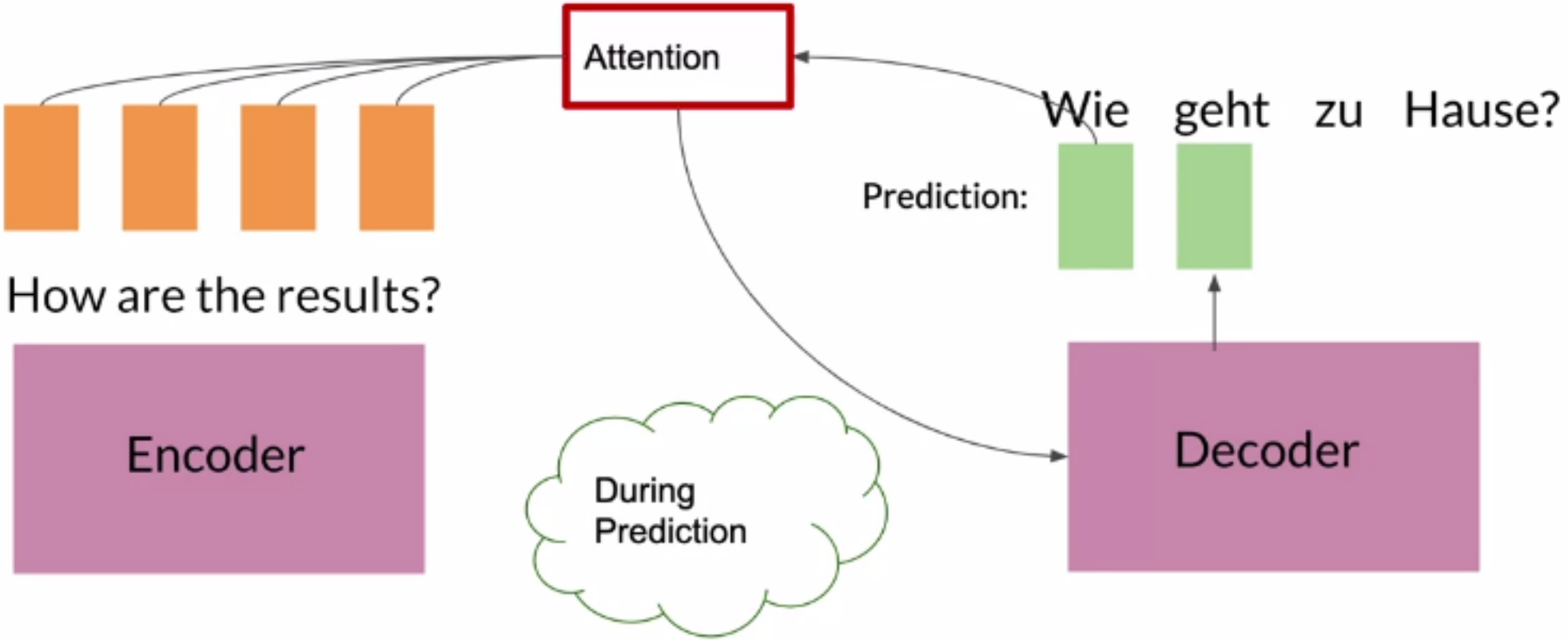
Encoder

Decoder

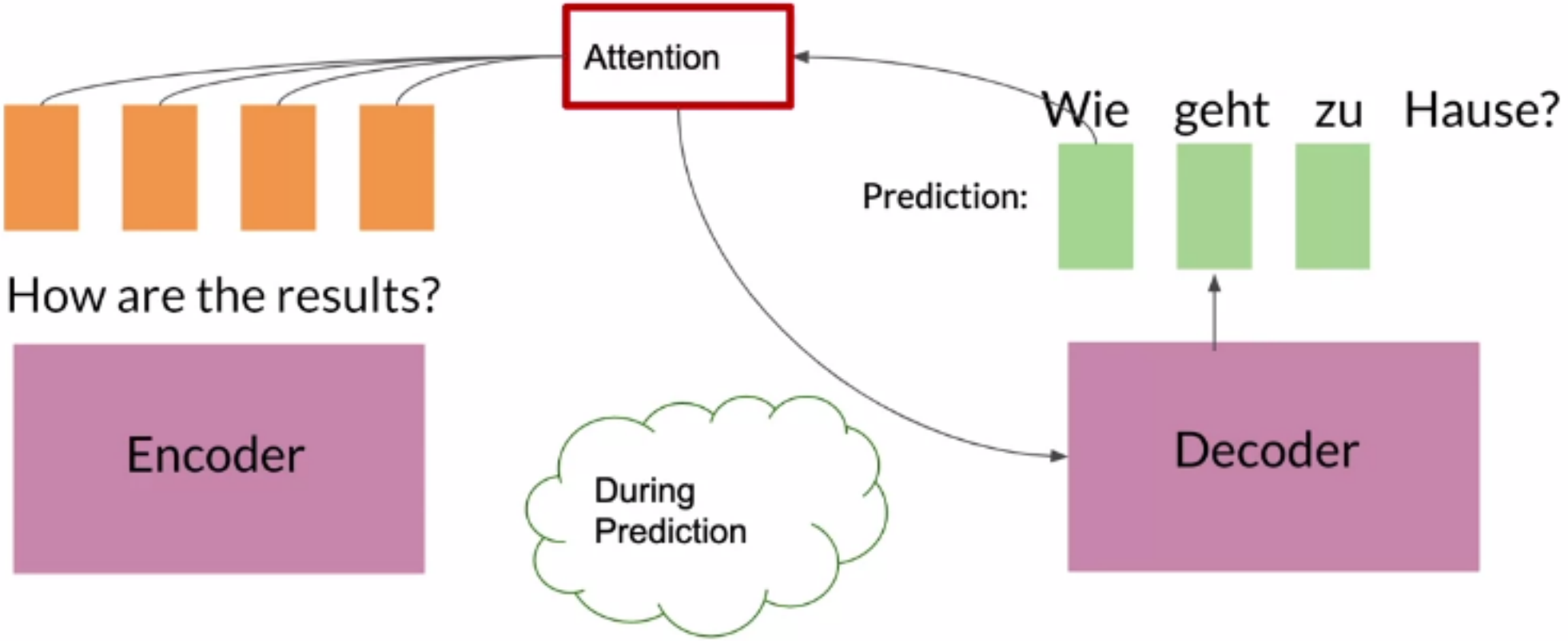
Prediction


Attention
Wie geht zu Hause?
Prediction:
How are the results?
Encoder
During Prediction
Decoder

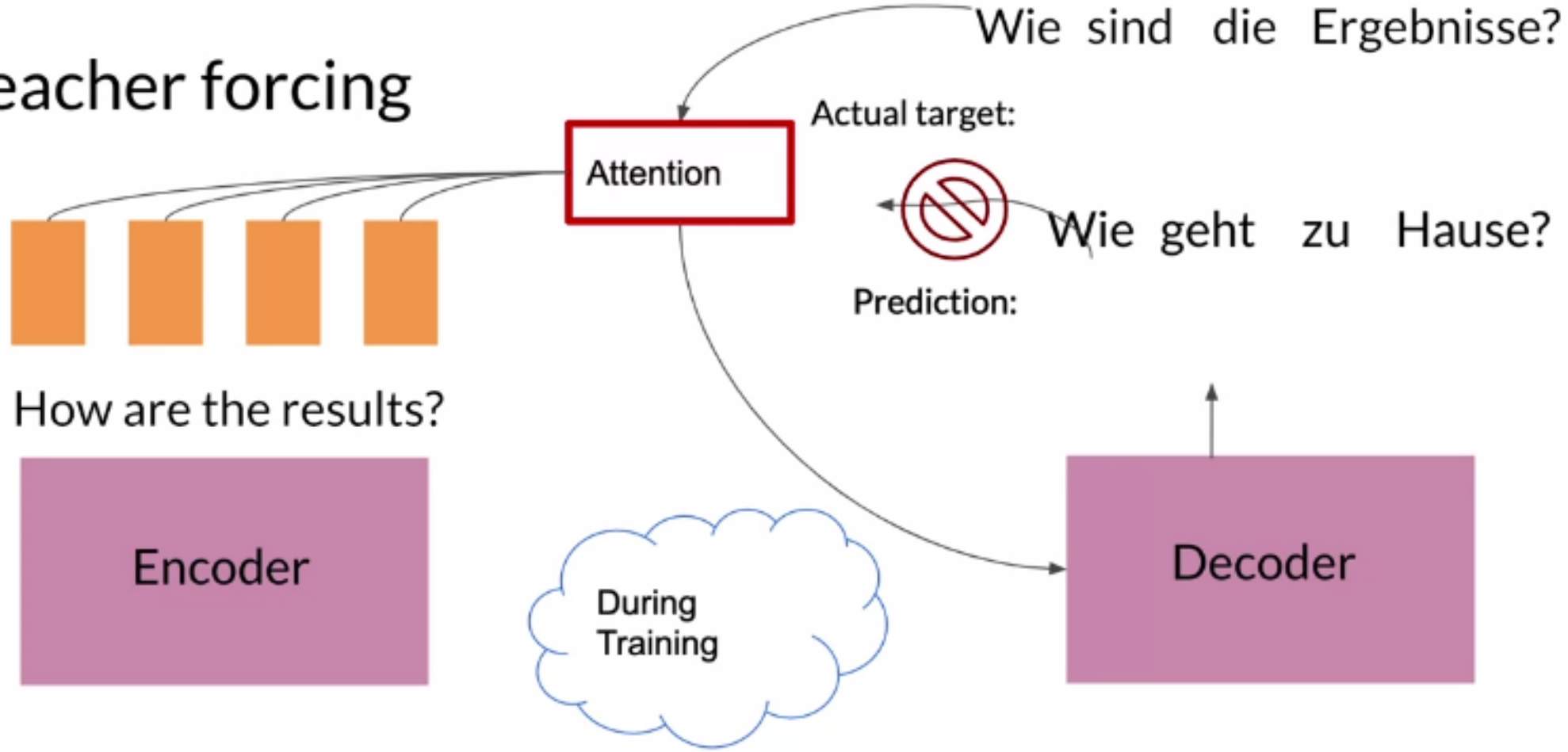
Prediction


Attention
Wie geht zu Hause?
Prediction:
How are the results?
Encoder
During Prediction
Decoder

