Understanding Attention Training via Output Relevance

Charlie Snell, Ruiqi Zhong, Jacob Steinhardt, Dan Klein

Berkeley

N L P

Outline

- Prior work on text classification:
  - Standardly trained attention ≈ Explanations .
  - Can train models to attend to irrelevant words, same accuracy.
- Our work studies how the attention evolves at training time, for text classification and machine translation.

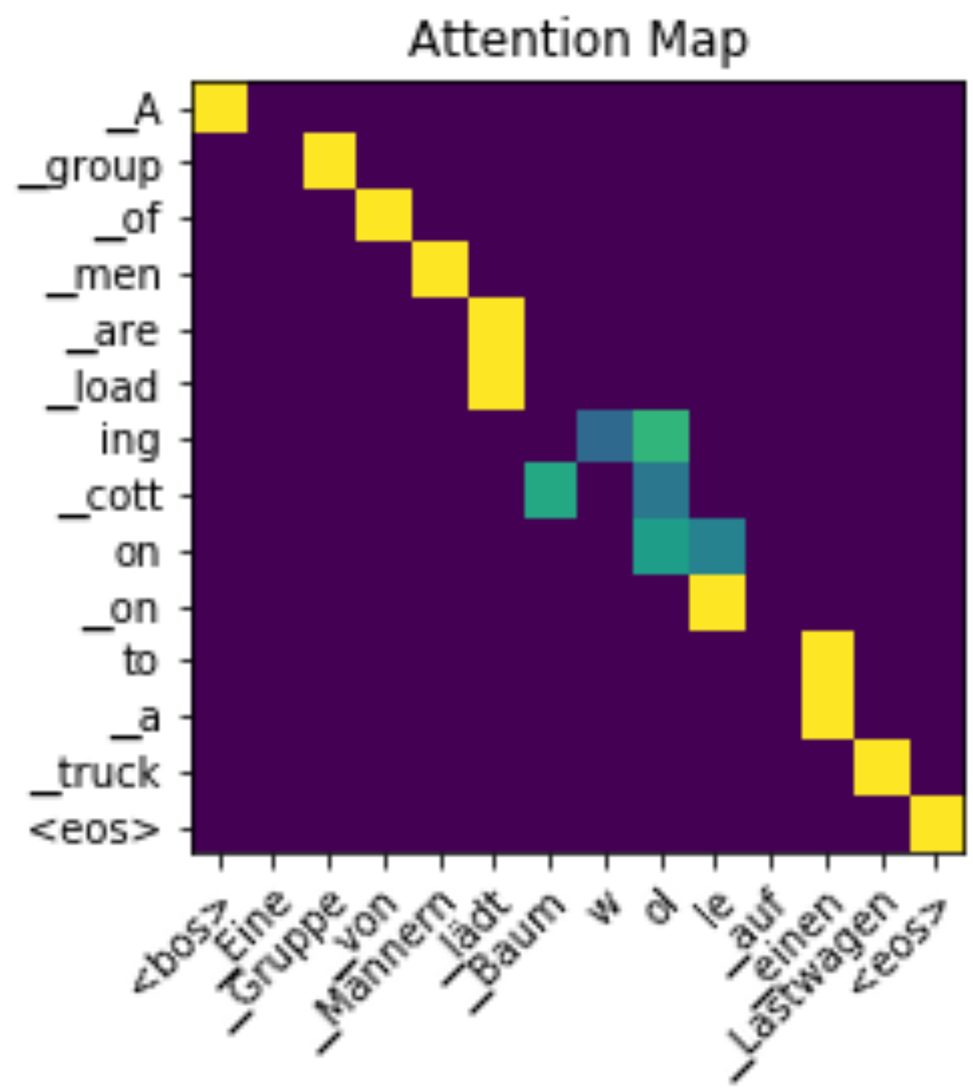
Attention Distribution over Inputs

Sentiment Classification

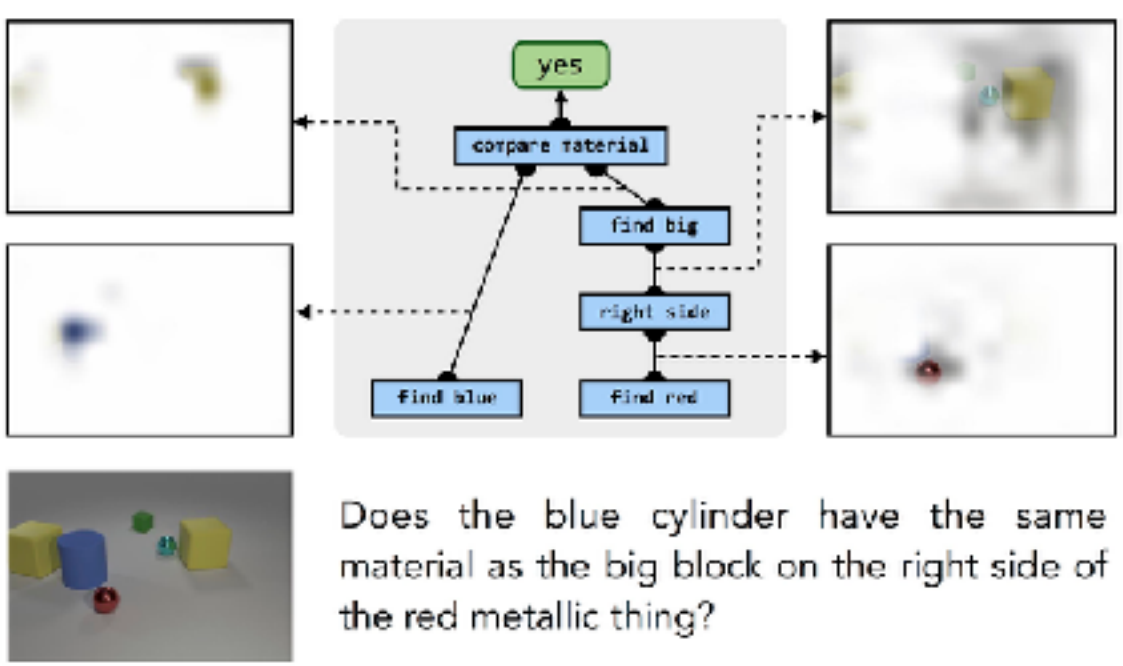
Negative

after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

Seems to provide explanations !!



German -> English Translation



Visual Question Answering

Attention ≈ Explanation?

