

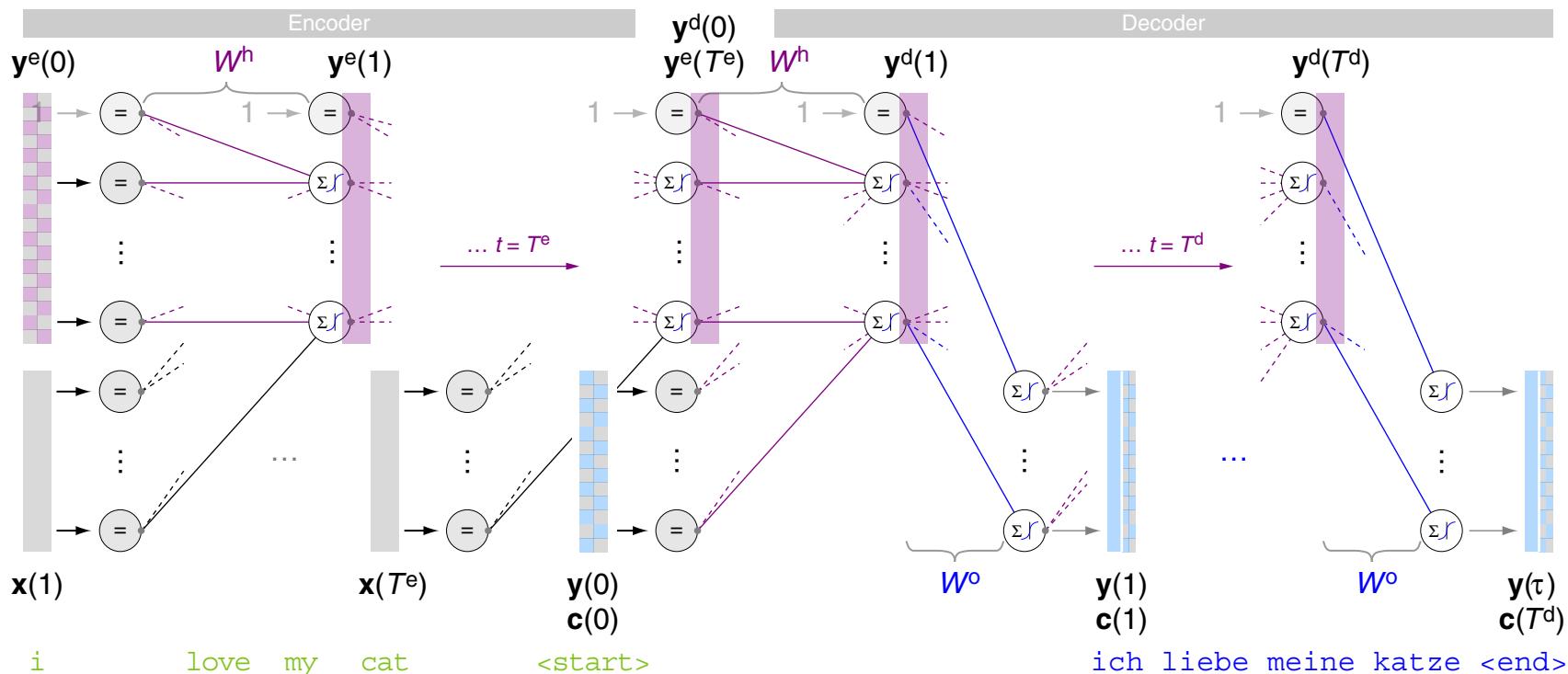
## IX. Deep Learning

- ❑ Elements of Deep Learning
- ❑ Convolutional Neural Networks
- ❑ Autoencoder Networks
- ❑ Recurrent Neural Networks
- ❑ Long-Term Dependencies
- ❑ RNNs for Machine Translation
- ❑ Attention Mechanism
- ❑ Self Attention and Transformers
- ❑ Transformer Language Models

# Attention Mechanism

## Notation III (language model view)

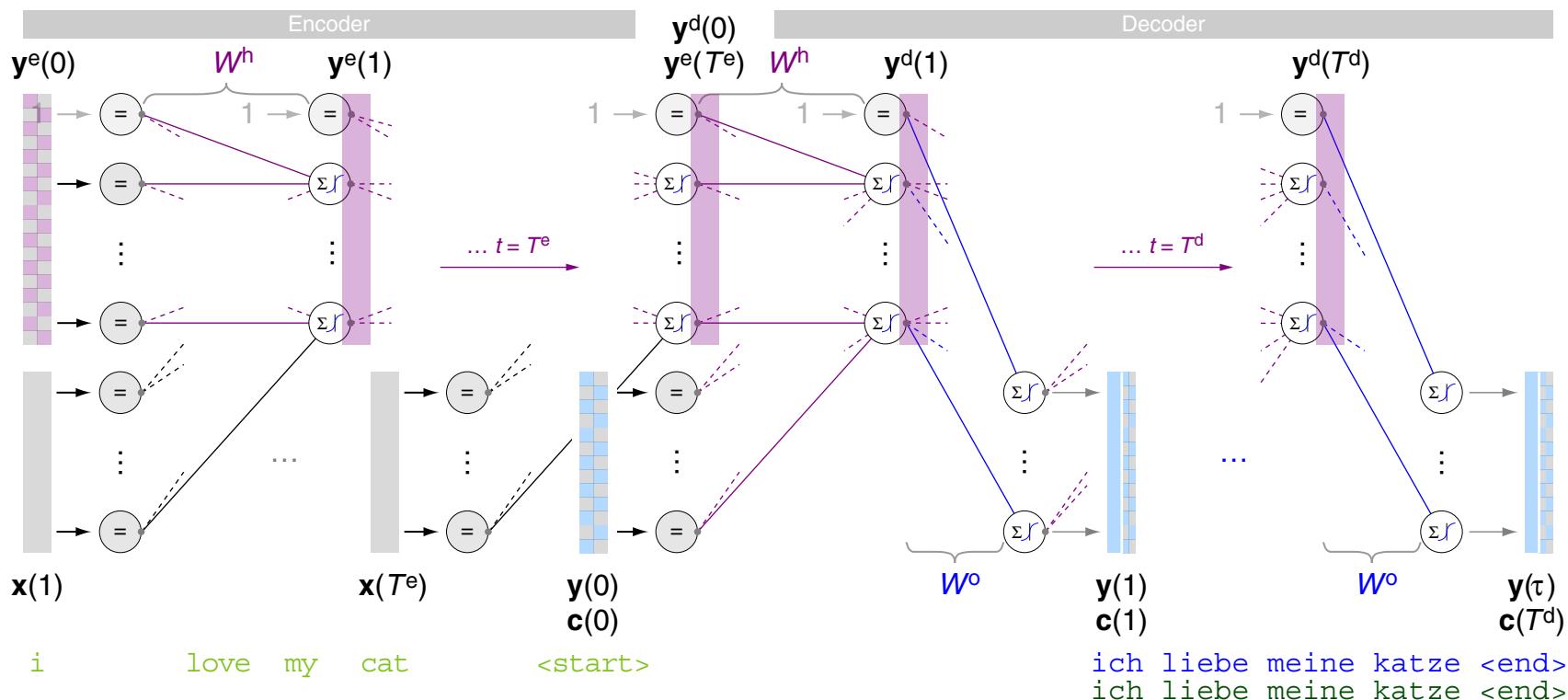
[notation: color, graph, language]



# Attention Mechanism

## Notation III (language model view)

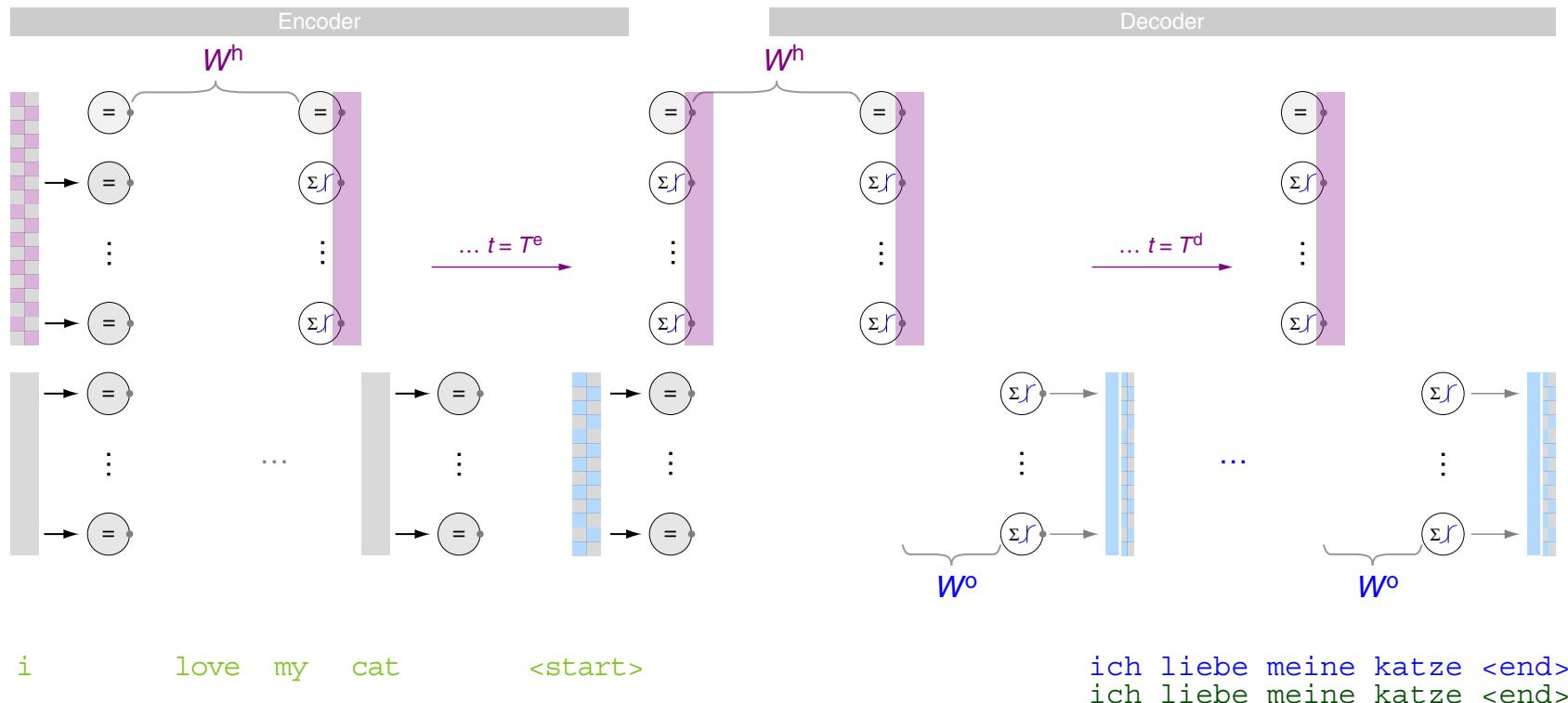
[notation: color, graph, language]