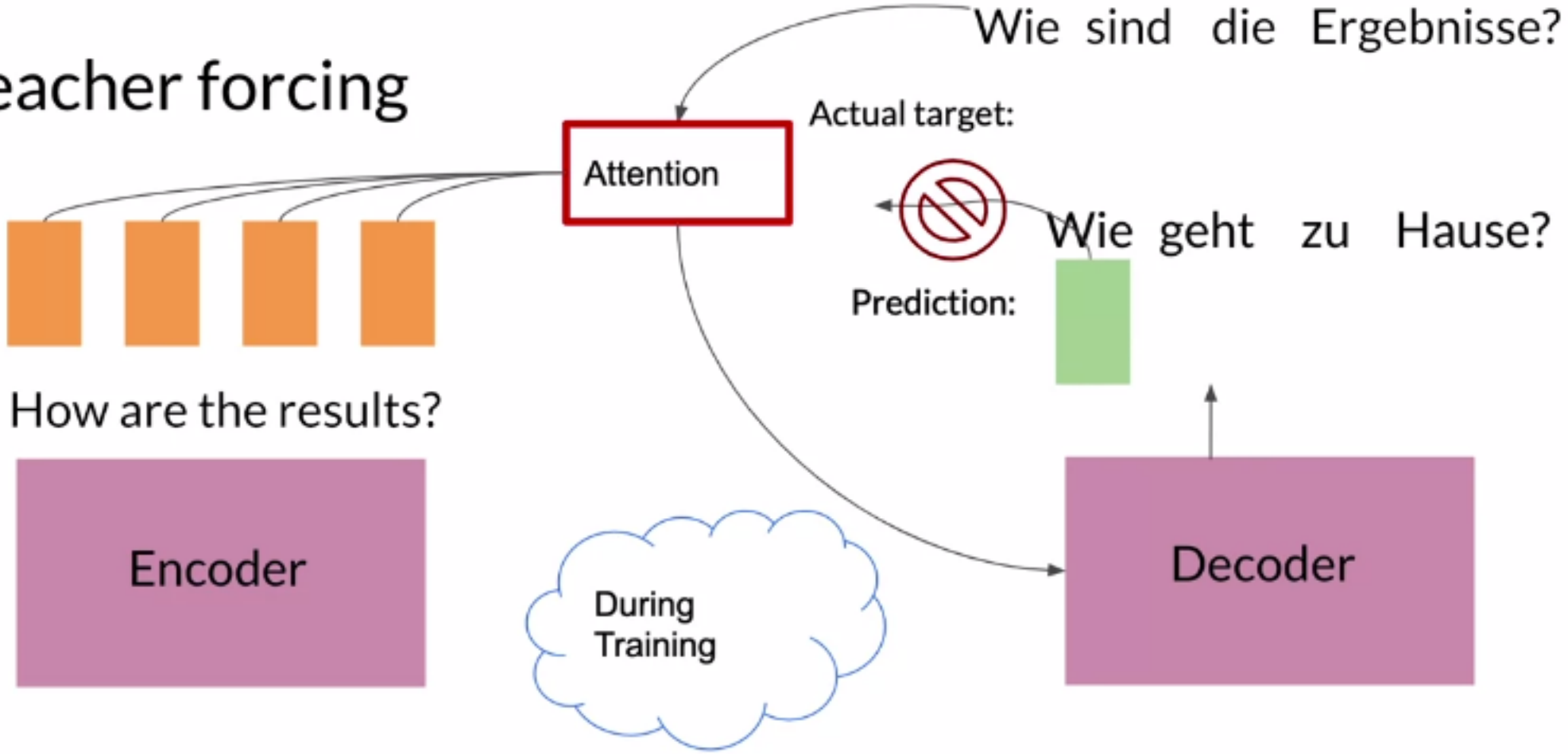
Prediction


Attention
Wie geht zu Hause?
Prediction:
How are the results?
Encoder
During Prediction
Decoder