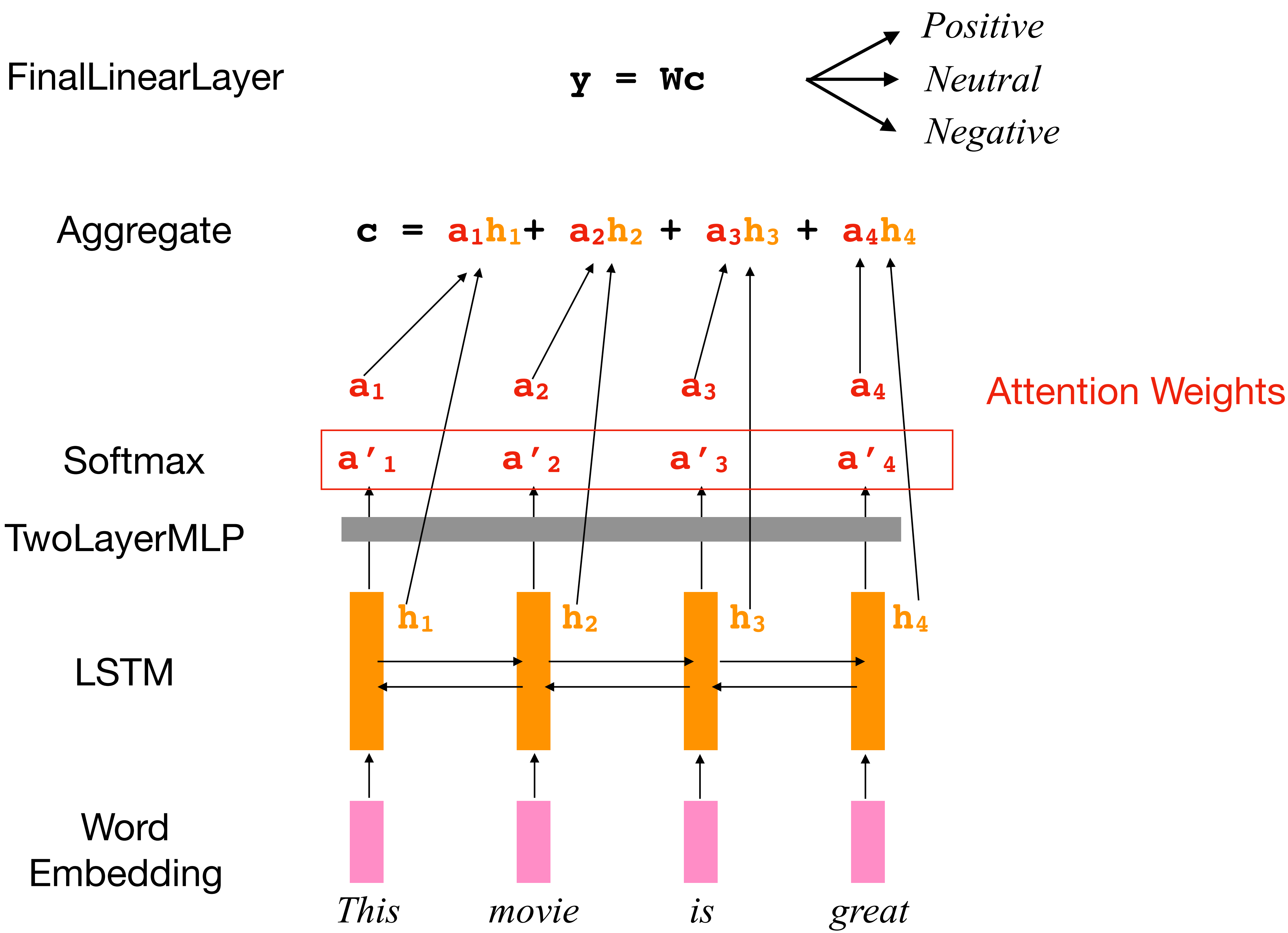
- Attention is not Explanation (Jain & Wallace 2019)
- Is Attention Interpretable? (Serrano & Smith 2019)
- Attention is not not Explanation (Wiegreffe & Pinter 2019)
- Attention Interpretability Across Tasks (Shikhar et al. 2019)
- Understanding Attention Training via Output Relevance (Snell et al. 2020)

A series of work trying to understand attention mechanism.

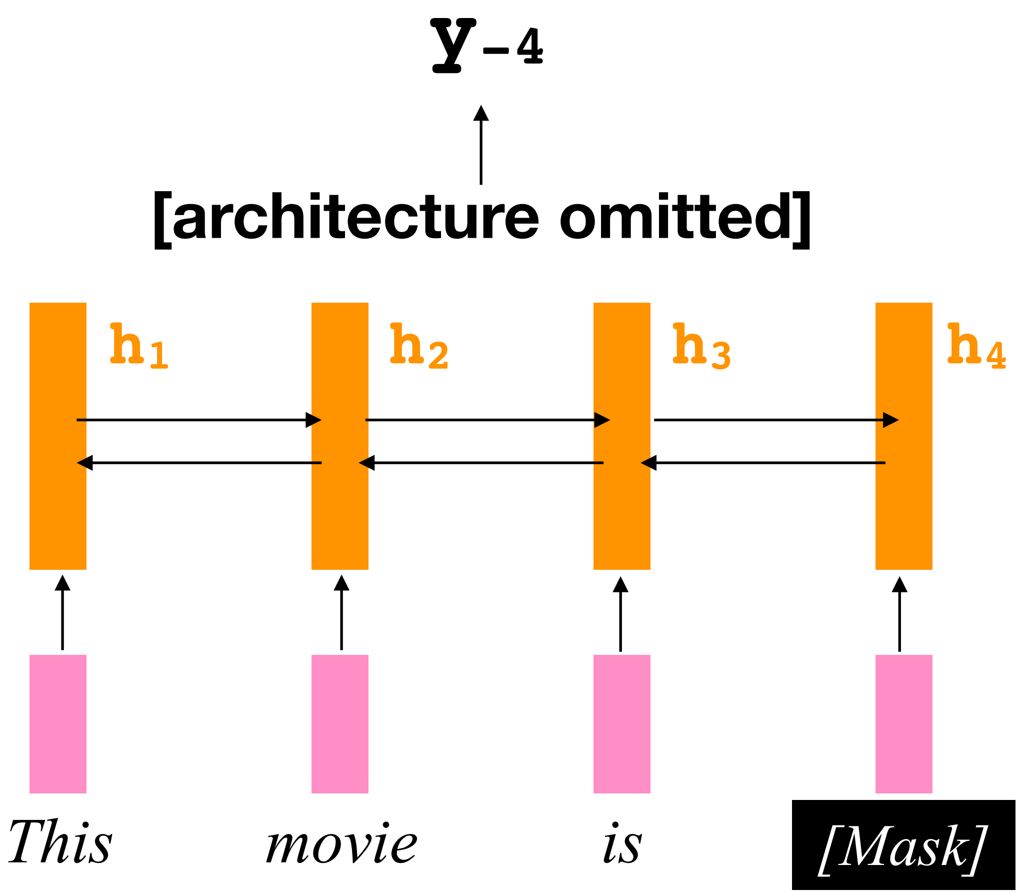
Prior Works on Classification

- If the model is trained in a standard way, attention weights correlate strongly with individual token influence.
- Uniform attention gives the same accuracy (attention does not matter) .
- Models can be trained to attend to irrelevant words, *without harming accuracy*.