# Attention Mechanism

## Notation III (language model view)

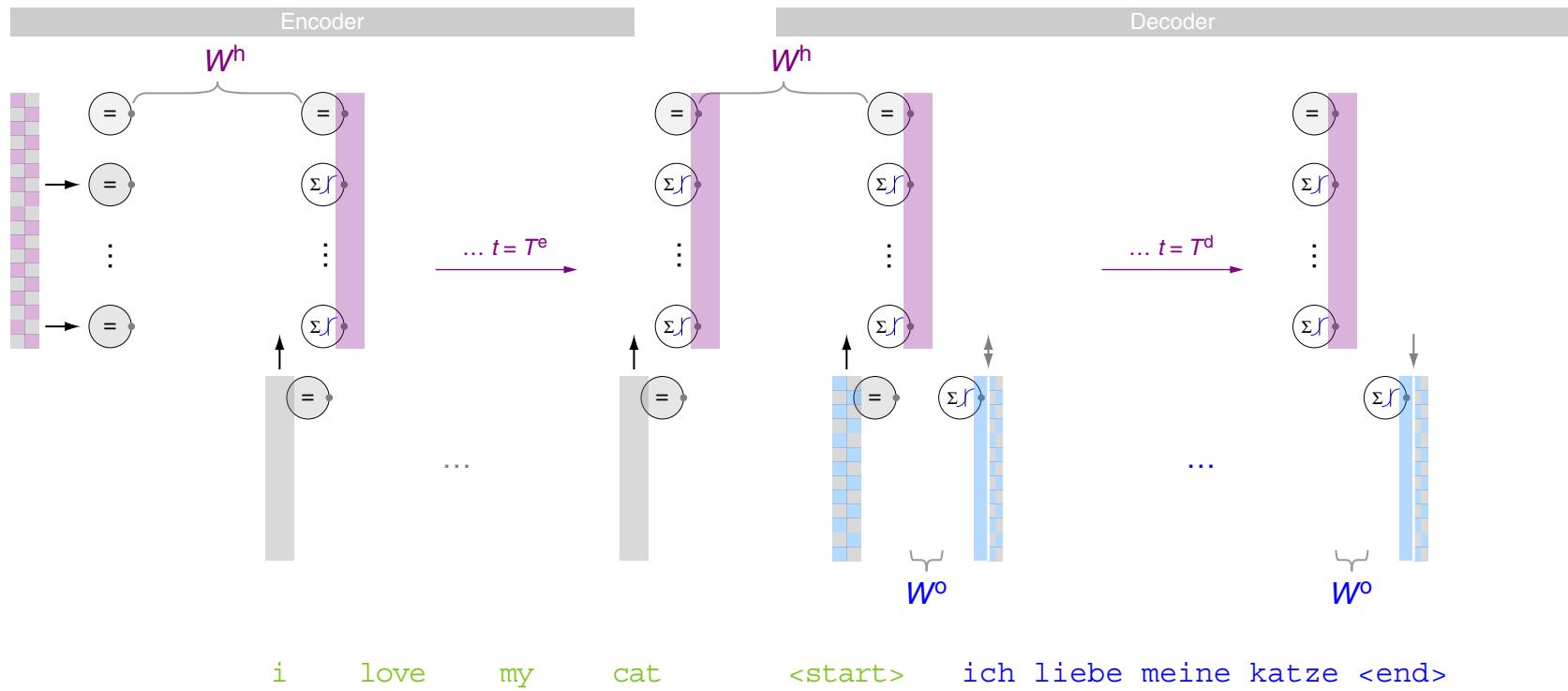
[notation: color, graph, language]



# Attention Mechanism

## Notation III (language model view)

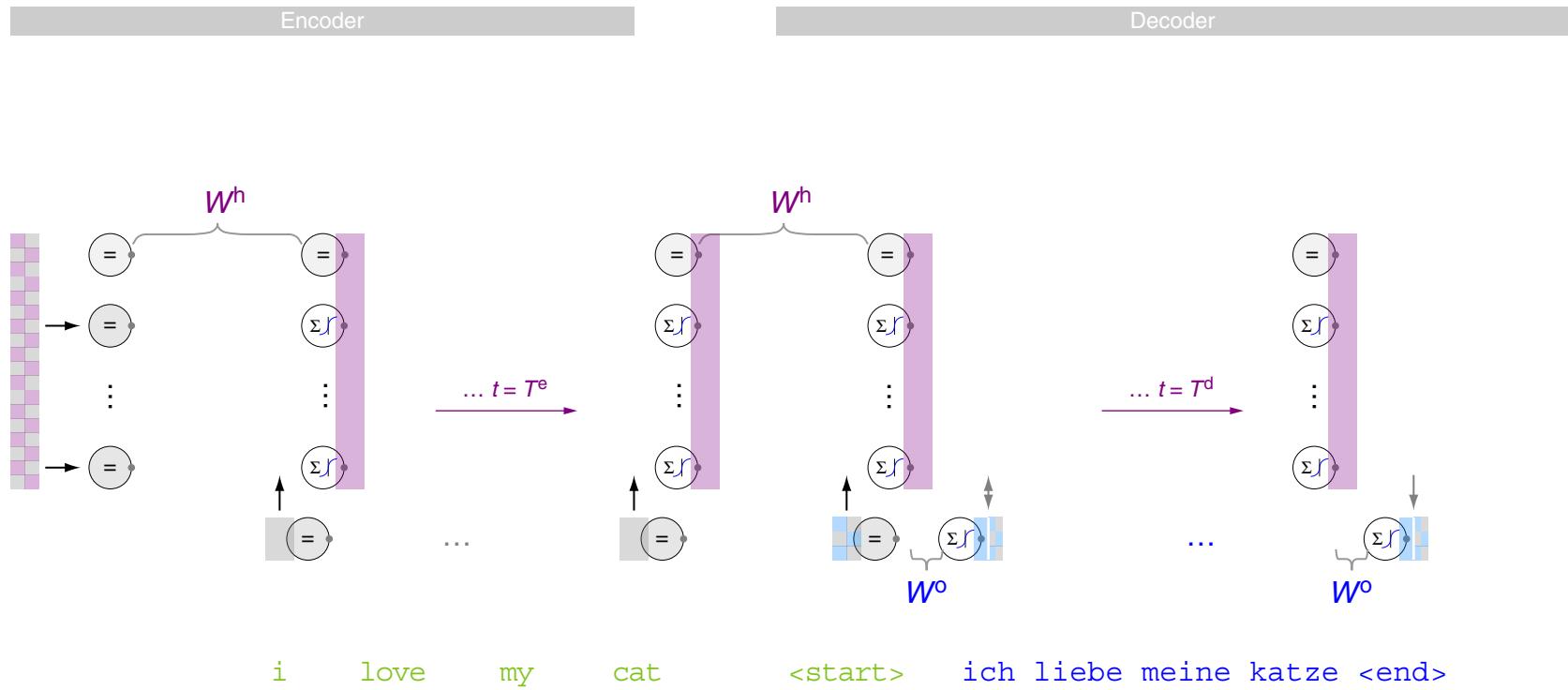
[notation: color, graph, language]



# Attention Mechanism

Notation III (language model view)