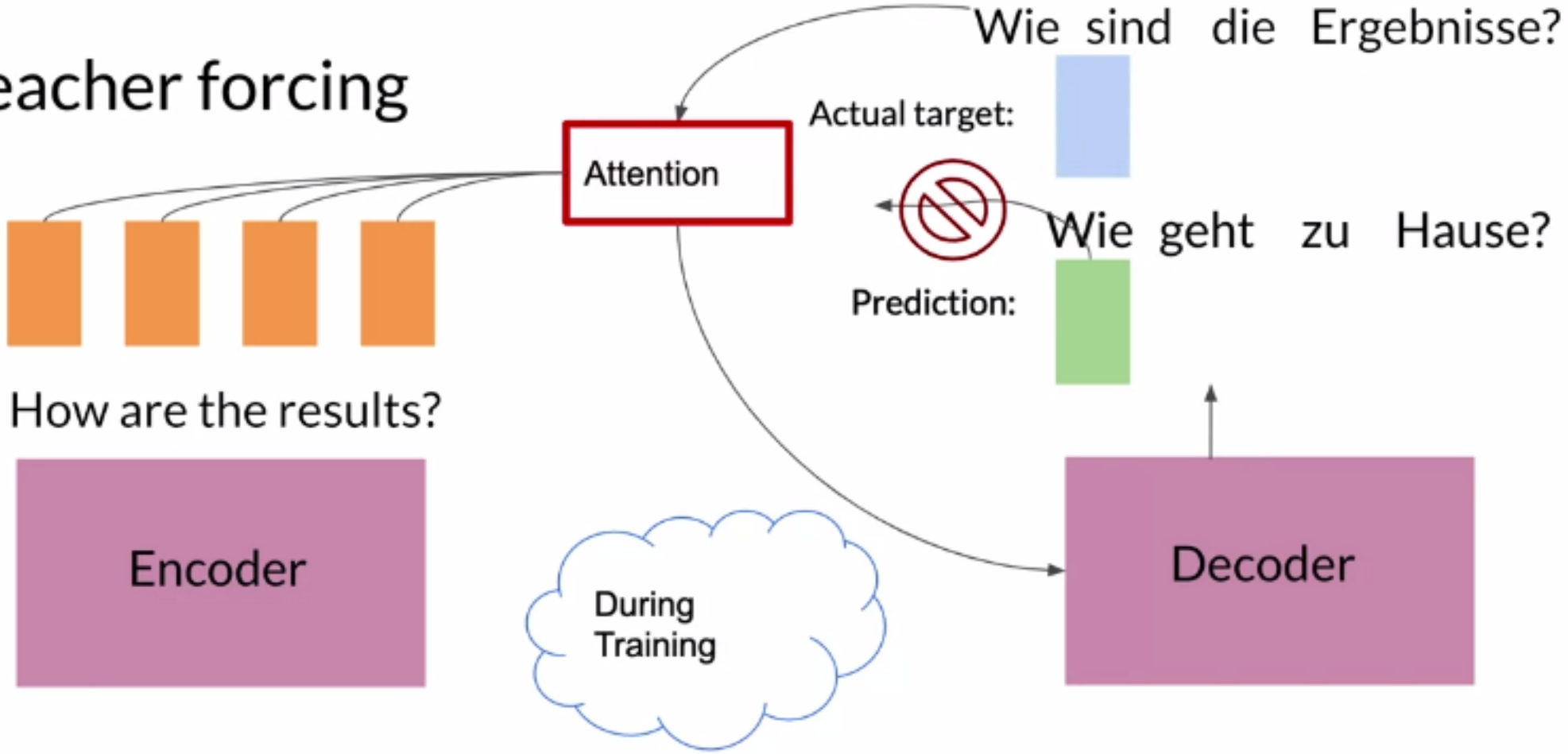
Teacher forcing



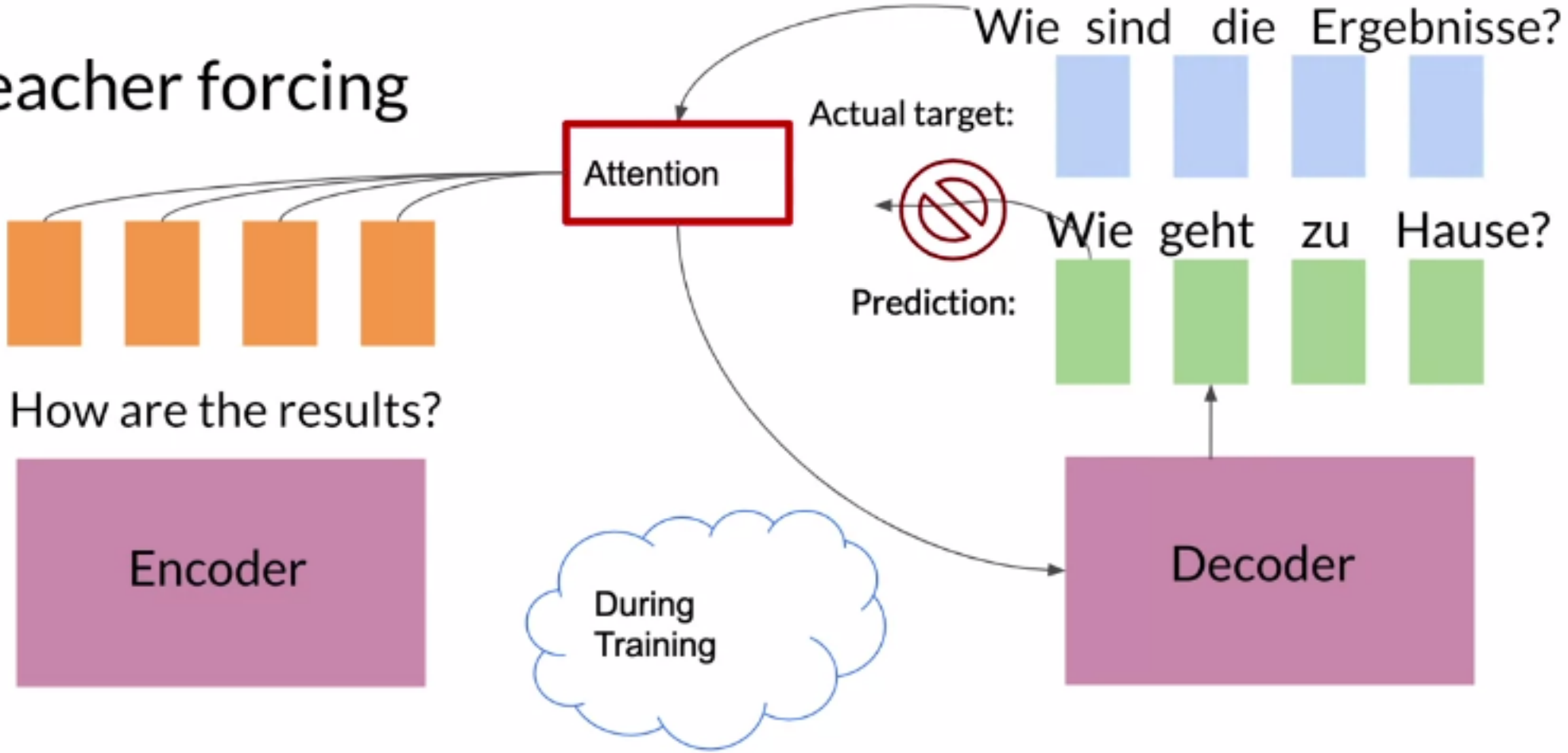
Teacher forcing



Teacher forcing

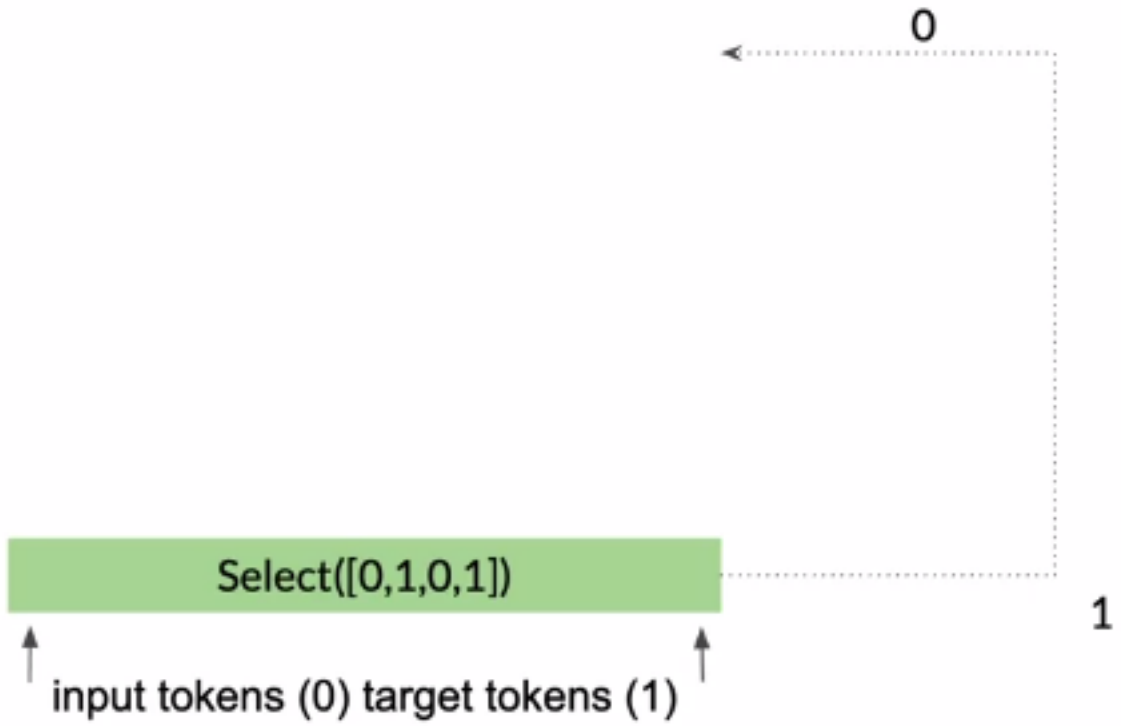


Teacher forcing

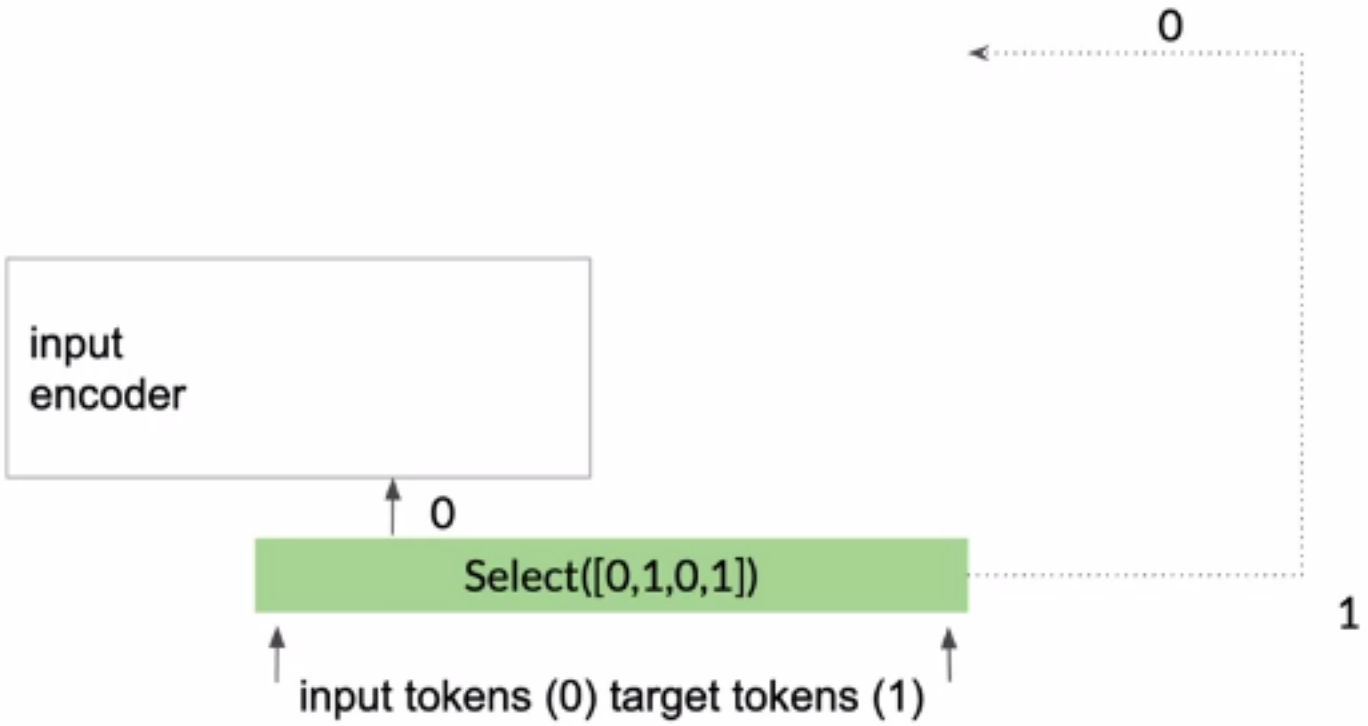


Training NMT

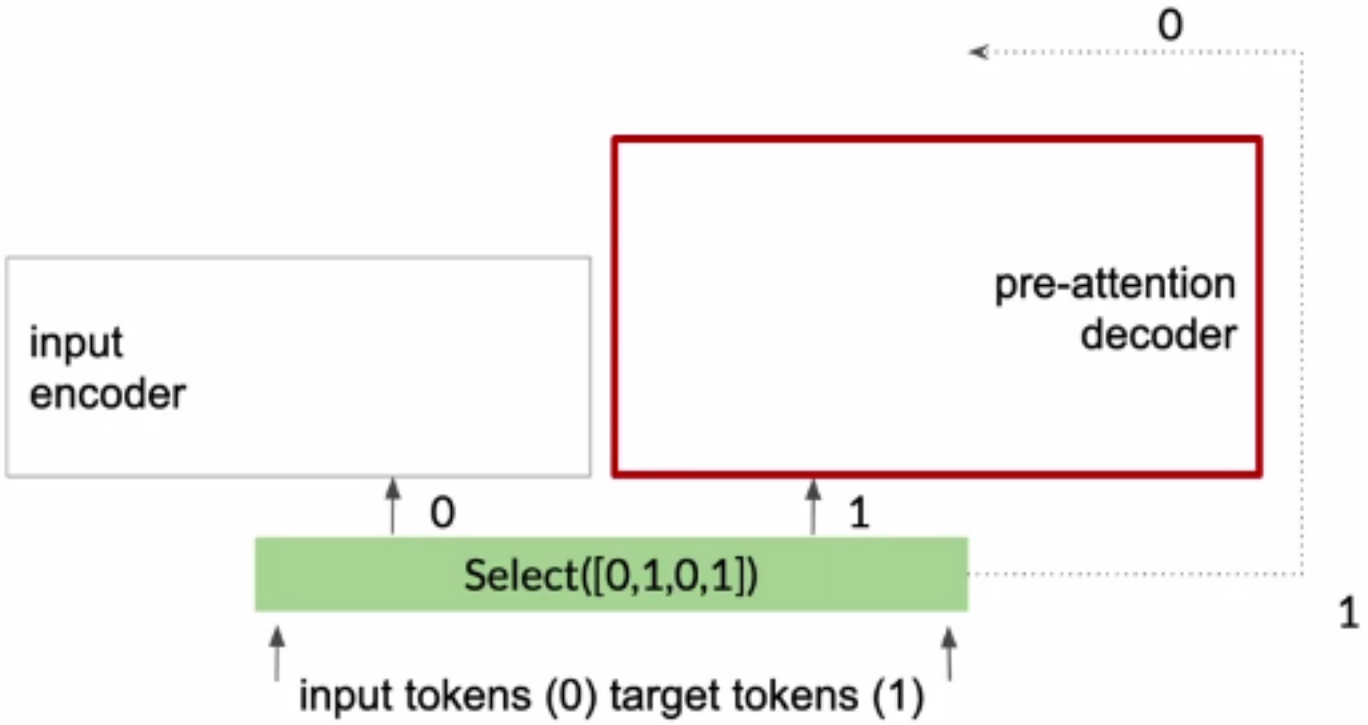
Training NMT



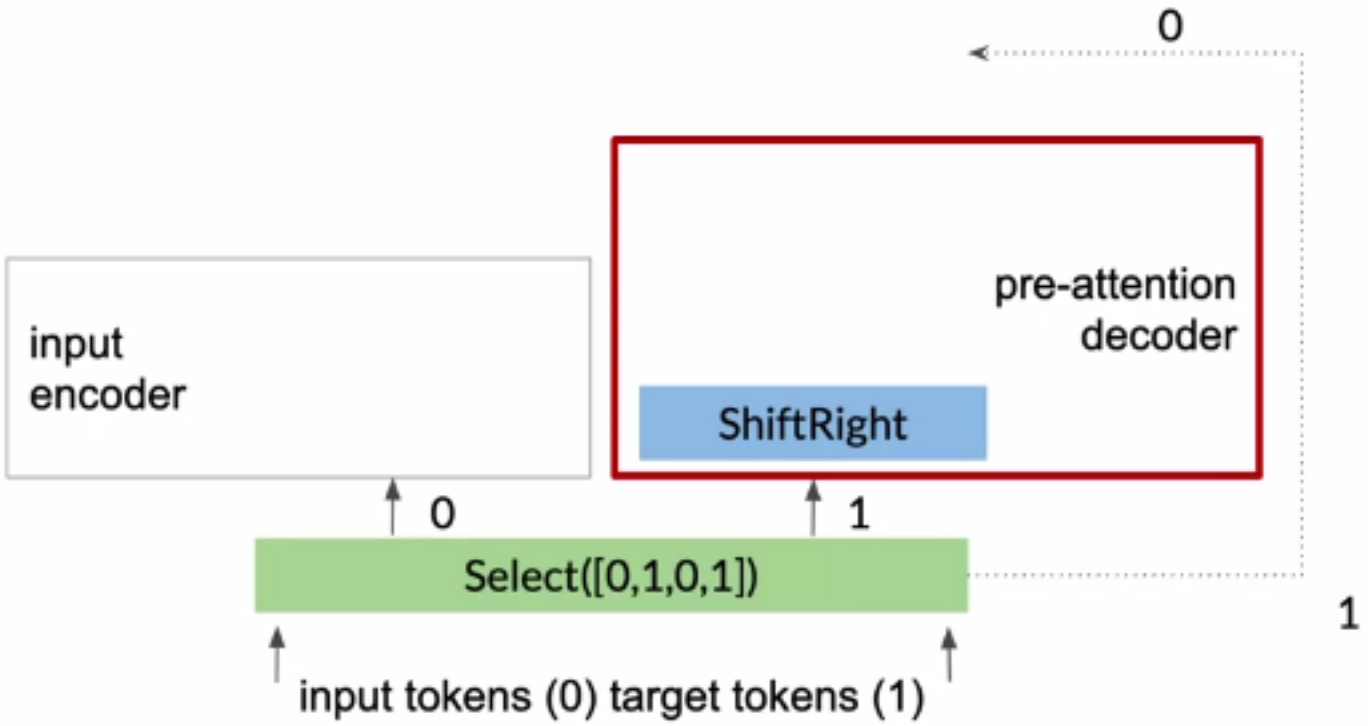
Training NMT



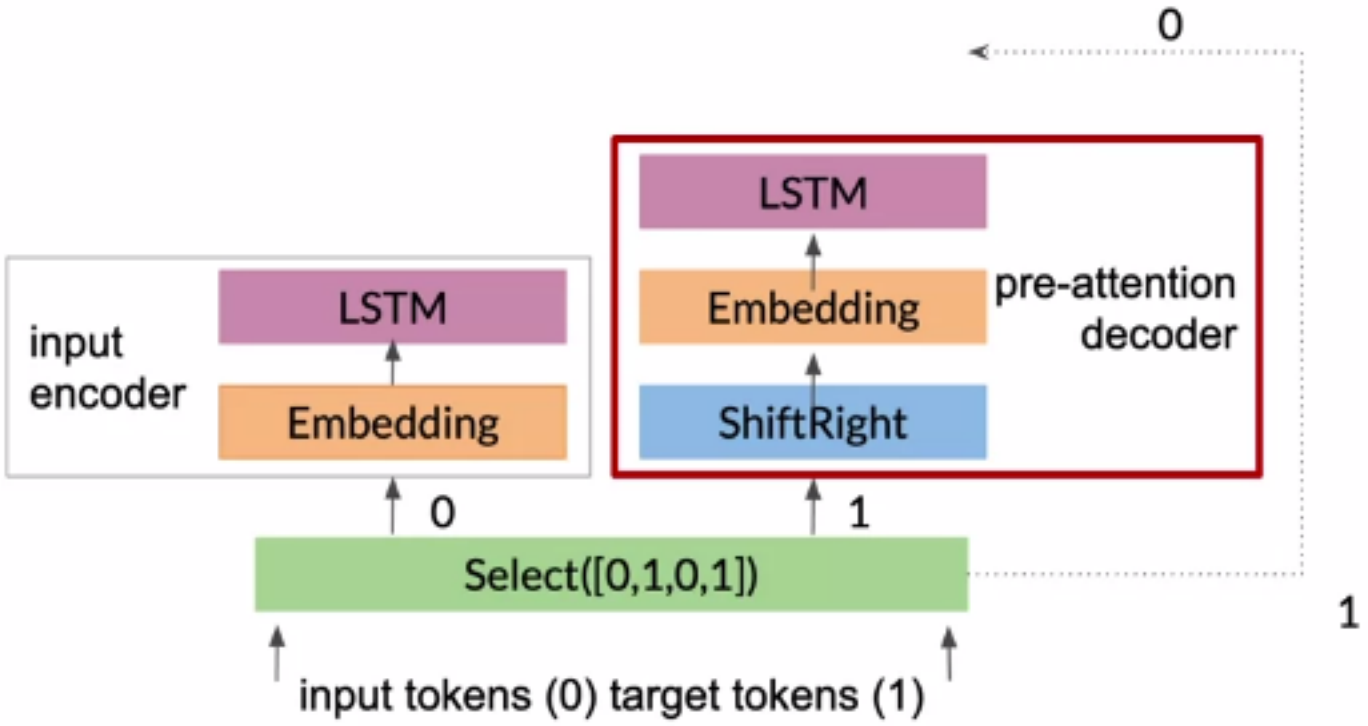
Training NMT



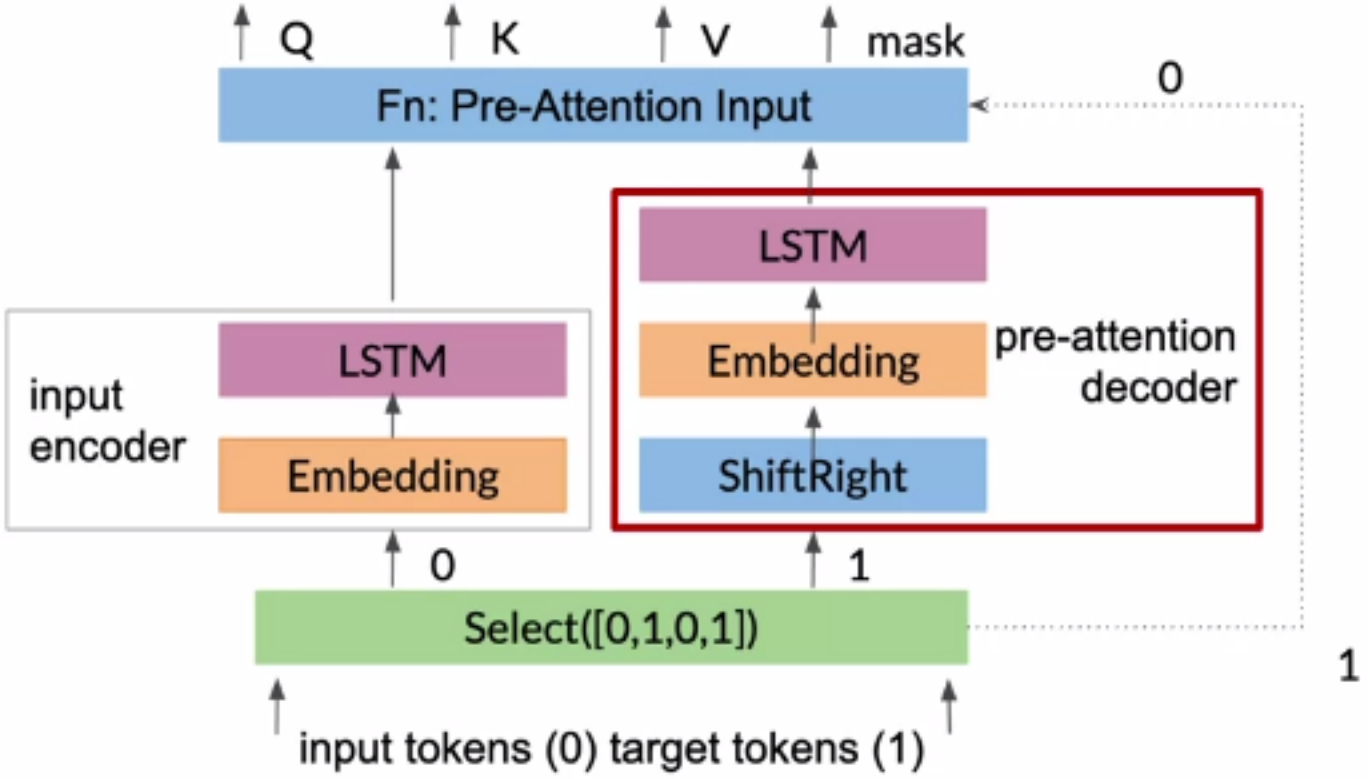
Training NMT



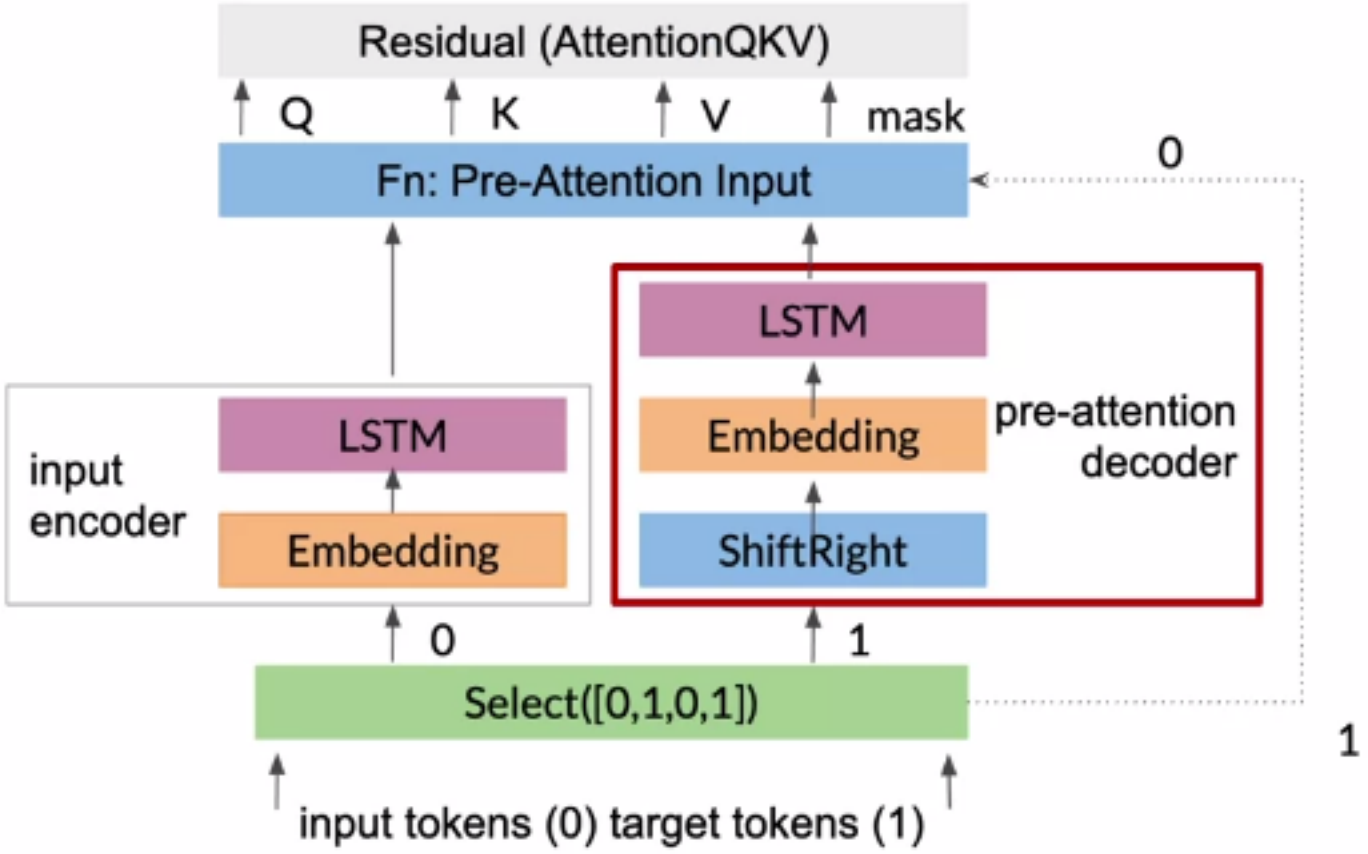
Training NMT



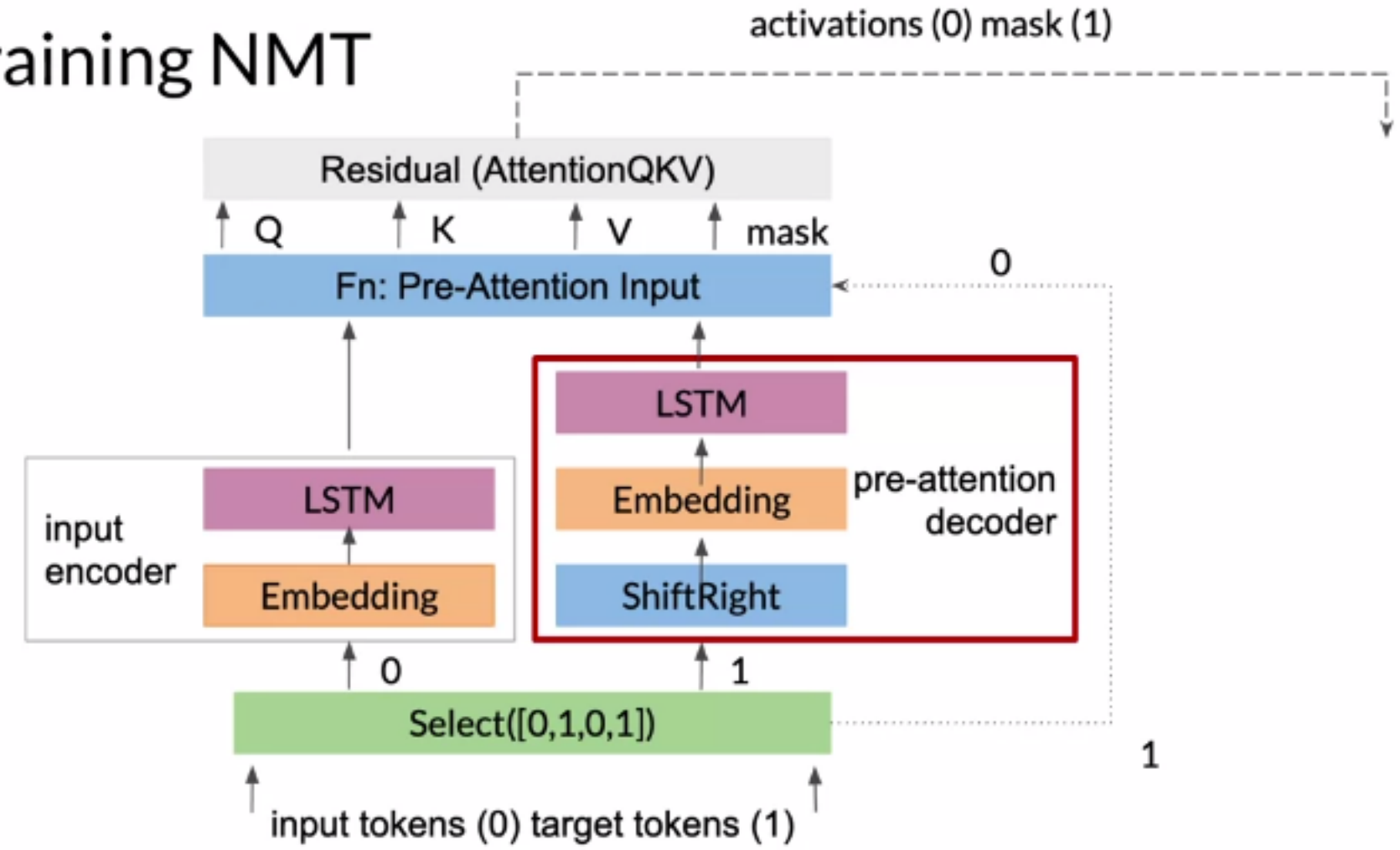
Training NMT



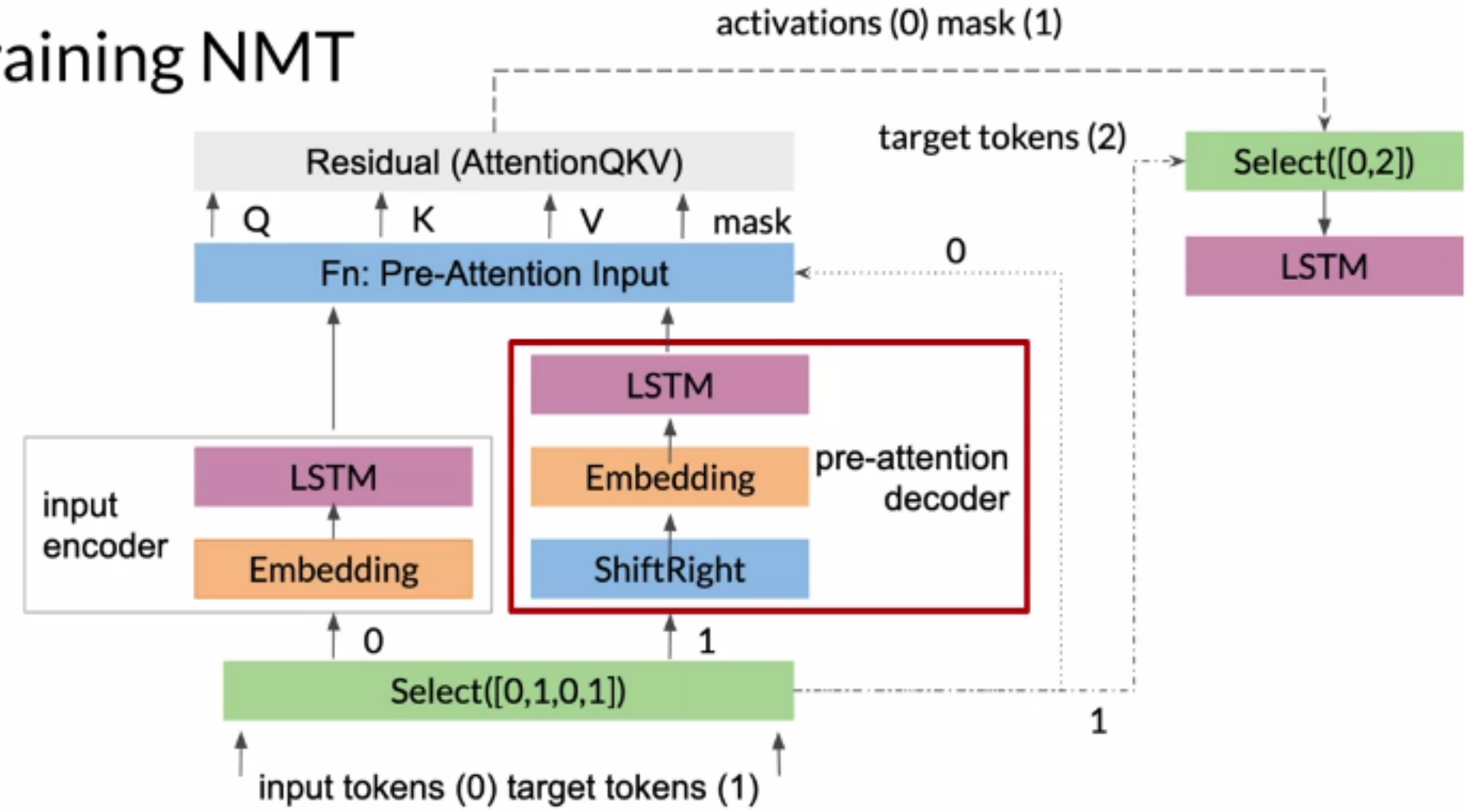
Training NMT



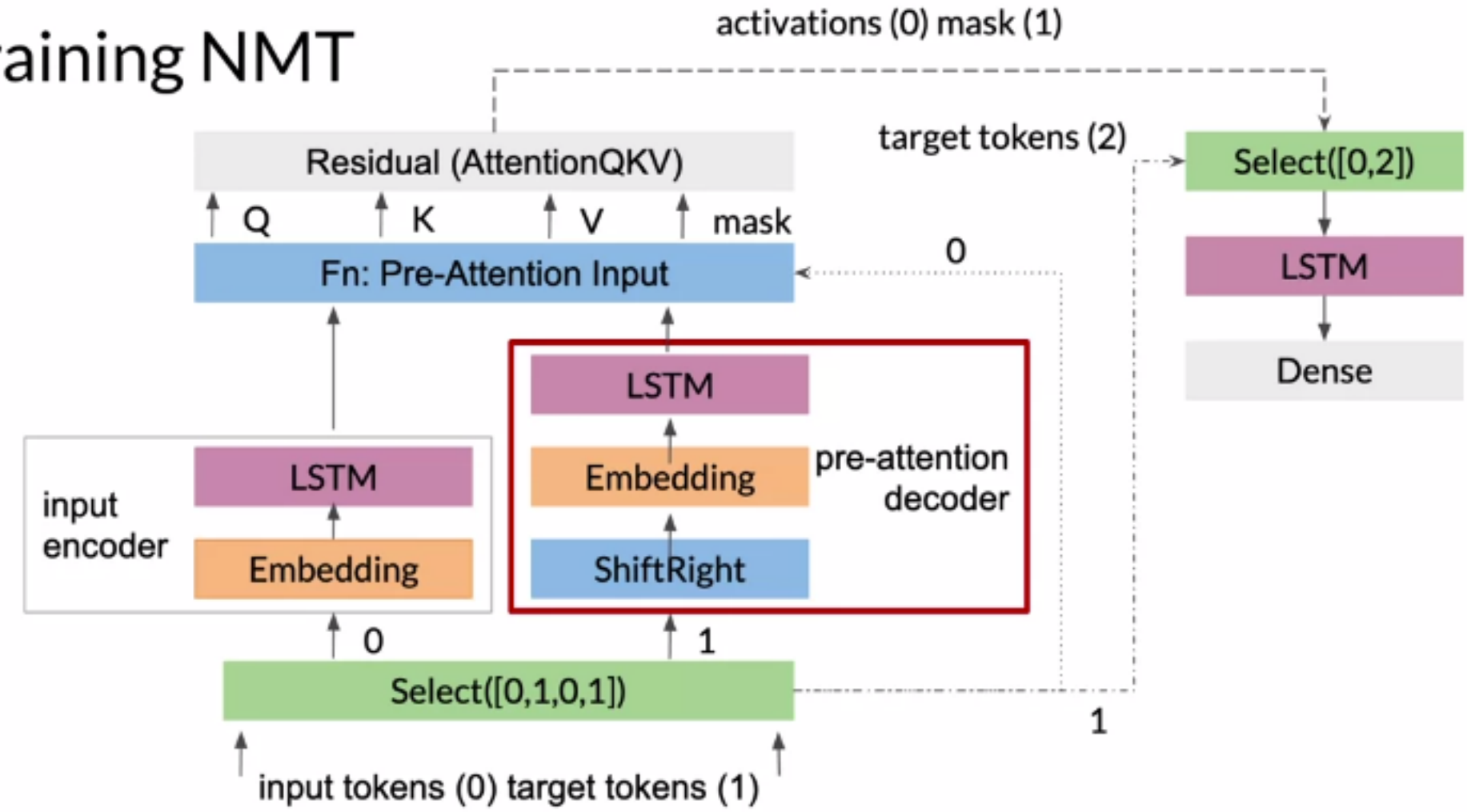
Training NMT



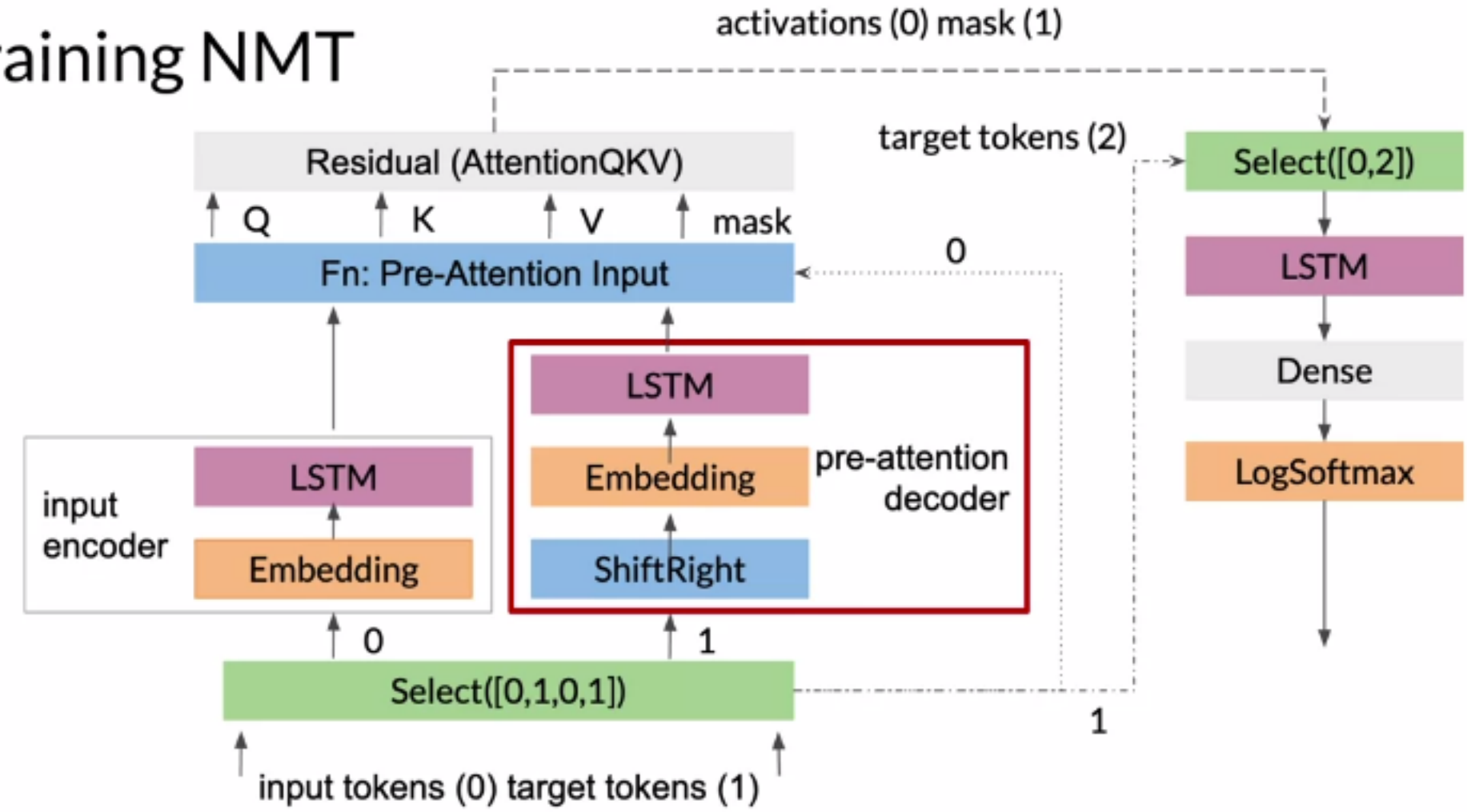
Training NMT



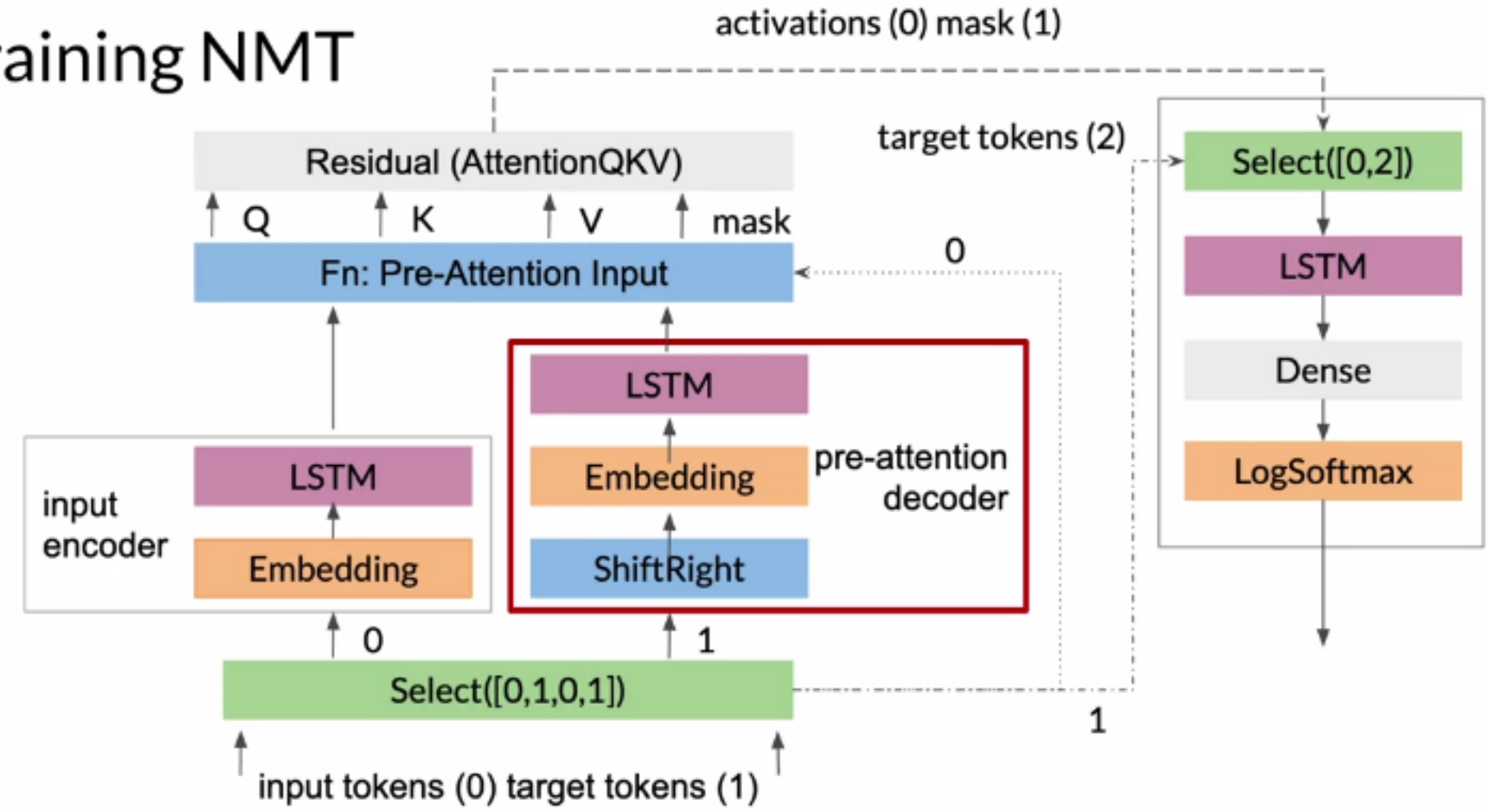
Training NMT



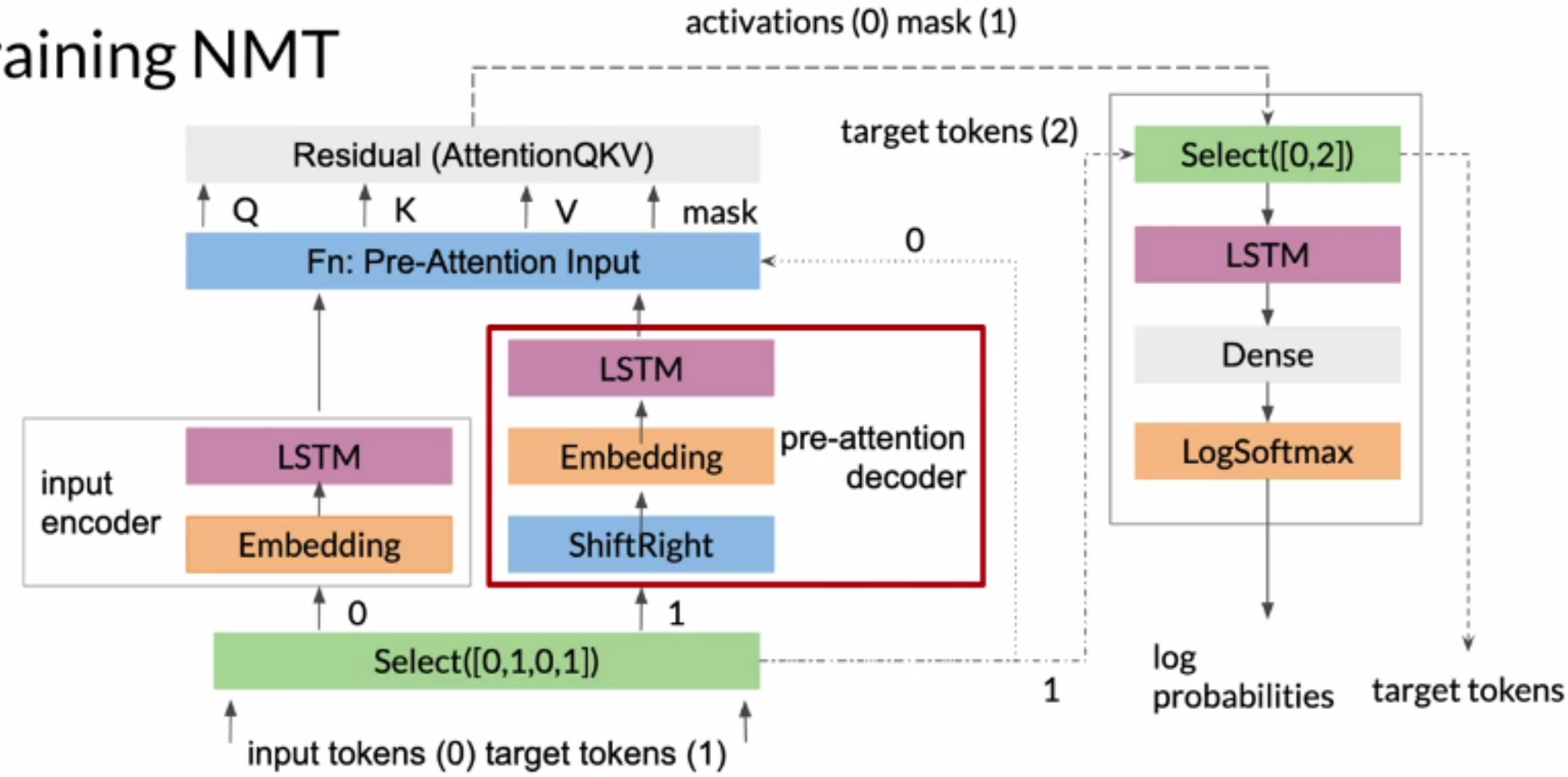
Training NMT



Training NMT



Training NMT



BLEU Score

Stands for Bilingual Evaluation Understudy

BLEU Score