Text Classification



Attention ≈ Explanations



Leave-one-out influence:
$\Delta_4 := y - y_{-4}$

	This	*movie*	*is*	*great*
Attention	a_1	a_2	a_3	$\uparrow a_4$
Influence	Δ_1	Δ_2	Δ_3	$\uparrow \Delta_4$

Strongly Correlates

Train “Deceptive” Attention

Call a standard model P.
Now train the deceptive model Q.

$$\mathcal{L}(\mathcal{M}_P, \mathcal{M}_Q) = TVD(\hat{y_P}, \hat{y_Q}) - \lambda KL(a_P || a_Q)$$

Q makes similar predictions as P.

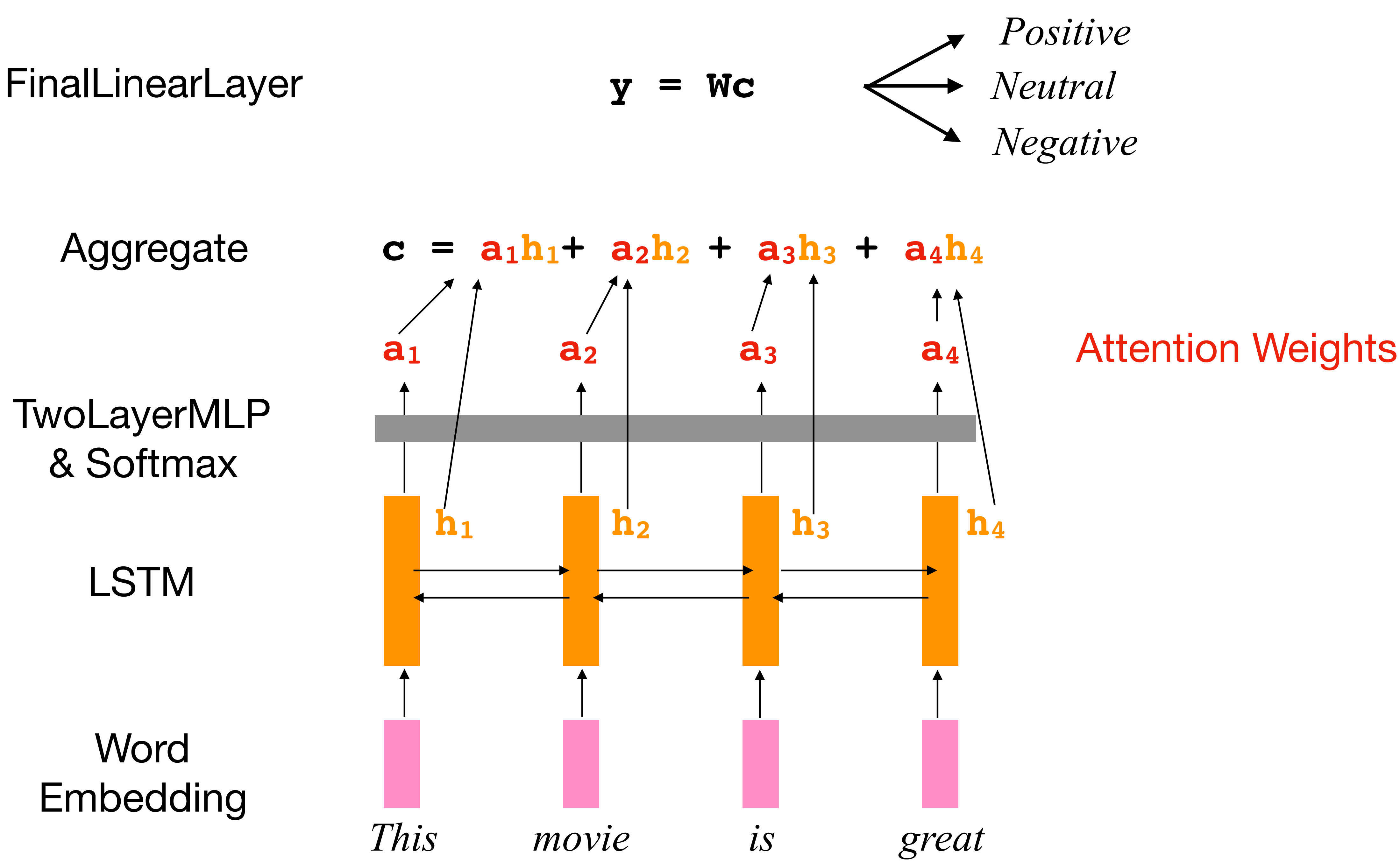
Q attends differently from P.

Attention becomes anti-correlated with explanations !