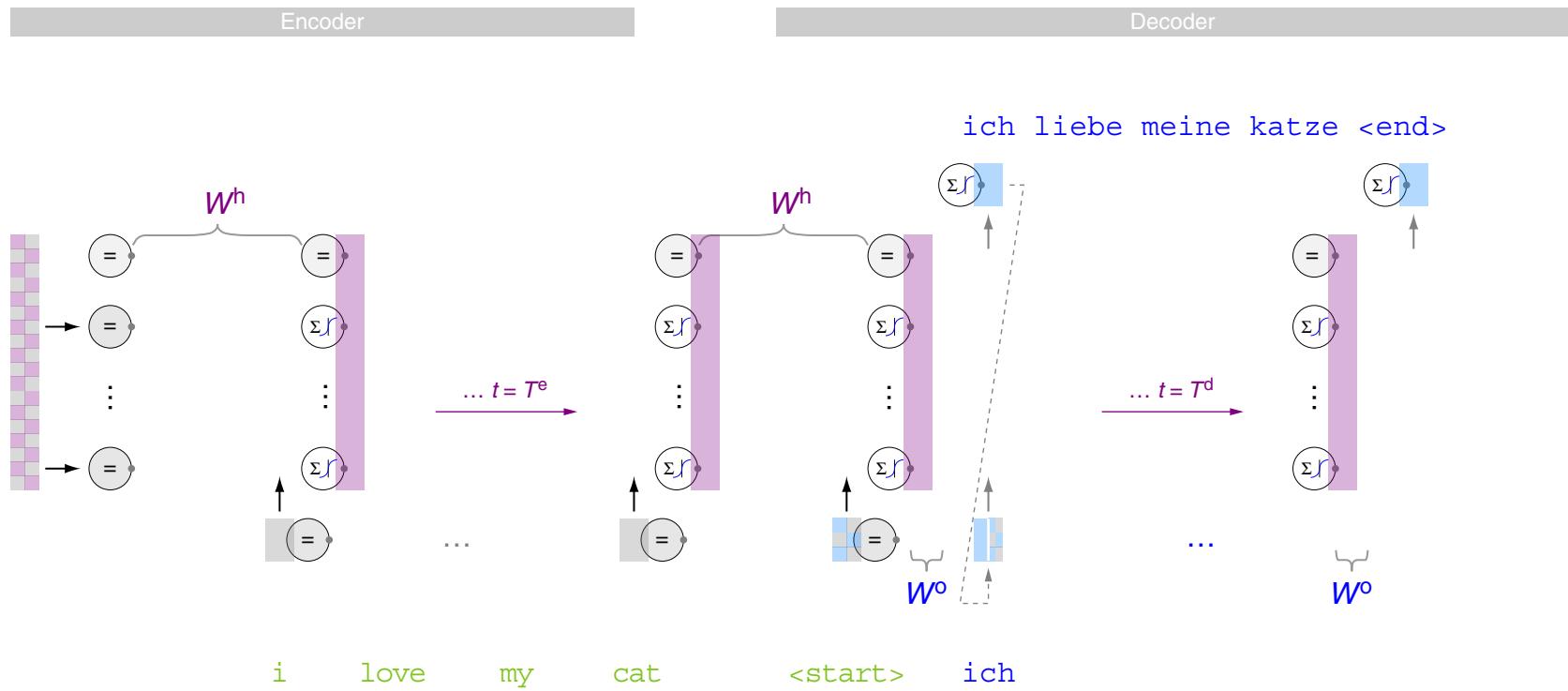
[notation: color, graph, language]



# Attention Mechanism

Notation III (language model view)

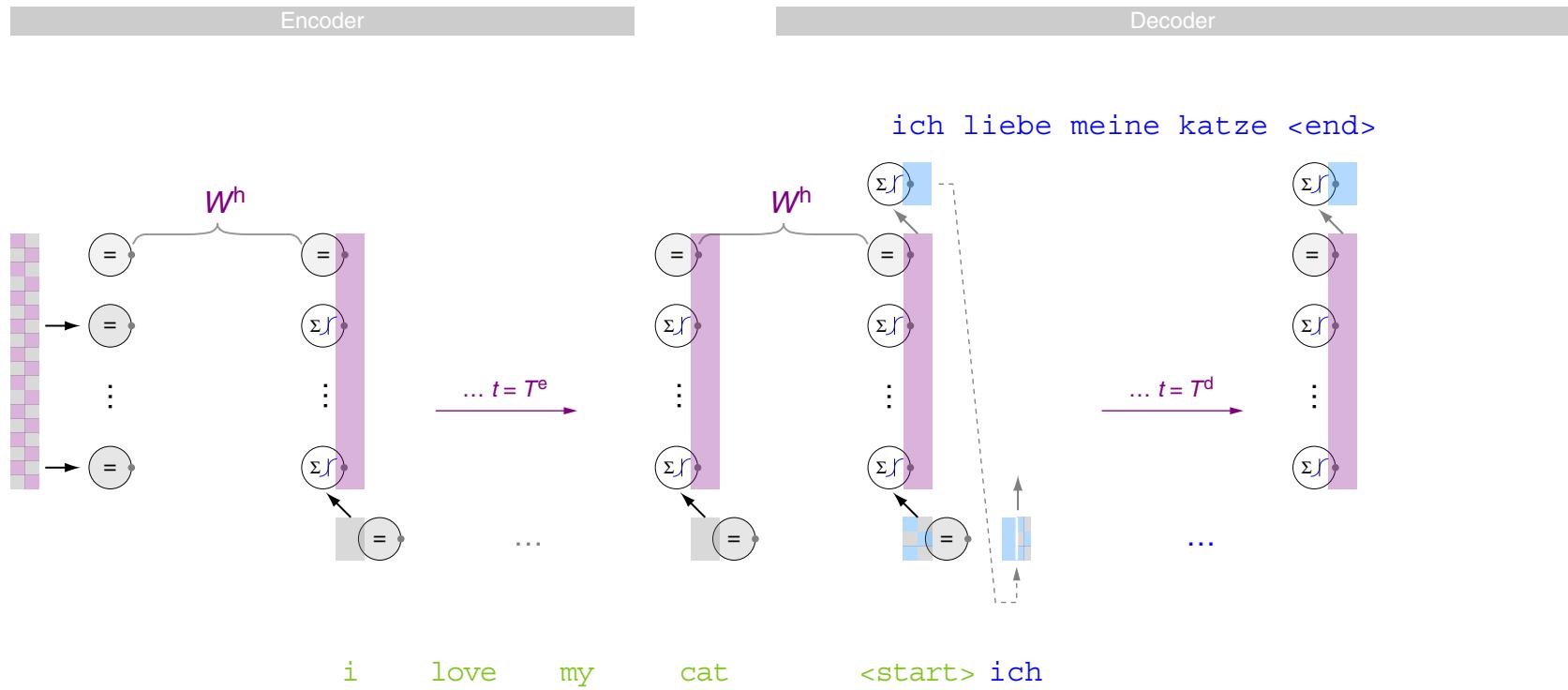
[notation: color, graph, language]



# Attention Mechanism

Notation III (language model view)

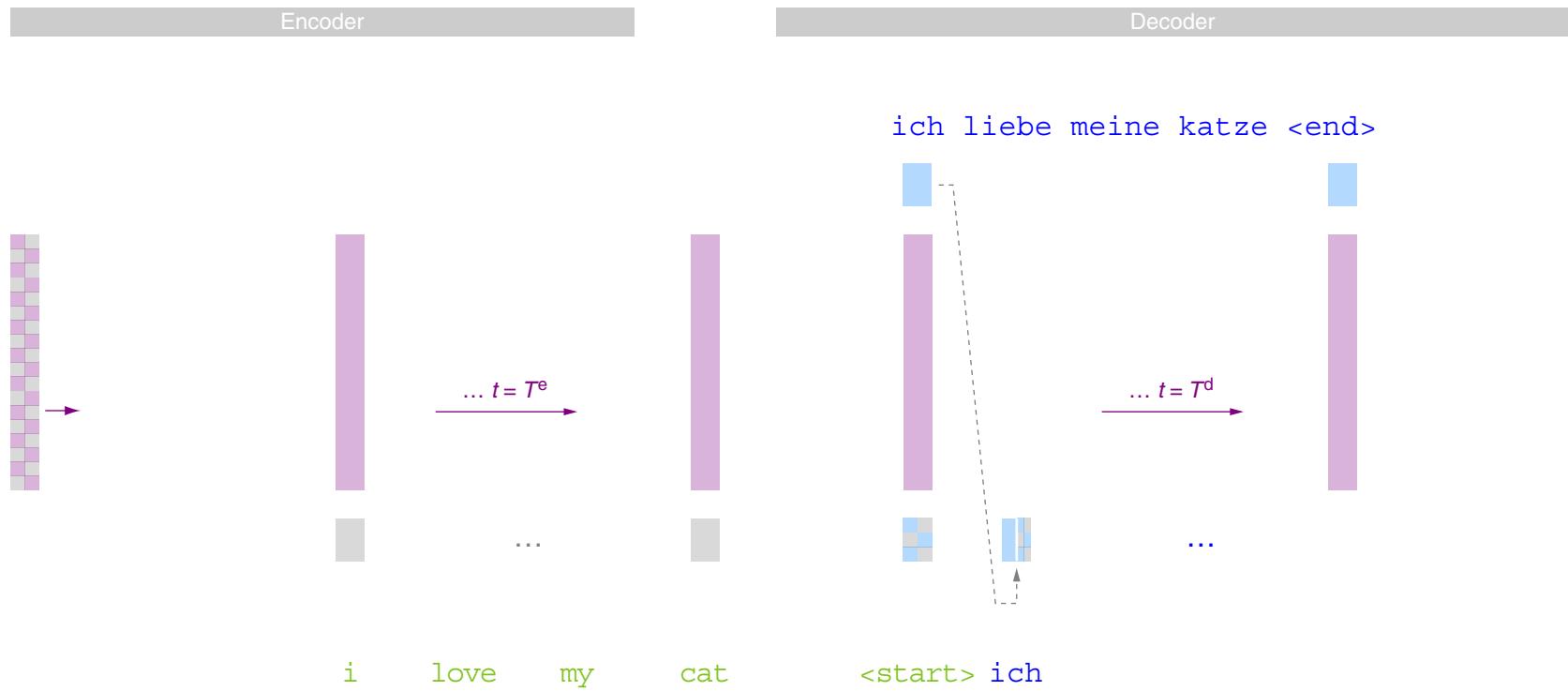
[notation: color, graph, language]



# Attention Mechanism

Notation III (language model view)

[notation: color, graph, language]



# Attention Mechanism

Notation III (language model view)

[notation: color, graph, language]



# Attention Mechanism

Notation III (language model view)

[notation: color, graph, language]



# Attention Mechanism

Notation III (language model view)