Stands for Bilingual Evaluation Understudy

Evaluates the quality of machine-translated text by comparing “candidate” text to one or more “reference” translations.

BLEU Score

Stands for Bilingual Evaluation Understudy

Evaluates the quality of machine-translated text by comparing “candidate” text to one or more “reference” translations.

Scores: the closer to 1, the better, and vice versa:

BLEU Score

Candidate	I	I	am	I	I
Reference 1	Younes	said	I	am	hungry
Reference 2	He	said	I	am	hungry

BLEU Score

Candidate	I	I	am	I	I
Reference 1	Younes	said	I	am	hungry
Reference 2	He	said	I	am	hungry

BLEU Score

Candidate	I	I	am	I	I
Reference 1	Younes	said	I	am	hungry
Reference 2	He	said	I	am	hungry

BLEU Score

Candidate	I	I	am	I	I
Reference 1	Younes	said	I	am	hungry
Reference 2	He	said	I	am	hungry

BLEU Score

Candidate	I	I	am	I	I
Reference 1	Younes	said	I	am	hungry
Reference 2	He	said	I	am	hungry

How many words in the candidate column appear in the reference translations?

BLEU Score

Candidate	I	I	am	I	I
Reference 1	Younes	said	I	am	hungry
Reference 2	He	said	I	am	hungry

"I" appears at most once in both, so clip to one: $m_w = 1$

BLEU Score

Candidate	I	I	am	I	I
Reference 1	Younes	said	I	am	hungry
Reference 2	He	said	I	am	hungry

"I" appears at most once in both, so clip to one: $m_w = 1$

$$\frac{\text{(Sum over unique n-gram counts in the candidate)}}{\text{(total \# of words in candidate)}}$$

BLEU Score

2 divide by total (5) = 2/5

Candidate	I	I	am	I	I
Reference 1	Younes	said	I	am	hungry
Reference 2	He	said	I	am	hungry

"I" appears at most once in both, so clip to one: $m_w = 1$

$$\frac{\text{(Sum over unique n-gram counts in the candidate)}}{\text{(total \# of words in candidate)}}$$

BLEU score is great, but...

BLEU score is great, but...

Consider the following:

- BLEU doesn't consider semantic meaning

BLEU score is great, but...

Consider the following:

- BLEU doesn't consider semantic meaning