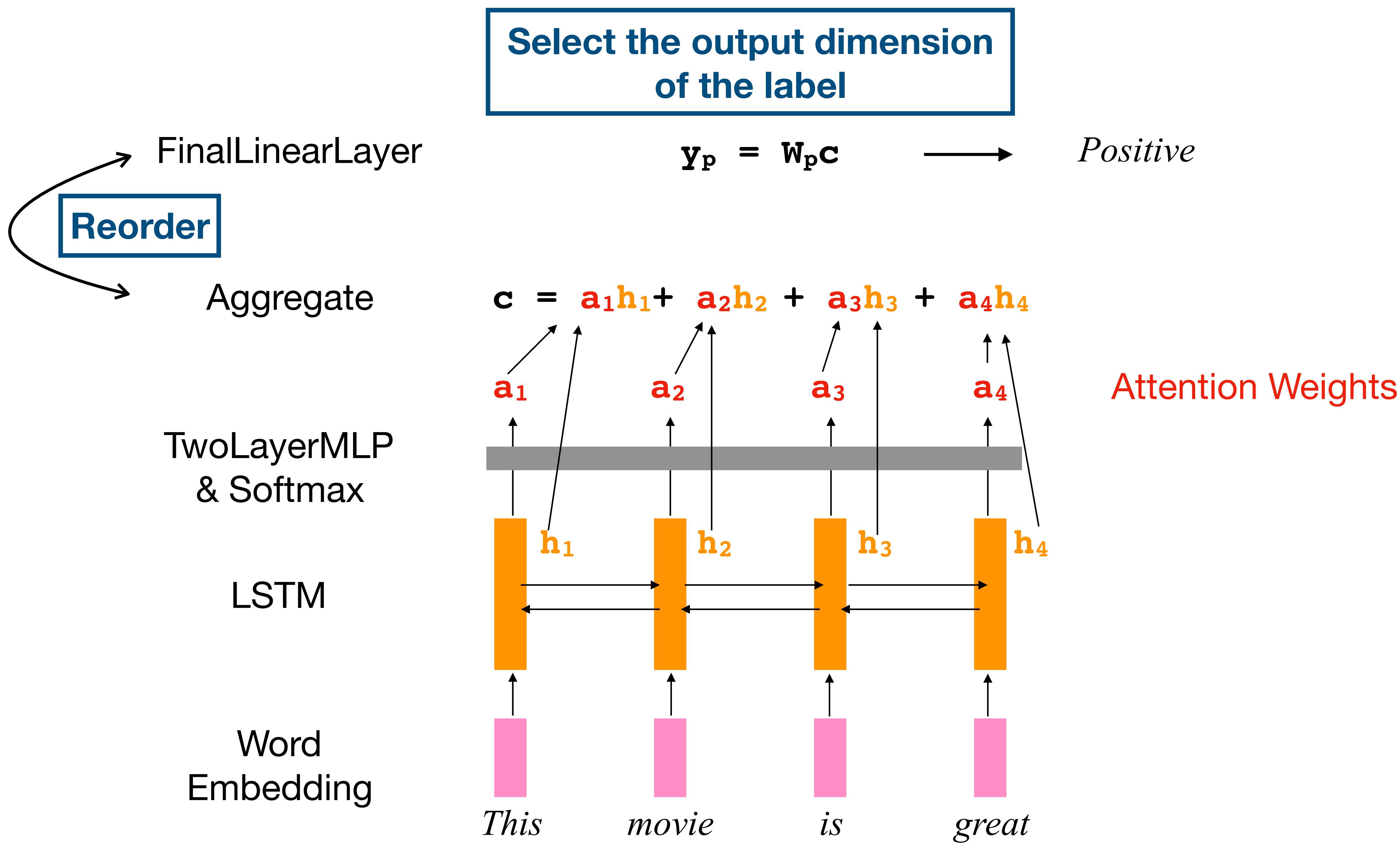
Our Work: Understand Attention Training

- We can construct two models with the same loss but different attention weights.
- Need to open the black-box of standard training.

Defining Output Relevance



Defining Output Relevance

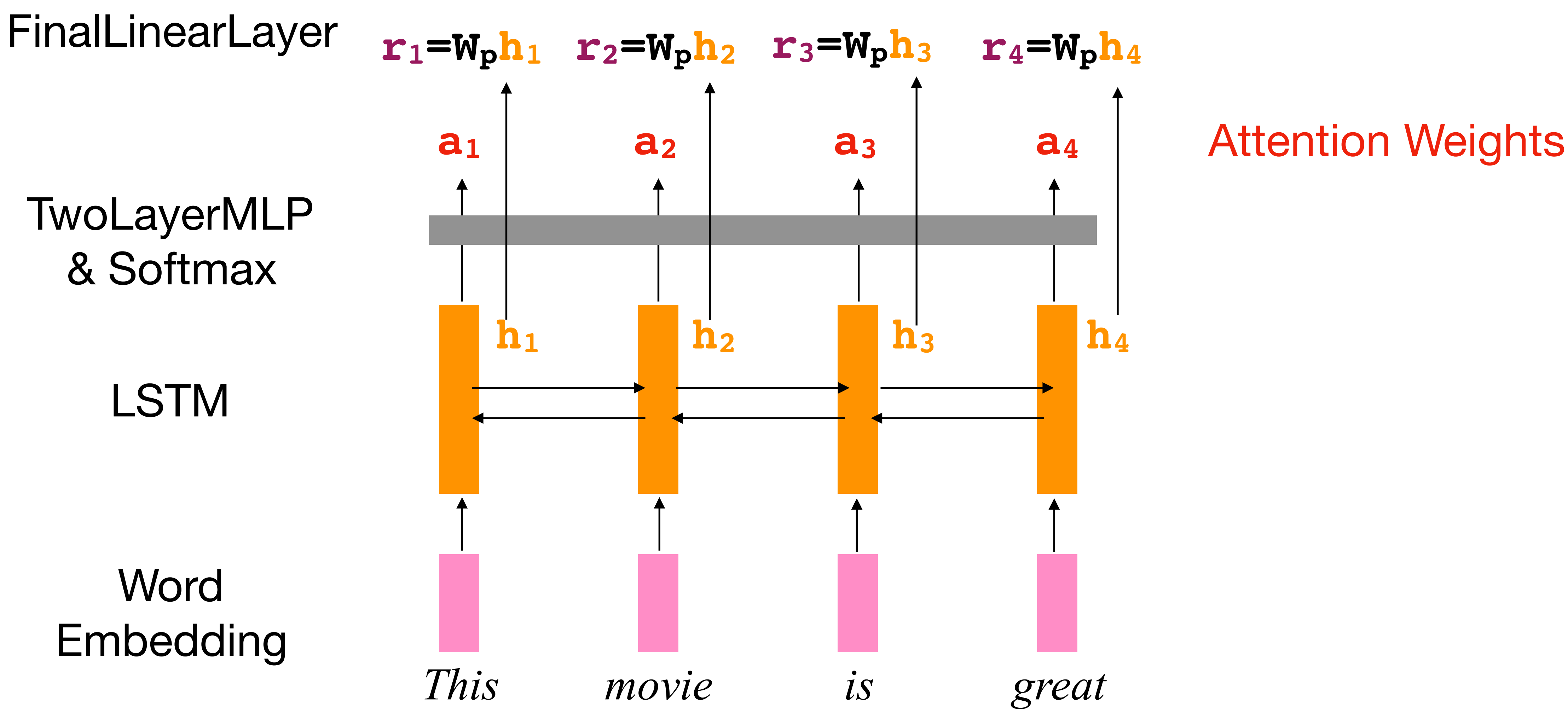


Defining Output Relevance

Training objective optimizes $\mathbf{y_p}$

Output Relevance r_4: how much does the model associate h_4 with the positive label

Aggregate $\quad \mathbf{y_p} = a_1 r_1 + a_2 r_2 + a_3 r_3 + a_4 r_4 \longrightarrow$ *Positive* Logit



Attention a and Output Relevance r

Training objective maximizes y_p

Output Relevance r_4: how much does the model associate h_4 with the positive label

$$y_p = a_1 r_1 + a_2 r_2 + a_3 r_3 + a_4 r_4 \longrightarrow \textit{Positive}\ \text{Logit}$$

- a attracted to larger r
- What is a and r when initialized? => both uniform
- How will a and r interact? Near the beginning, a remains uniform. Under uniform attention, r increases faster at “keyword" positions, then attracts a