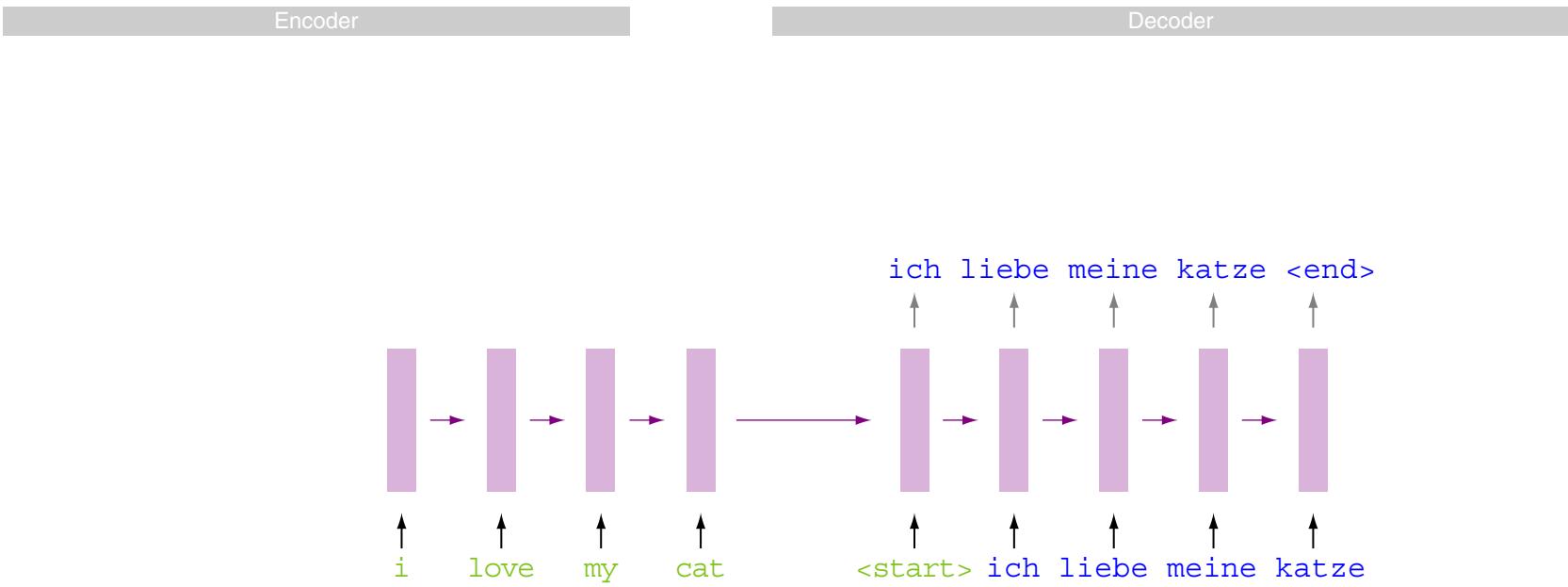
[notation: color, graph, language]



# Attention Mechanism

Notation III (language model view)

[notation: color, graph, language]

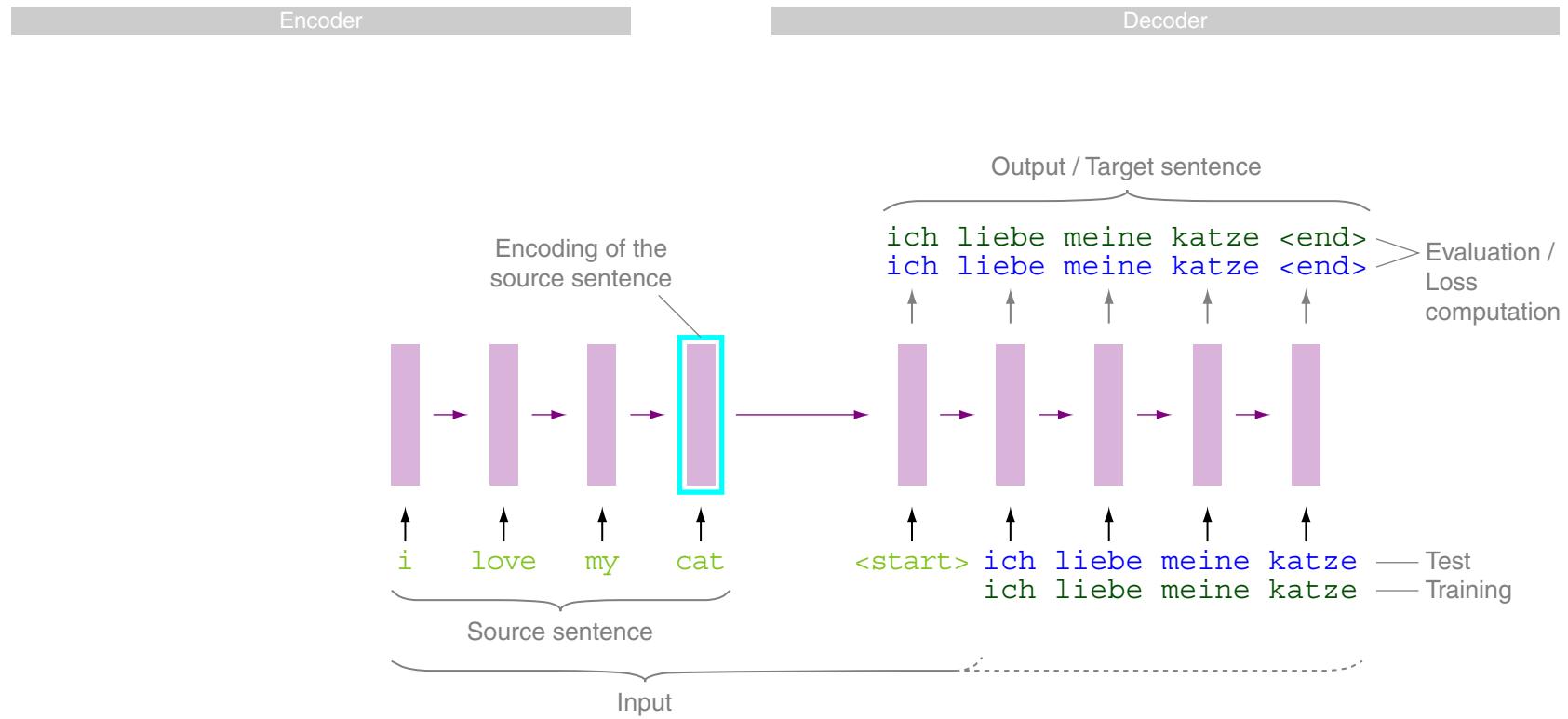


This level of abstraction is used among others by [Manning 2021, lecture CS224N].

# Attention Mechanism

## Notation III (language model view)

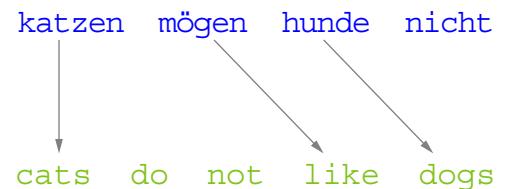
[notation: color, graph, language]



This level of abstraction is used among others by [Manning 2021, lecture CS224N].

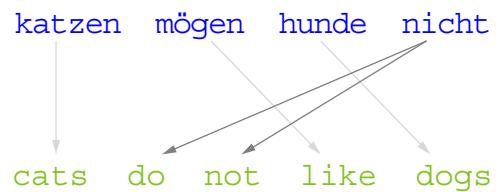
# Attention Mechanism

## Rationale of the Attention Mechanism



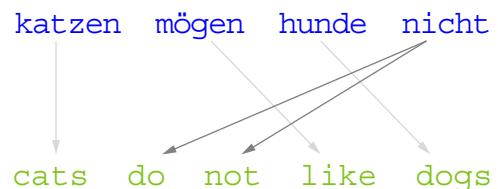
# Attention Mechanism

## Rationale of the Attention Mechanism (continued)



# Attention Mechanism

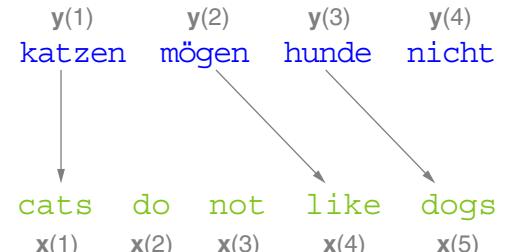
## Rationale of the Attention Mechanism (continued)



The trained RNN ( $\sim W^h, W^o$ ) has to generate in each time step the correct word.

# Attention Mechanism

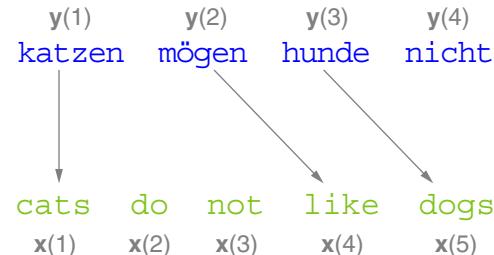
## Rationale of the Attention Mechanism (continued)



The trained RNN ( $\sim W^h, W^o$ ) has to generate in each time step the correct word.

# Attention Mechanism

## Rationale of the Attention Mechanism (continued)

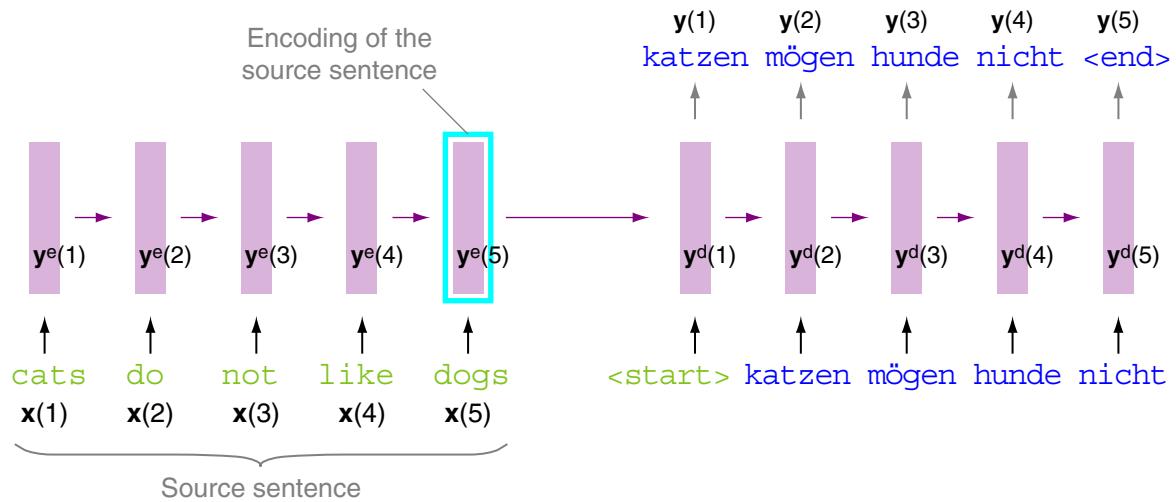


The trained RNN ( $\sim W^h, W^o$ ) has to generate in each time step the correct word.