- BLEU doesn't consider sentence structure:

"Ate I was hungry because!"

ROUGE

Recall-Oriented Understudy for Gisting Evaluation

ROUGE

Recall-Oriented Understudy for Gisting Evaluation

Evaluates quality of machine text

ROUGE

Recall-Oriented Understudy for Gisting Evaluation

Evaluates quality of machine text

Measures precision and recall between generated text and human-created text

ROUGE evaluation

Model	The	cat	had	striped	orange	fur
Reference	The	cat	had	orange	fur	

ROUGE evaluation

Model	The	cat	had	striped	orange	fur
Reference	The	cat	had	orange	fur	

Recall = How much of the reference text is the system text capturing?

ROUGE evaluation

Model	The	cat	had	striped	orange	fur
Reference	The	cat	had	orange	fur	

Recall = How much of the reference text is the system text capturing?

Precision = How much of the model text was relevant?

Recall in ROUGE

Model	The	cat	had	striped	orange	fur
Reference	The	cat	had	orange	fur	

$$\frac{\text{(Sum of overlapping unigrams in model and reference)}}{\text{(total \# of words in reference)}}$$

Recall in ROUGE

Model	The	cat	had	striped	orange	fur
Reference	The	cat	had	orange	fur	

$$\frac{\text{(Sum of overlapping unigrams in model and reference)}}{\text{(total \# of words in reference)}} \quad \frac{5}{5}$$

Recall in ROUGE

Model	The	cat	had	striped	orange	fur
Reference	The	cat	had	orange	fur	

$$\frac{\text{(Sum of overlapping unigrams in model and reference)}}{\text{(total \# of words in reference)}} \qquad \frac{5}{5} \quad \text{Recall} = 1$$

Precision in ROUGE

Model	The	cat	had	striped	orange	fur
Reference	The	cat	had	orange	fur	

$$\frac{\text{(Sum of overlapping unigrams in model and reference)}}{\text{(total \# of words in model)}}$$

$$\frac{5}{6} \quad \text{Precision} = 0.83$$

Problems in ROUGE

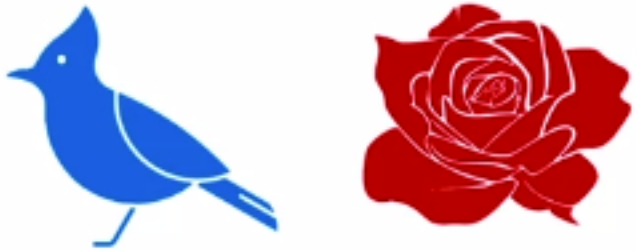
Problems in ROUGE

- Doesn't take themes or concepts into consideration (i.e., a low ROUGE score doesn't necessarily mean the translation is bad)