Attention a and Output Relevance r

Training objective maximizes y_p

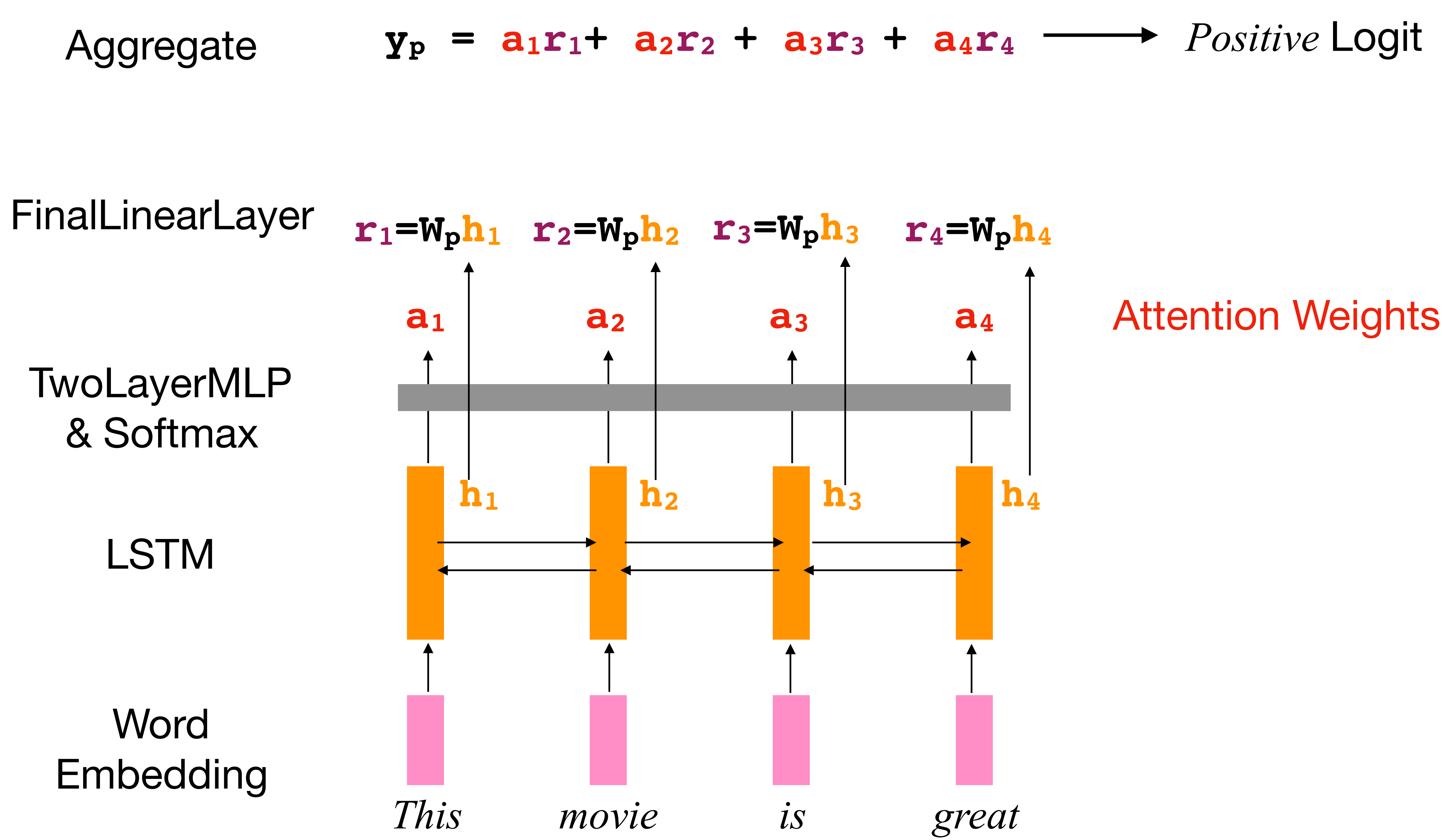
Output Relevance r_4: how much does the model associate h_4 with the positive label

$y_p = a_1r_1 + a_2r_2 + a_3r_3 + a_4r_4 \longrightarrow$ *Positive* Logit

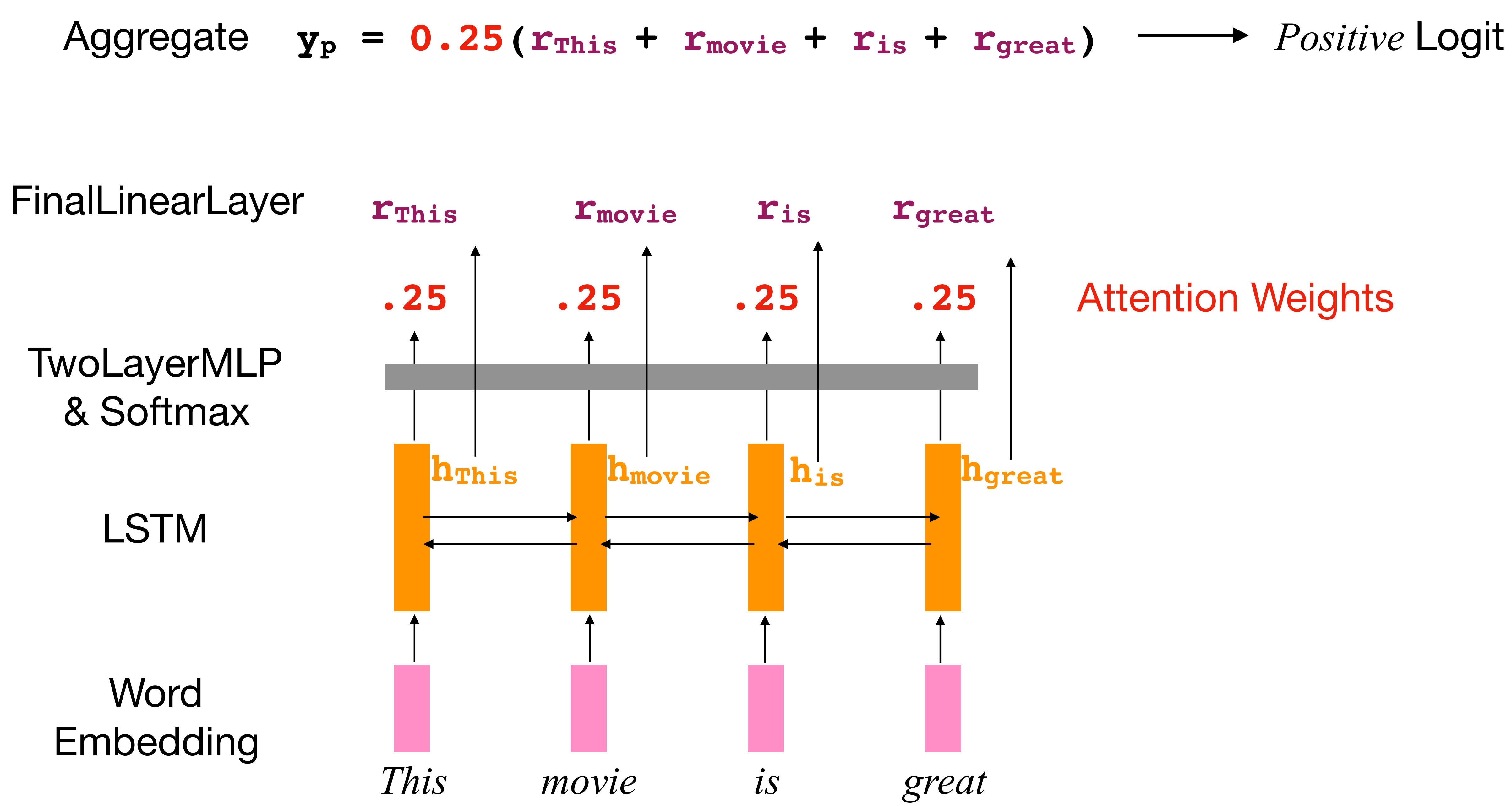
- a attracted to larger r
- What is a and r when initialized? => both uniform
- How will a and r interact? Near the beginning, a remains uniform. Under uniform attention, <u>r increases faster at "keyword" positions</u>, then attracts a