- By short-circuiting the decoder output with the encoder, alleviate the problem of creating suitable hidden vectors  $y^d(t)$  from  $y^e(T^e)$ . This is called attention.
- Attention tackles (1) the bottleneck problem, (2) the vanishing gradient problem, and (3) the alignment problem of the translation task.

# Attention Mechanism

Query the Encoder to Tweak the Decoder Output



## Remarks (observations) :

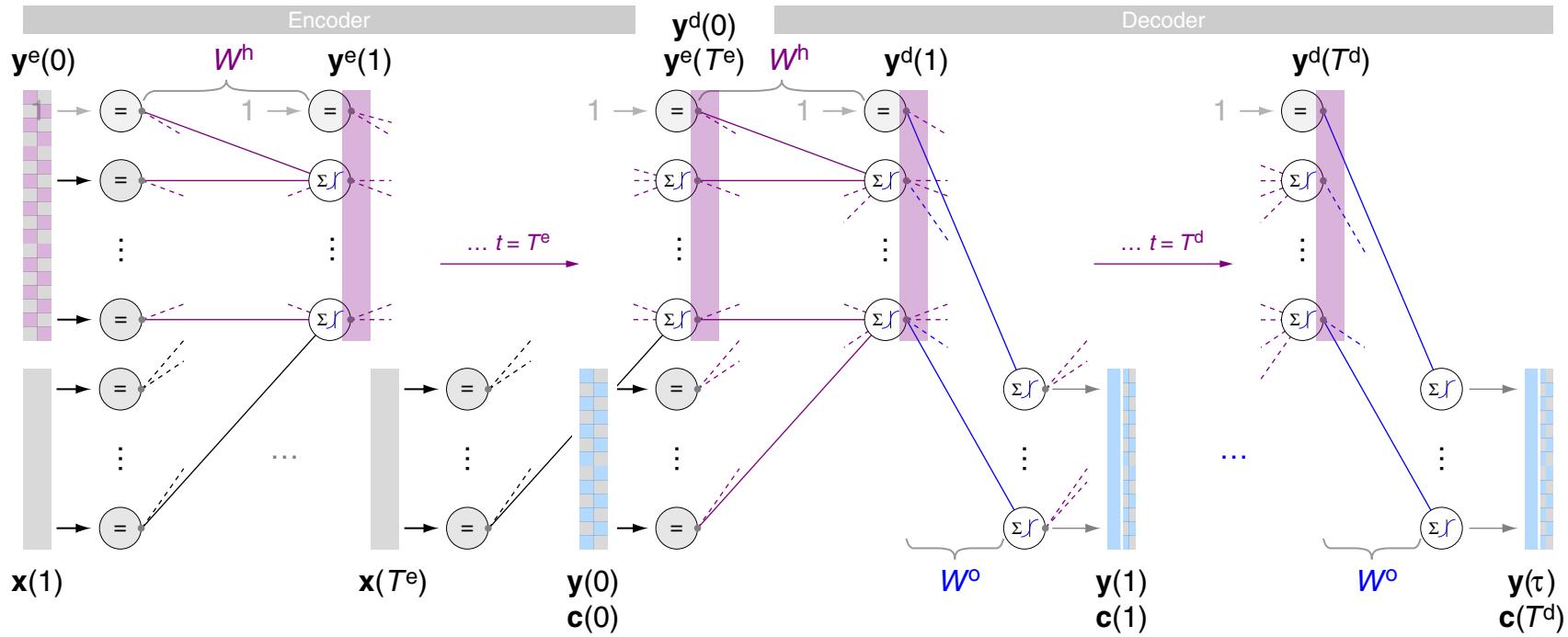
- Regarding the bottleneck problem. Attention allows the decoder to get additional information directly from the source sentence, this way bypassing the bottleneck  $y^e(T^e)$  (here:  $y^e(5)$ ).
- Regarding the vanishing gradient problem. Attention allows the decoder to look back into the past as far as it wants.
- Regarding the alignment problem. By inspecting the attention distribution  $a_1(t)$  for the output at time  $t$ , we can see what the decoder was focusing on and hence consider changes of the word order.

For example, without attention one has to manage to encode the sequence “cats do not like dogs <start> katzen” as a hidden vector  $y^d(2)$  from which  $y(2)$ , “mögen”, is decoded. I.e., in addition to looking back in time (keyword: vanishing gradient) the decoder, being at  $t=2$ , has to focus on the encoder at  $t=4$ , since the negation in the English sentence is before “like”, while the negation in the German sentence is after “mögen”.

With attention, the decoder can observe the importance of  $y^e(4)$ , “cats do not like”, for  $y^d(2)$ , and it will consider  $y^e(4)$  with a high weight to compute the output  $y(2)$ .

# Attention Mechanism

## RNN with Simple Attention



Input:

$$\mathbf{x}, \mathbf{y}^e(0), \mathbf{y}(t), \mathbf{y}(0) \hat{=} \langle \text{start} \rangle$$

Output:

$$\mathbf{y}(t) = \sigma_1(W^o \mathbf{y}^d(t)), t = 1, \dots, \tau$$

Hidden:

$$\mathbf{y}^e(t) = \sigma \left( W^h \begin{pmatrix} \mathbf{y}^e(t-1) \\ \mathbf{x}(t) \end{pmatrix} \right), t = 1, \dots, T^e$$

$$\mathbf{y}^d(t) = \sigma \left( W^h \begin{pmatrix} \mathbf{y}^d(t-1) \\ \mathbf{y}(t-1) \end{pmatrix} \right), t = 1, \dots, \tau$$

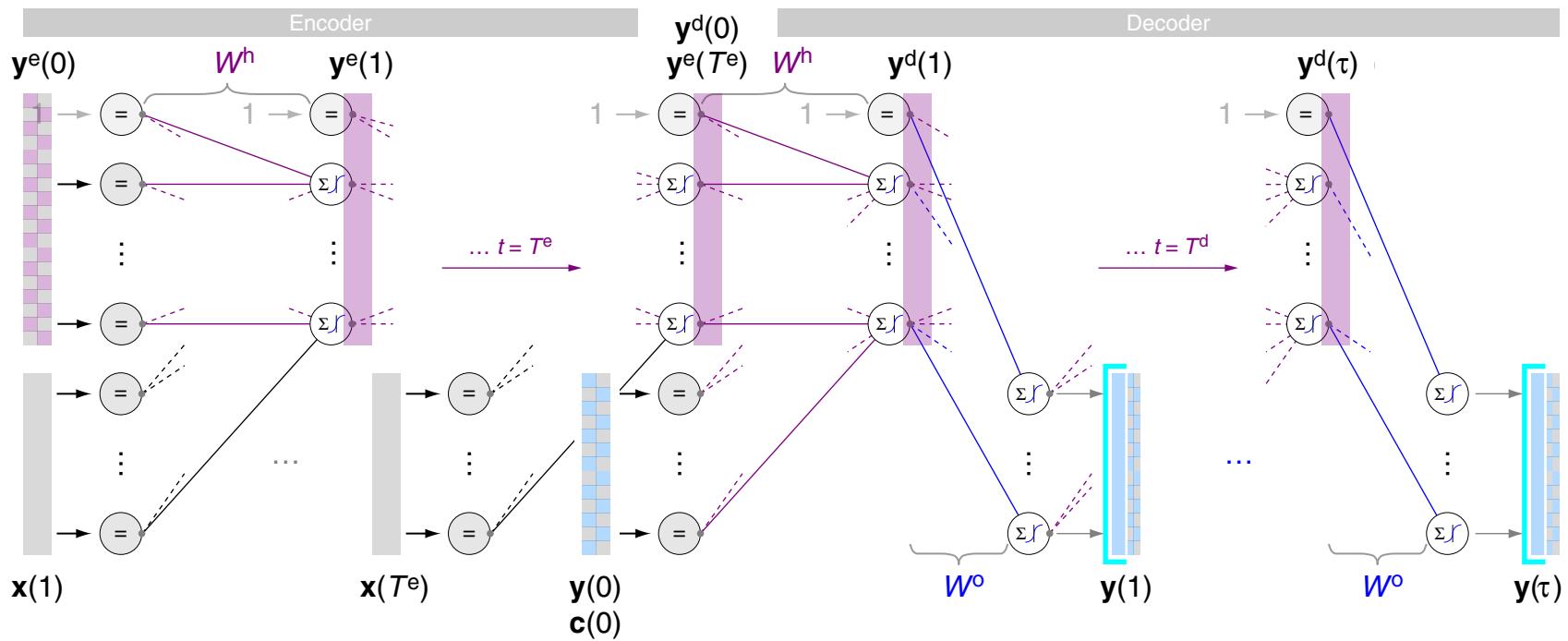
Target:

$$[\mathbf{c}(1), \dots, \mathbf{c}(T)]$$

$$\mathbf{c}(T) \hat{=} \langle \text{end} \rangle$$

# Attention Mechanism

## RNN with Simple Attention (continued)



Output:

$$\mathbf{y}(t) = \sigma_1 (W^o \mathbf{y}^d(t))$$

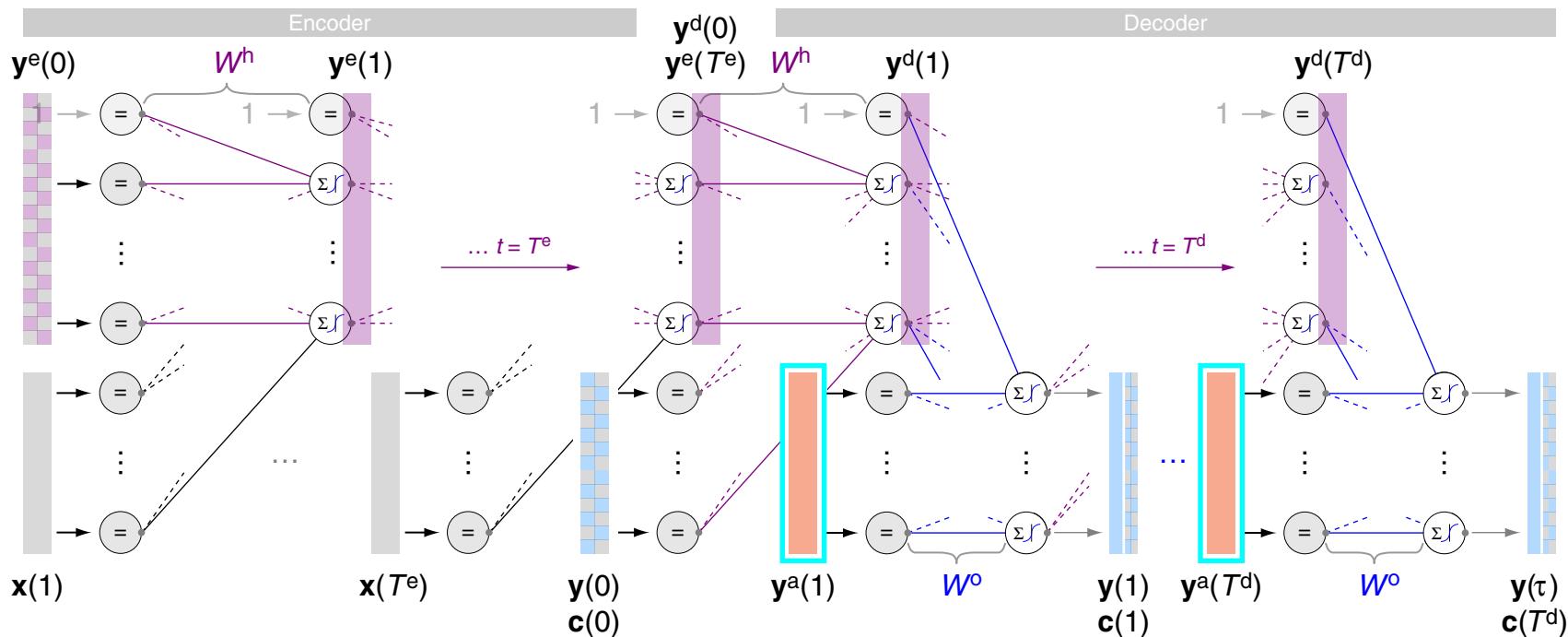
Attention:

$$\mathbf{y}^a(t) = [\mathbf{y}^e(1), \dots, \mathbf{y}^e(T^e)] \mathbf{a}(t)$$

$$\mathbf{a}(t) = [\mathbf{y}^e(1), \dots, \mathbf{y}^e(T^e)]^T \mathbf{y}^d(t), \quad t = 1, \dots, T^d$$

# Attention Mechanism

## RNN with Simple Attention (continued)



Output:

$$y(t) = \sigma_1 \left( W^o \begin{pmatrix} y^d(t) \\ y^a(t) \end{pmatrix} \right)$$

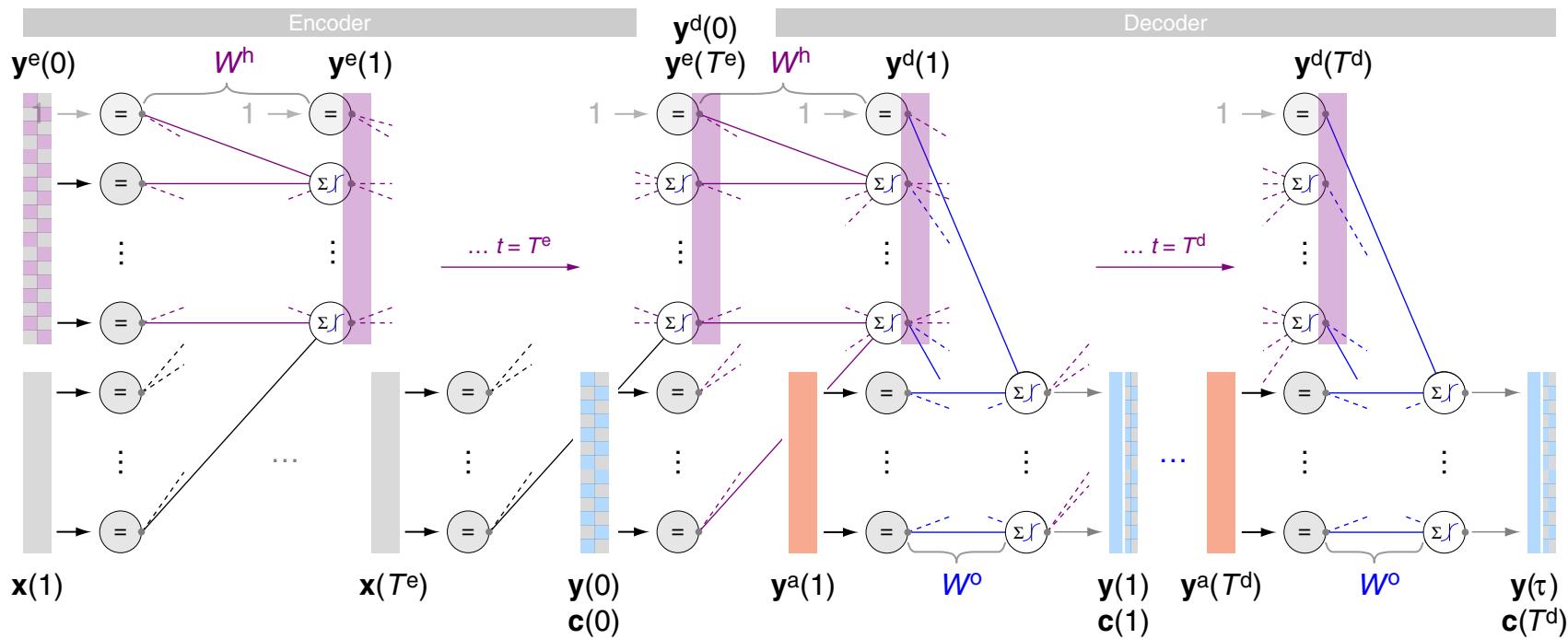
Attention:

$$y^a(t) = \left[ y^e(1), \dots, y^e(T^e) \right]^T \mathbf{a}(t)$$

$$\mathbf{a}(t) = \left[ y^e(1), \dots, y^e(T^e) \right]^T y^d(t), \quad t = 1, \dots, T^d$$

# Attention Mechanism

## RNN with Simple Attention (continued)



Output:

$$y(t) = \sigma_1 \left( W^o \begin{pmatrix} y^d(t) \\ y^a(t) \end{pmatrix} \right)$$

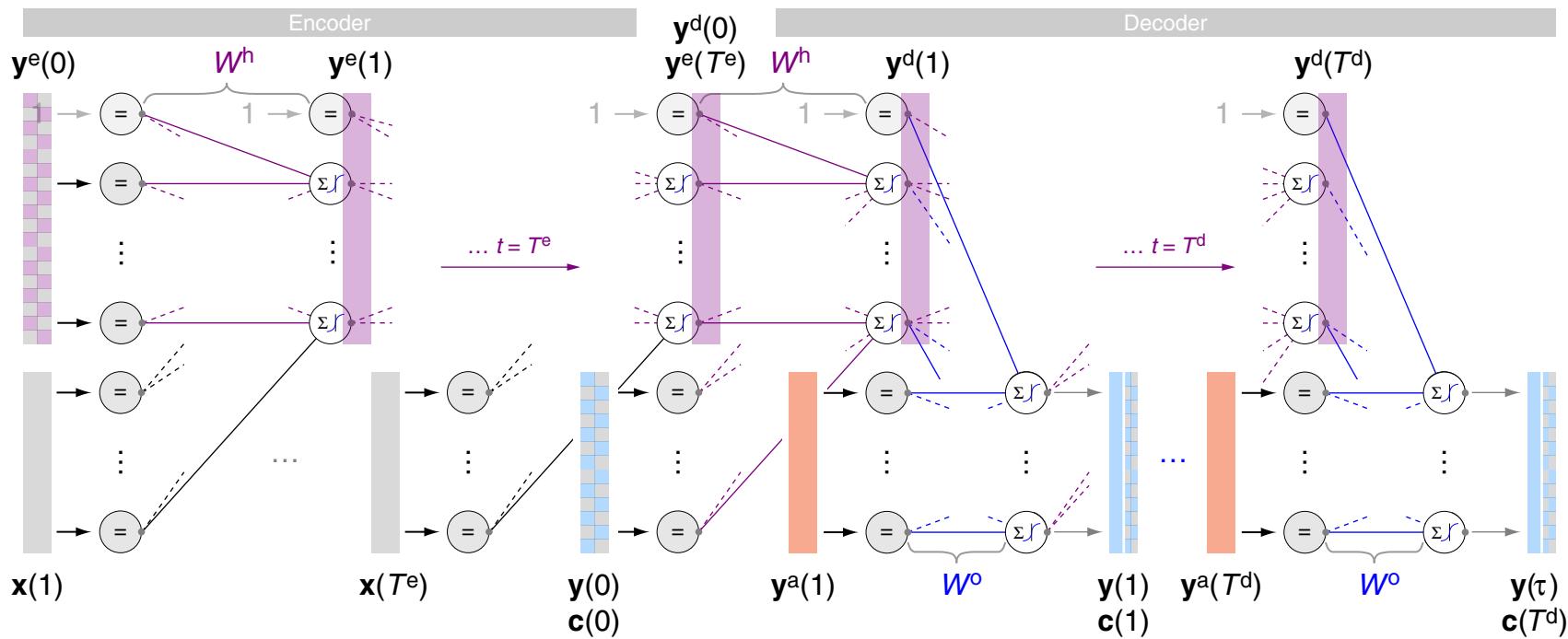
Attention:

$$y^a(t) = \left[ y^e(1), \dots, y^e(T^e) \right]^\top a(t)$$

$$a(t) = \left[ y^e(1), \dots, y^e(T^e) \right]^\top y^d(t), \quad t = 1, \dots, T^d$$