Model	I	am	a	fruit-filled	pastry
Reference	I	am	a	jelly	donut

Summary

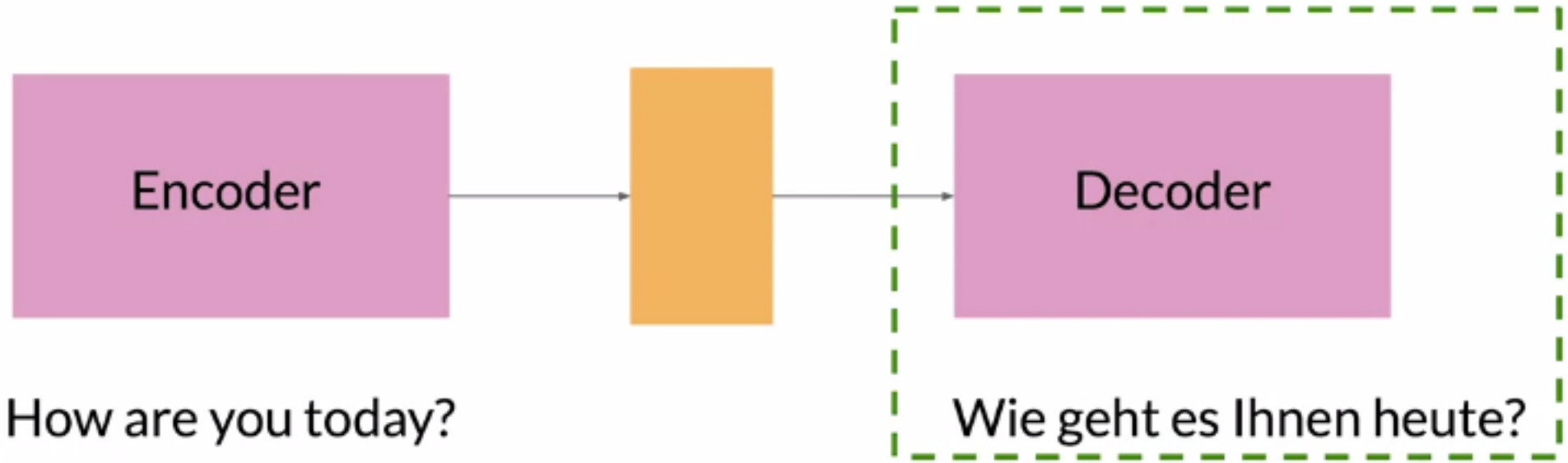
- BLEU score compares “candidate” against “references” using an n-gram average
- BLEU doesn’t consider meaning or structure
- ROUGE measures machine-generated text against an “ideal” reference



Outline

- Random sampling
- Temperature in sampling
- Greedy decoding
- Beam search
- Minimum Bayes' risk (MBR)

Seq2Seq model



Greedy decoding

Greedy decoding

Selects the most probable word at each step

Greedy decoding

Selects the most probable word at each step

But the best word at each step may not be the best for longer sequences...

Ich habe Hunger.

I am hungry .

I am, am, am, am...

Random sampling

am	full	hungry	I	the
0.05	0.3	0.15	0.25	0.25

Random sampling

am	full	hungry	I	the
0.05	0.3	0.15	0.25	0.25

Often a little too random for accurate translation!

Random sampling

am	full	hungry	I	the
0.05	0.3	0.15	0.25	0.25

Often a little too random for accurate translation!

Solution: Assign more weight to more probable words, and less weight to less probable words.

Temperature

In sampling, temperature is a parameter allowing for more or less randomness in predictions

Temperature

In sampling, temperature is a parameter allowing for more or less randomness in predictions

Lower temperature setting = More confident, conservative network

Temperature

In sampling, temperature is a parameter allowing for more or less randomness in predictions

Lower temperature setting = More confident, conservative network

Higher temperature setting = More excited, random network (and more mistakes)

Beam search decoding

Beam search decoding

A broader, more exploratory decoding alternative

Beam search decoding

A broader, more exploratory decoding alternative

Selects multiple options for the best input based on conditional probability

Beam search decoding

A broader, more exploratory decoding alternative

Selects multiple options for the best input based on conditional probability

Number of options depends on a predetermined beam width parameter B

Beam search decoding

A broader, more exploratory decoding alternative

Selects multiple options for the best input based on conditional probability

Number of options depends on a predetermined beam width parameter B

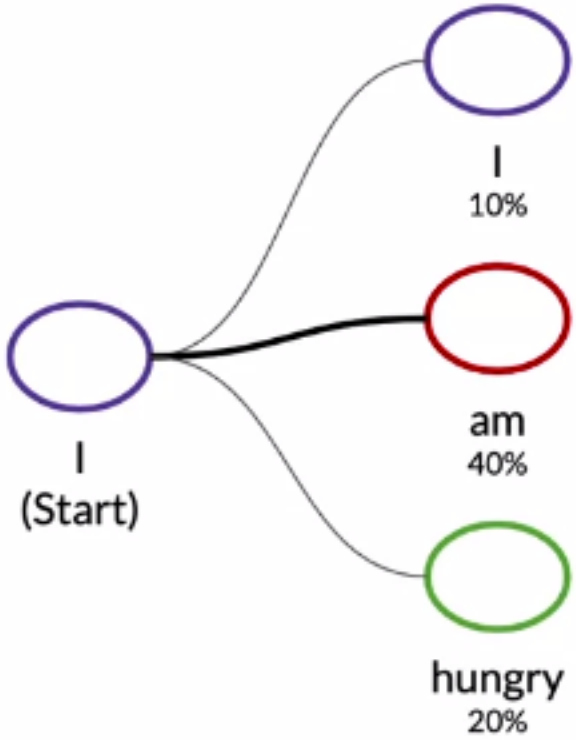
Selects B number of best alternatives at each time step

Beam search example

I
(Start)

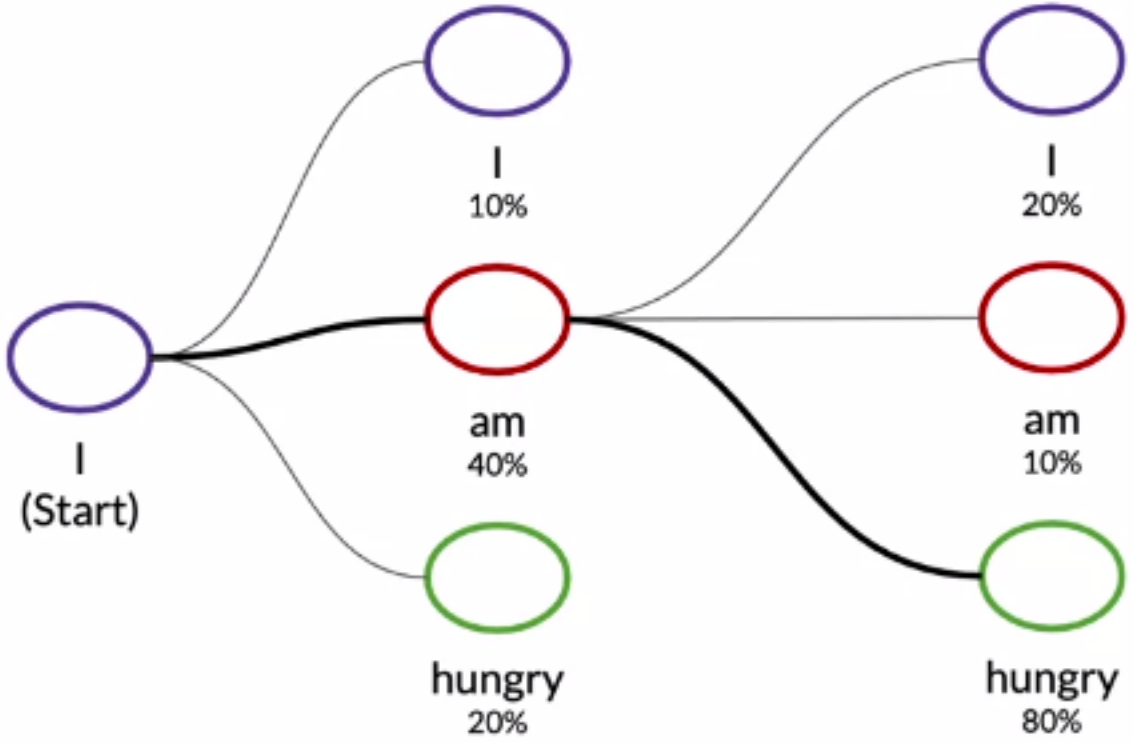
B = 3

Beam search example



B = 3

Beam search example



B = 3

Problems with beam search

Since the model learns a distribution, that tends to carry more weight than single tokens

Problems with beam search

Since the model learns a distribution, that tends to carry more weight than single tokens

Can cause translation problems, i.e. in a speech corpus that hasn't been cleaned

Problems with beam search

“Ich mag die Vereinigten Staaten, weil die Vereinigten Staaten groß sind.”

Problems with beam search

“Ich mag die Vereinigten Staaten, weil die Vereinigten Staaten groß sind.”

Even with 11 good English translations of “Vereinigten Staaten,” but a ~1% probability of the non-word “Uhm” occurring, you might get this as a translation:

“I like the United States, because the Uhm is big. ”

Problems with beam search

"Ich mag die Vereinigten Staaten, weil die Vereinigten Staaten groß sind."

Even with 11 good English translations of "Vereinigten Staaten," but a ~1% probability of the non-word "Uhm" occurring, you might get this as a translation:

"I like the United States, because the Uhm is big. "

Even with 11^2 good translations, the most probable one will still be "Uhm."

Minimum Bayes Risk (MBR)

Compares many samples against one another. To implement MBR:

- Generate several random samples

Minimum Bayes Risk (MBR)

Compares many samples against one another. To implement MBR:

- Generate several random samples
- Compare each sample against all the others and assign a similarity score (such as ROUGE!)

Example: MBR Sampling

To generate the scores for 4 samples:

Example: MBR Sampling

To generate the scores for 4 samples:

1. Calculate similarity score between sample 1 and sample 2

Example: MBR Sampling

To generate the scores for 4 samples:

1. Calculate similarity score between sample 1 and sample 2
2. Calculate similarity score between sample 1 and sample 3

Example: MBR Sampling

To generate the scores for 4 samples:

1. Calculate similarity score between sample 1 and sample 2
2. Calculate similarity score between sample 1 and sample 3
3. Calculate similarity score between sample 1 and sample 4

Example: MBR Sampling

To generate the scores for 4 samples:

1. Calculate similarity score between sample 1 and sample 2
2. Calculate similarity score between sample 1 and sample 3
3. Calculate similarity score between sample 1 and sample 4
4. Average the score of the first 3 steps (Usually a weighted average)

Example: MBR Sampling

To generate the scores for 4 samples:

1. Calculate similarity score between sample 1 and sample 2
2. Calculate similarity score between sample 1 and sample 3
3. Calculate similarity score between sample 1 and sample 4
4. Average the score of the first 3 steps (Usually a weighted average)
5. Repeat until all samples have overall scores

Summary

- Beam search uses conditional probabilities and the beam width parameter
- MBR (Minimum Bayes Risk)takes several samples and compares them against each other to find the golden one ✨
- Go forth to the coding assignment!