?

Intuition of Increasing r



Intuition of Increasing r



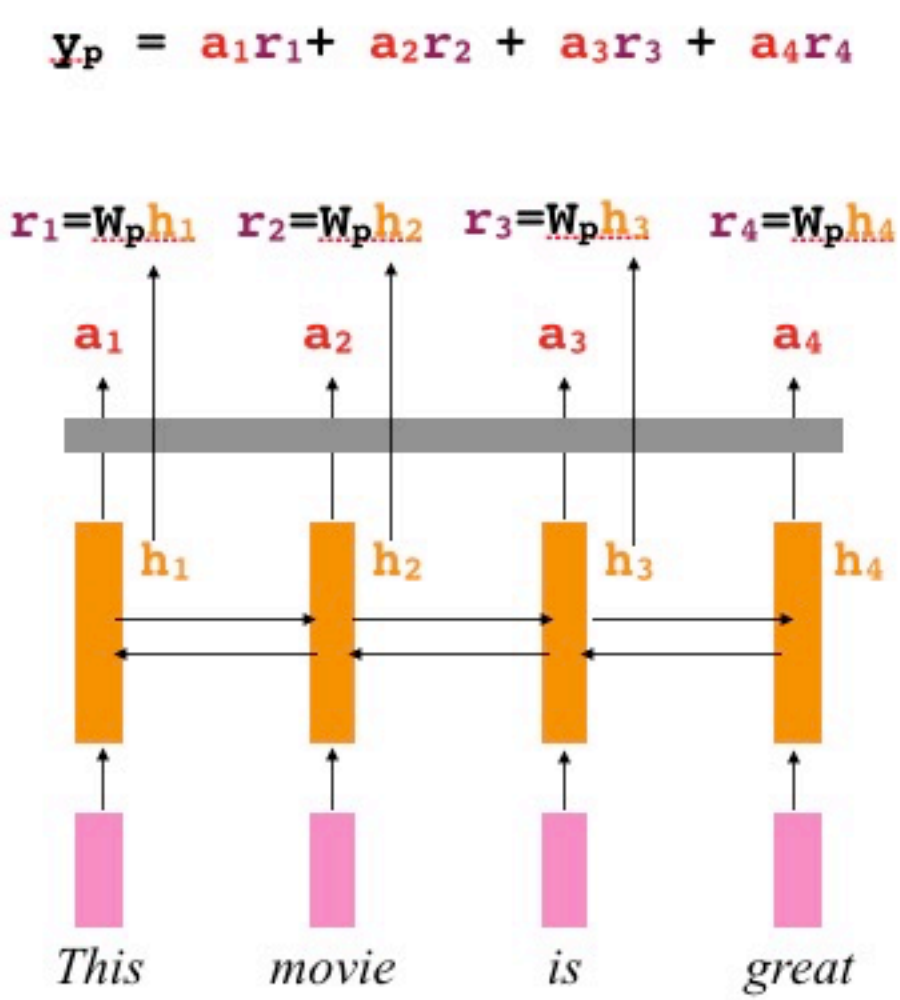
Intuition of Increasing r

$$y_p = 0.25(r_{This} + r_{movie} + r_{is} + r_{great}) \longrightarrow \textit{Positive}\text{ Logit}$$

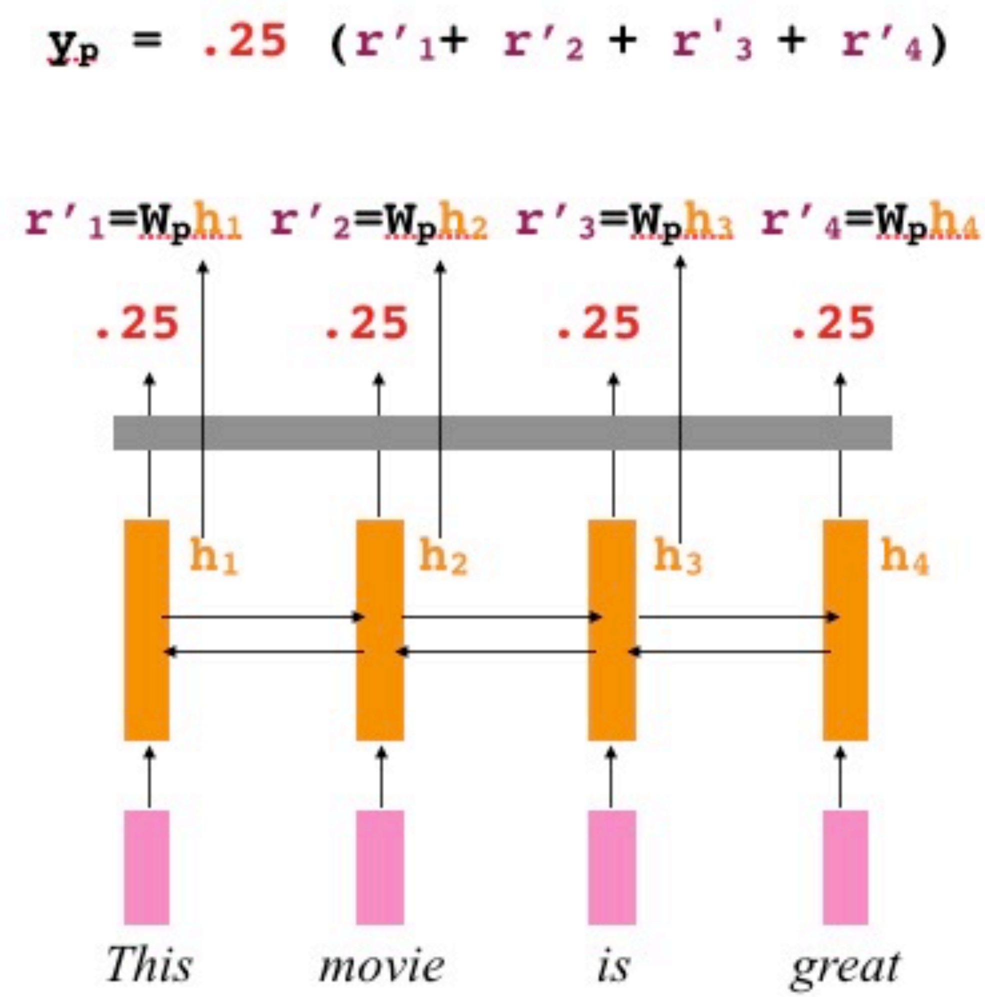
$$y_n = -0.25(r_{This} + r_{movie} + r_{is} + r_{bad}) \longrightarrow \textit{Negative}\text{ Logit}$$