# Attention Mechanism

## RNN with Simple Attention (continued)



Output:

$$y(t) = \sigma_1 \left( W^o \left( y^d(t), y^a(t) \right) \right)$$

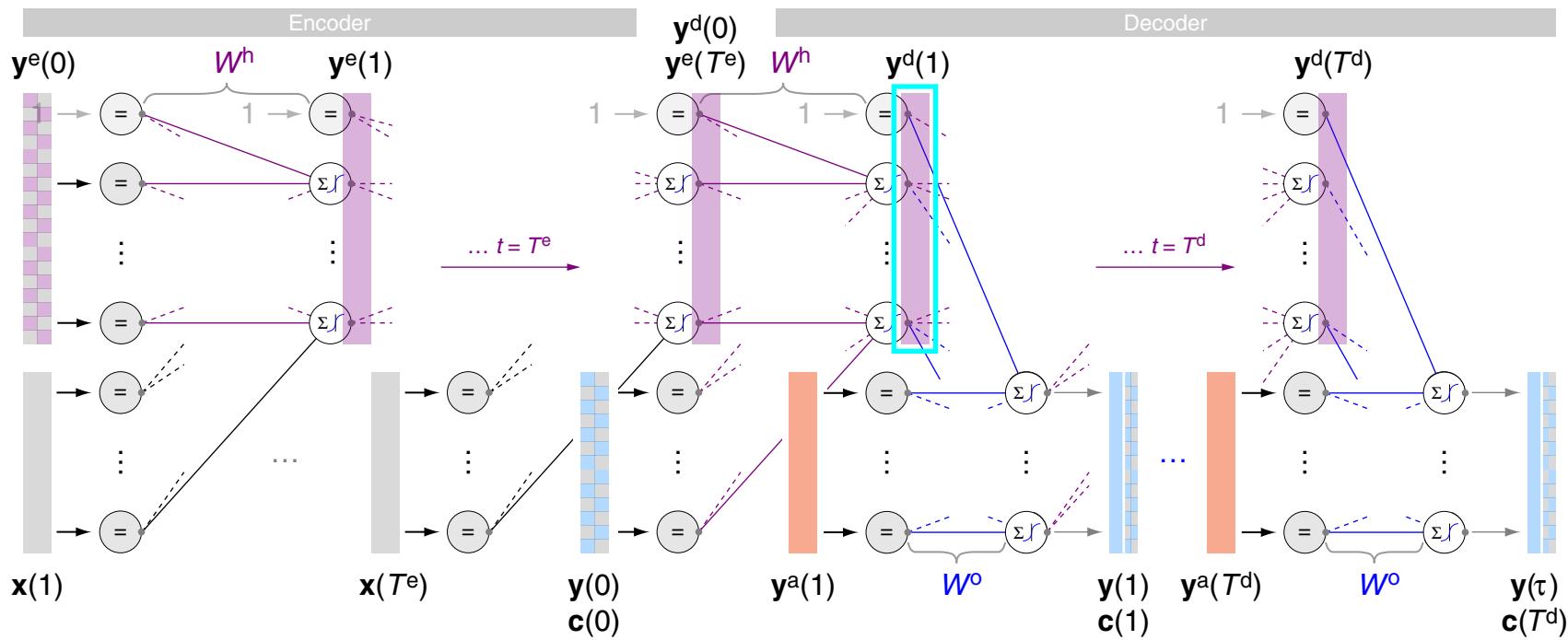
Attention:

$$y^a(t) = \left[ y^e(1), \dots, y^e(T^e) \right]^T a_1(t), \quad a_1(t) = \sigma_1(a(t))$$

$$a(t) = \left[ y^e(1), \dots, y^e(T^e) \right]^T y^d(t), \quad t = 1, \dots, T^d$$

# Attention Mechanism

## RNN with Simple Attention (continued)



Output:

$$\mathbf{y}(t) = \sigma_1 \left( W^o \begin{pmatrix} \mathbf{y}^d(t) \\ \mathbf{y}^a(t) \end{pmatrix} \right)$$

Attention:

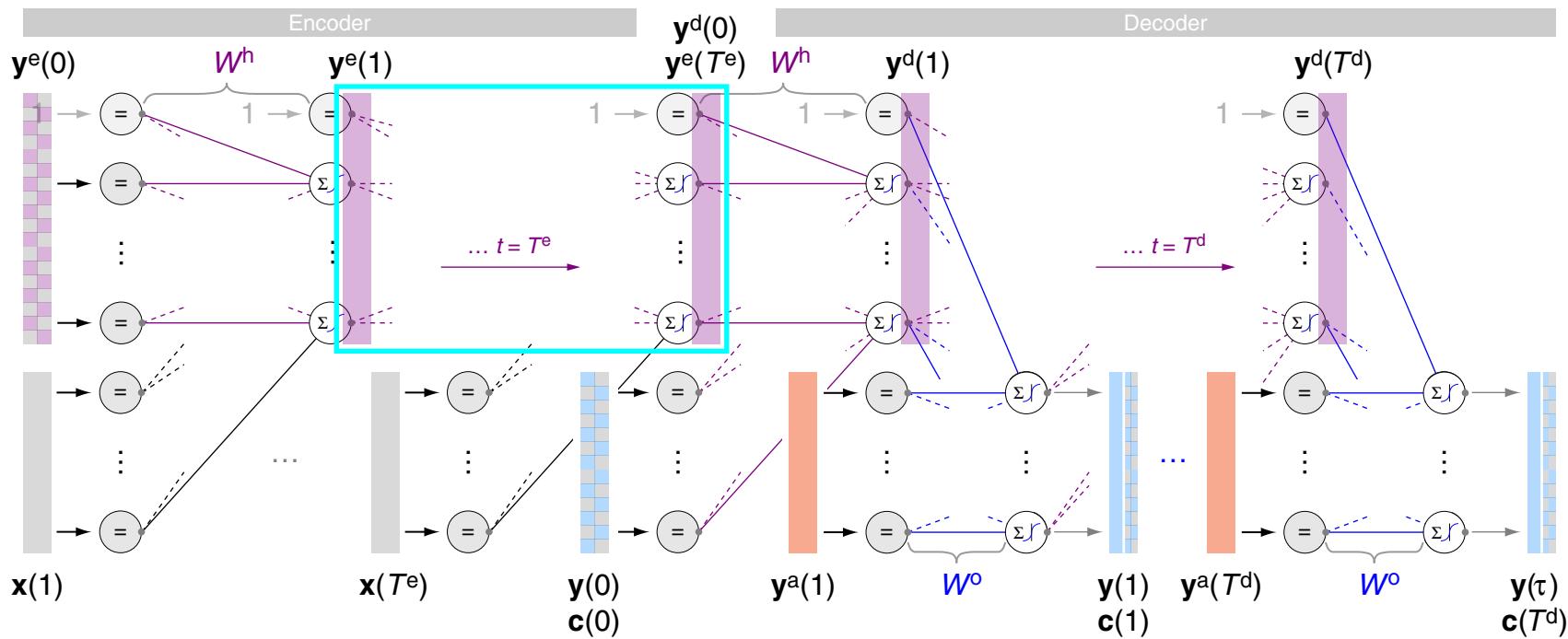
$$\mathbf{y}^a(t) = \left[ \mathbf{y}^e(1), \dots, \mathbf{y}^e(T^e) \right] \mathbf{a}_1(t), \quad \mathbf{a}_1(t) = \sigma_1(\mathbf{a}(t))$$

$$\mathbf{a}(t) = \left[ \mathbf{y}^e(1), \dots, \mathbf{y}^e(T^e) \right]^T \boxed{\mathbf{y}^d(t)}, \quad t = 1, \dots, T^d$$

“Queries”

# Attention Mechanism

## RNN with Simple Attention (continued)



Output:

$$y(t) = \sigma_1 \left( W^o \begin{pmatrix} y^d(t) \\ y^a(t) \end{pmatrix} \right)$$

Attention:

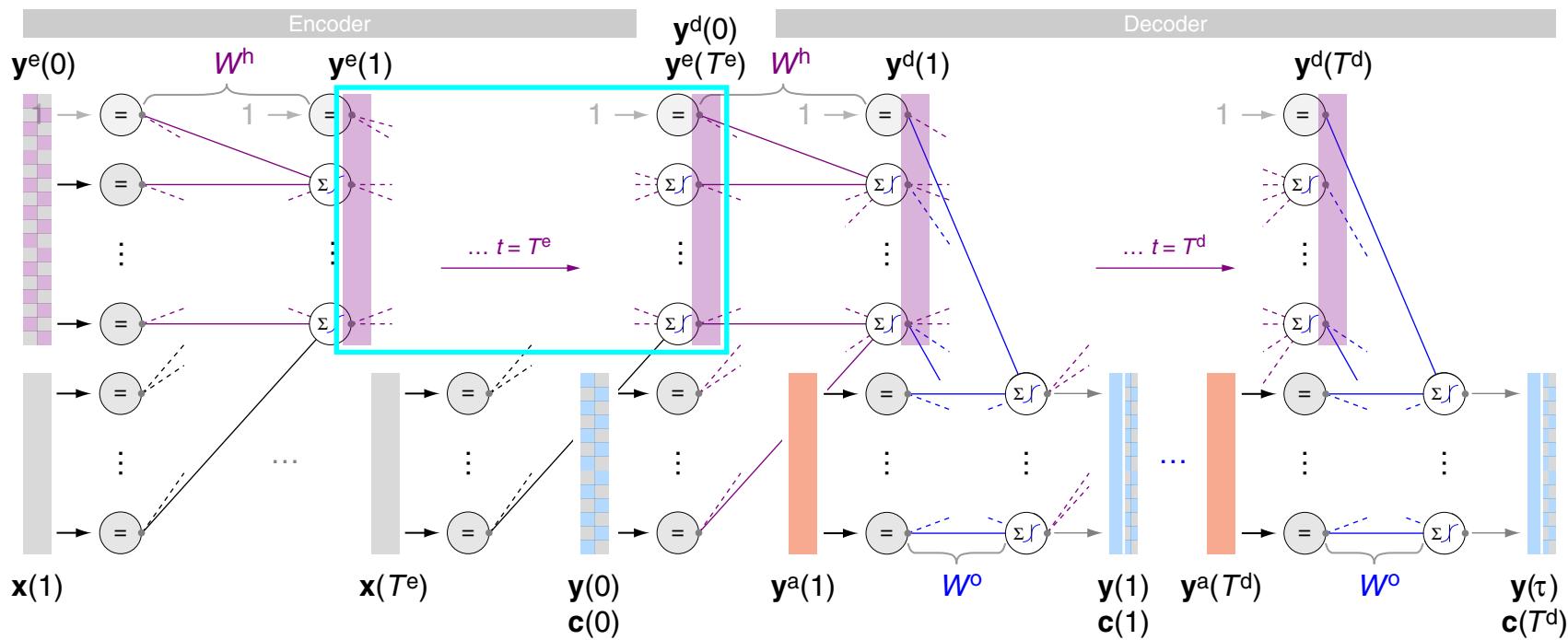
$$y^a(t) = [y^e(1), \dots, y^e(T^e)] \quad a_1(t), \quad a_1(t) = \sigma_1(a(t))$$

$$a(t) = \boxed{[y^e(1), \dots, y^e(T^e)]^T} y^d(t), \quad t = 1, \dots, T^d$$

“Keys”

# Attention Mechanism

## RNN with Simple Attention (continued)



Output:

$$y(t) = \sigma_1 \left( W^o \begin{pmatrix} y^d(t) \\ y^a(t) \end{pmatrix} \right)$$

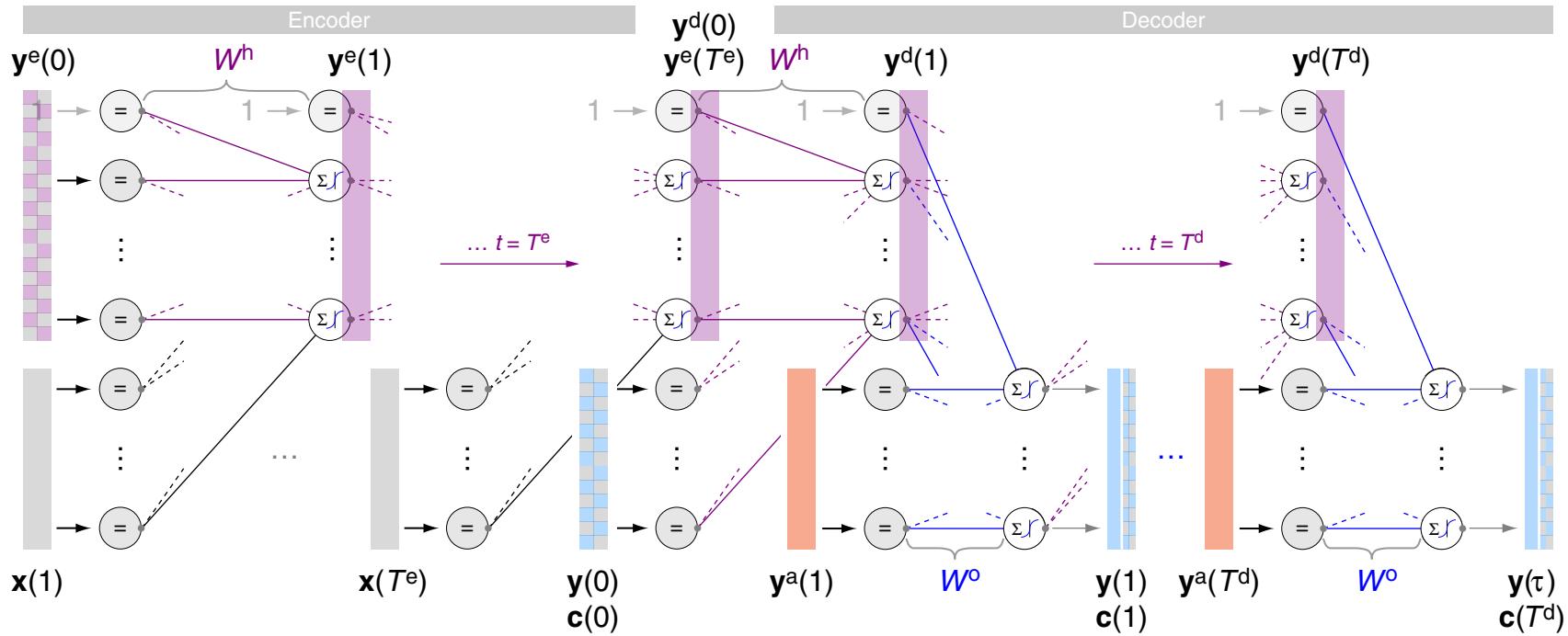
Attention:

$$y^a(t) = \boxed{\begin{bmatrix} y^e(1), \dots, y^e(T^e) \end{bmatrix}} \quad \mathbf{a}_1(t), \quad a_1(t) = \sigma_1(a(t)) \quad \text{"Values"}$$

$$\mathbf{a}(t) = \begin{bmatrix} y^e(1), \dots, y^e(T^e) \end{bmatrix}^T y^d(t), \quad t = 1, \dots, T^d$$

# Attention Mechanism

## RNN with Parameterized Attention



Output:

$$y(t) = \sigma_1 \left( W^o \begin{pmatrix} y^d(t) \\ y^a(t) \end{pmatrix} \right)$$

Attention:

$$y^a(t) = \left( W^v \begin{bmatrix} y^e(1), \dots, y^e(T^e) \end{bmatrix} \right) a_1(t), \quad a_1(t) = \sigma_1 \left( a(t) \right)$$

$$a(t) = \left( W^K \begin{bmatrix} y^e(1), \dots, y^e(T^e) \end{bmatrix} \right)^T \left( W^Q y^d(t) \right)$$

## Remarks (attention calculus) :

- As before,  $\sigma_1()$  denotes the softmax function.
- $\mathbf{a}(t)$  is called vector of *attention scores*,  $\mathbf{a}_1(t) := \sigma_1(\mathbf{a}(t))$  is the *attention distribution*.  
[Manning 2021]
- The  $i$ th component  $a_i(t)$  of the attention score vector  $\mathbf{a}(t)$ ,  $i = 1, \dots, T^e$ , models the importance of the  $i$ th *encoder* hidden state  $\mathbf{y}^e(i)$  for the *decoder* hidden state  $\mathbf{y}^d(t)$ :  $a_i(t)$  is the scalar product of  $\mathbf{y}^e(i)$  and  $\mathbf{y}^d(t)$  (do not overlook the matrix transpose operation).
- $\mathbf{y}^a(t)$  is the result of weighting the encoder hidden state sequence  $[\mathbf{y}^e(1), \dots, \mathbf{y}^e(T^e)]$  with the attention distribution  $\mathbf{a}_1(t)$ . I.e., each vector  $\mathbf{y}^e(i)$  is considered as a “value” that is weighted with the relative importance stored in the respective dimension (= time step) of  $\mathbf{a}_1(t)$ .  
 $\mathbf{y}^a(t)$  is called *attention [vector]* for output vector  $\mathbf{y}(t)$  since it helps to pay attention to the most influential input states for  $\mathbf{y}(t)$ .
- Consider  $\mathbf{a}(t)$ . The scalar product of  $\mathbf{y}^e(i)$  and  $\mathbf{y}^d(t)$  becomes maximum if  $\mathbf{y}^e(i)$  and  $\mathbf{y}^d(t)$  are identical. The distribution  $\mathbf{a}_1(t)$  of the  $T^e$  weights of  $\mathbf{a}(t)$  reflects the distribution of absolute values among the  $\mathbf{y}^e$ .

Consider  $\mathbf{a}_1(t)$ . If some  $\mathbf{y}^e(i)$  has a high absolute value (compared to the other  $\mathbf{y}^e$ ) and if it is similar to  $\mathbf{y}^d(t)$ , it will push the weight of the  $i$ th dimension of  $\mathbf{a}_1(t)$  towards 1 (and the others towards zero).

In the extreme case,  $\mathbf{y}^d(t)$  is equal to  $\mathbf{y}^e(i)$ , and  $\mathbf{y}^a(t)$  is equal to  $\mathbf{y}^e(i)$ . I.e., as input for  $W^o$  only the  $i$ th *encoder* state,  $\mathbf{y}^e(i)$ , is used, say,  $W^o(\mathbf{y}^e(i), \mathbf{y}^e(i))$  is passed directly to position  $t$  of the output sequence.

## Remarks (parameterized attention) :

- $y^d(t)$  is also denoted as “query,” while the sequence of  $y^e(i)$  are denoted as “keys” in this setting. Since  $a(t)$  is normalized with a softmax operation it represents the importance of the  $T^e$  time steps as probabilities.
- *Parameterized* attention introduces three weight matrices,  $W^Q$ ,  $W^K$ , and  $W^V$ , in order to learn a more sophisticated version of the simple attention vector  $y^a(t)$ . In this regard the matrices are called “query projection”, “key projection”, and “value projection” [matrix] respectively.
- Inspired by nature, the structure of the model function  $y()$  has been developed in the form of a network that connects the matrices  $W^h$ ,  $W^o$ ,  $W^Q$ ,  $W^K$ , and  $W^V$  in a particular manner. Note that the shown model function is specified completely by the set of parameters  $w$ , which are organized in the aforementioned matrices.