- r_{This}, r_{movie} and r_{is} remains roughly unchanged (cancels out)
- r_{great} increases

Increasing r Under Uniform Attention



standard training

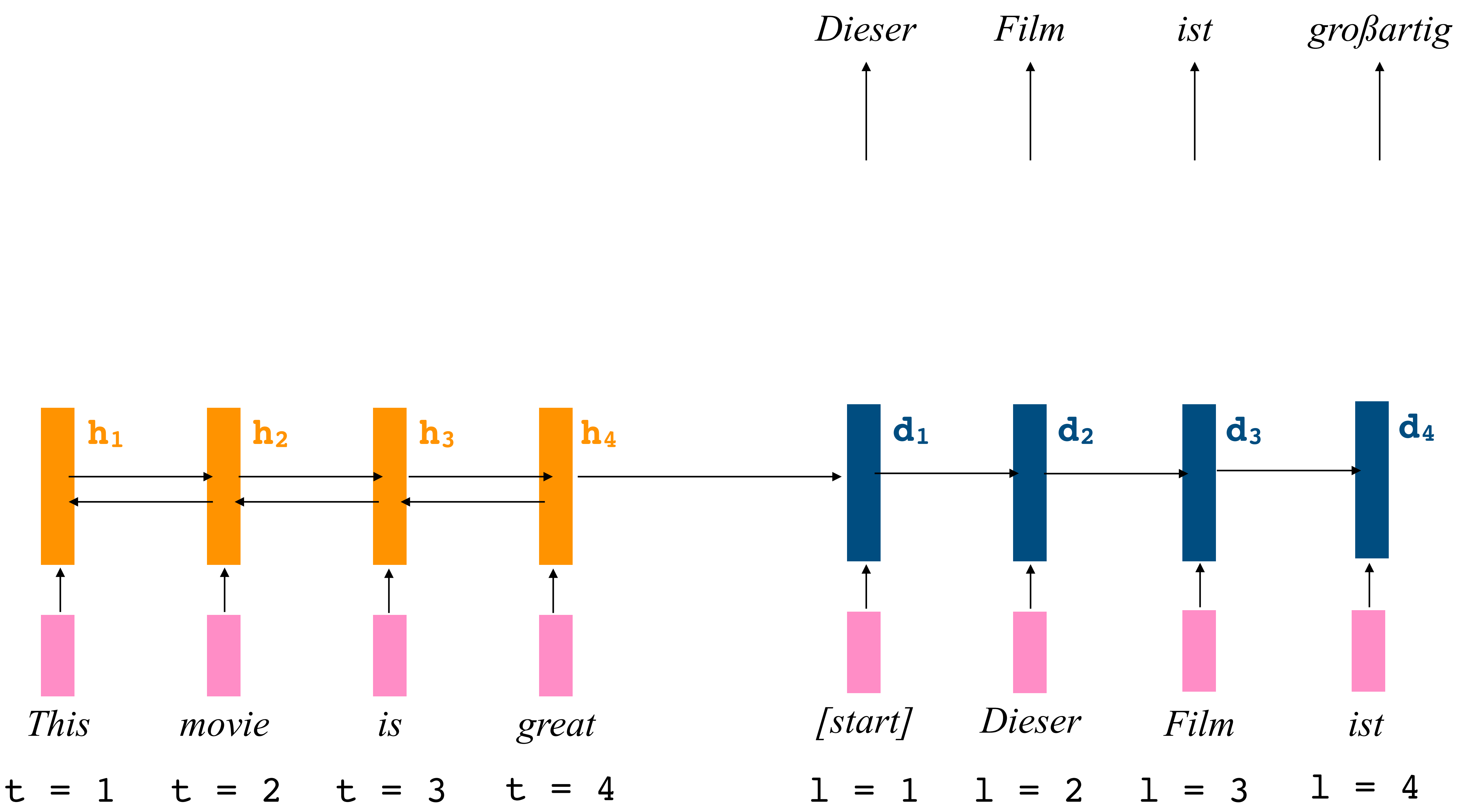


uniform attention training

- $\mathbf{r'}$ still correlates with standard attention $\mathbf{a}$
- $\mathbf{r'}$ correlates with individual token influence.

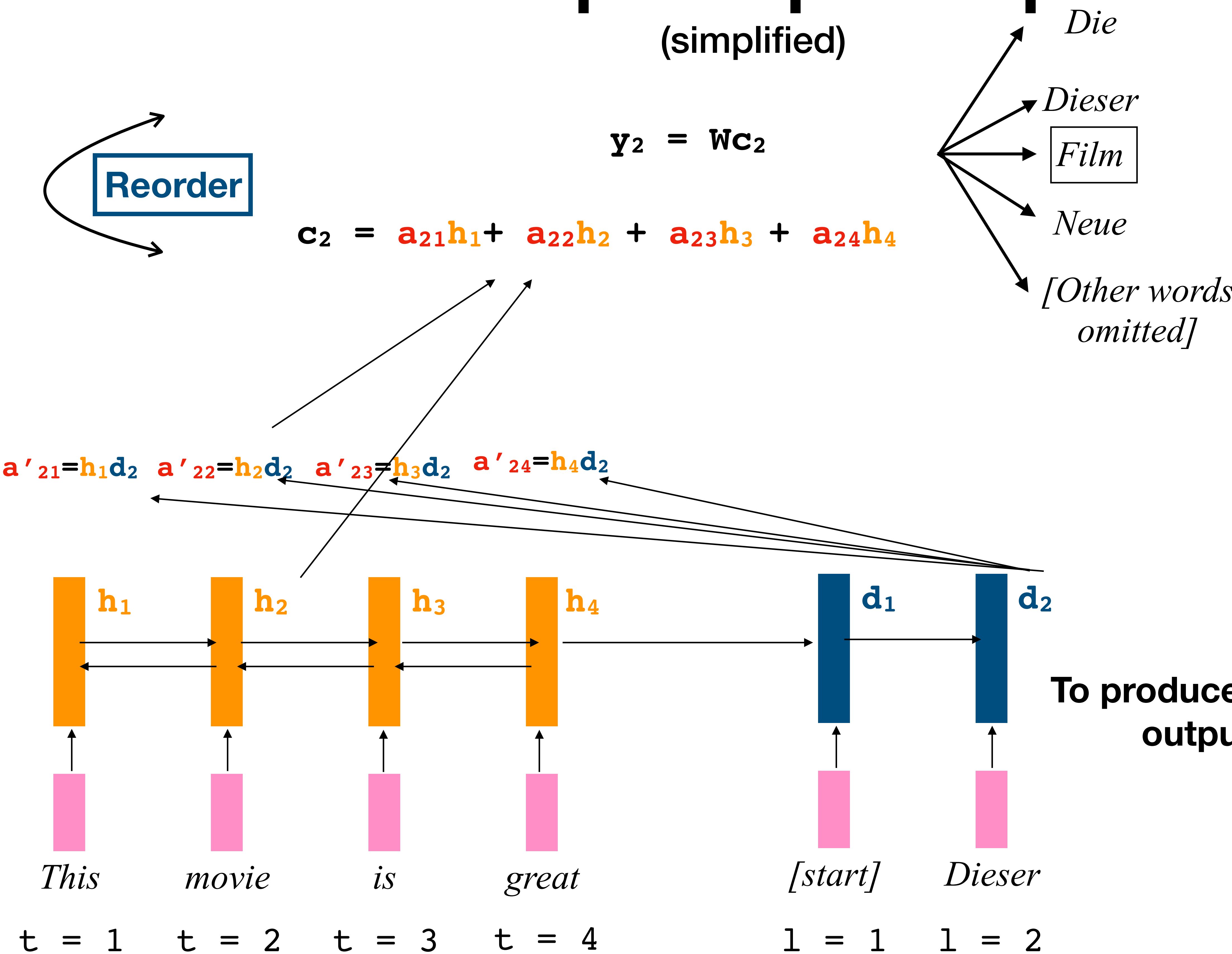
Seq2Seq Setup

(simplified)



Seq2Seq Setup

(simplified)



To produce the second output word

Seq2Seq Output Relevance

(simplified)

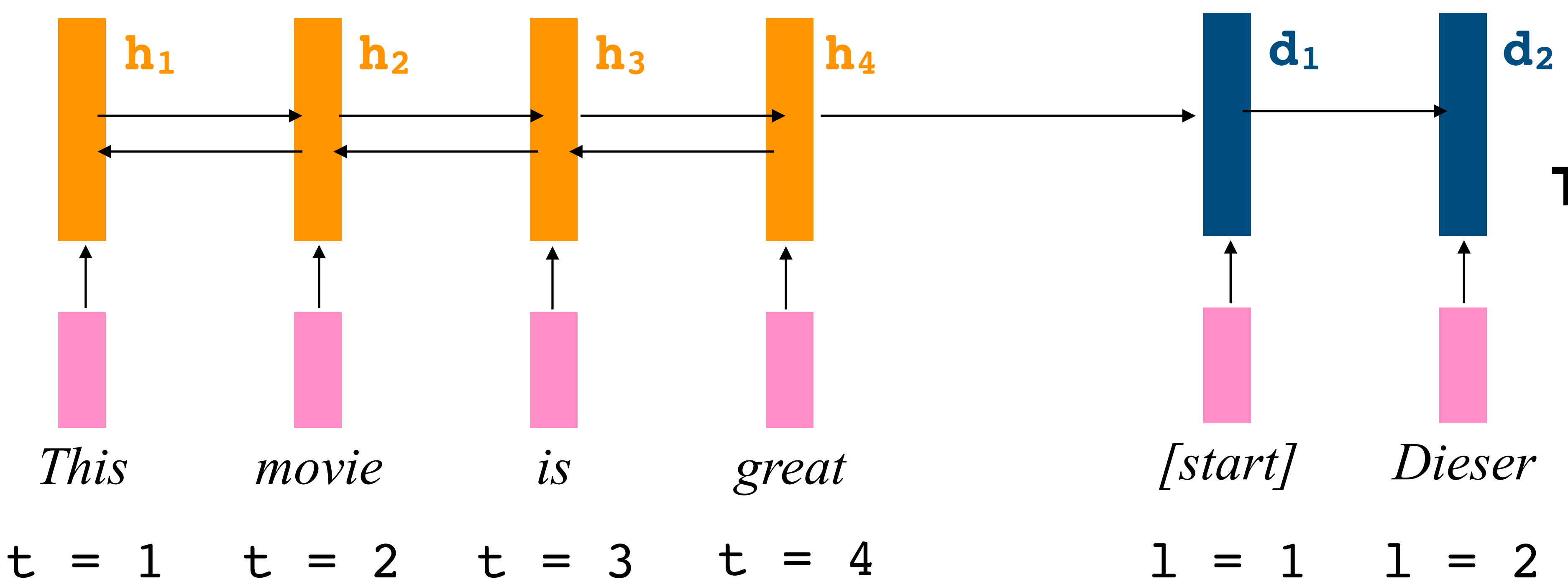
Training objective optimizes $\mathbf{y}_{2,\mathtt{Film}}$

Output Relevance $r_{\mathrm{Film},\,2}$: how much does the model associate h_2 with the word "Film"

$$\mathbf{y}_{2,\mathtt{Film}} = \mathbf{a}_{21}\mathbf{r}_{\mathtt{Film},1} + \mathbf{a}_{22}\mathbf{r}_{\mathtt{Film},2} + \mathbf{a}_{23}\mathbf{r}_{\mathtt{Film},3} + \mathbf{a}_{24}\mathbf{r}_{\mathtt{Film},4}$$

$$\mathbf{r}_{\mathtt{Film},t} := \mathbf{W}_{\mathtt{Film}}\, \mathbf{h}_t$$

$\mathbf{a}'_{21} = \mathbf{h}_1\mathbf{d}_2$ $\mathbf{a}'_{22} = \mathbf{h}_2\mathbf{d}_2$ $\mathbf{a}'_{23} = \mathbf{h}_3\mathbf{d}_2$ $\mathbf{a}'_{24} = \mathbf{h}_4\mathbf{d}_2$

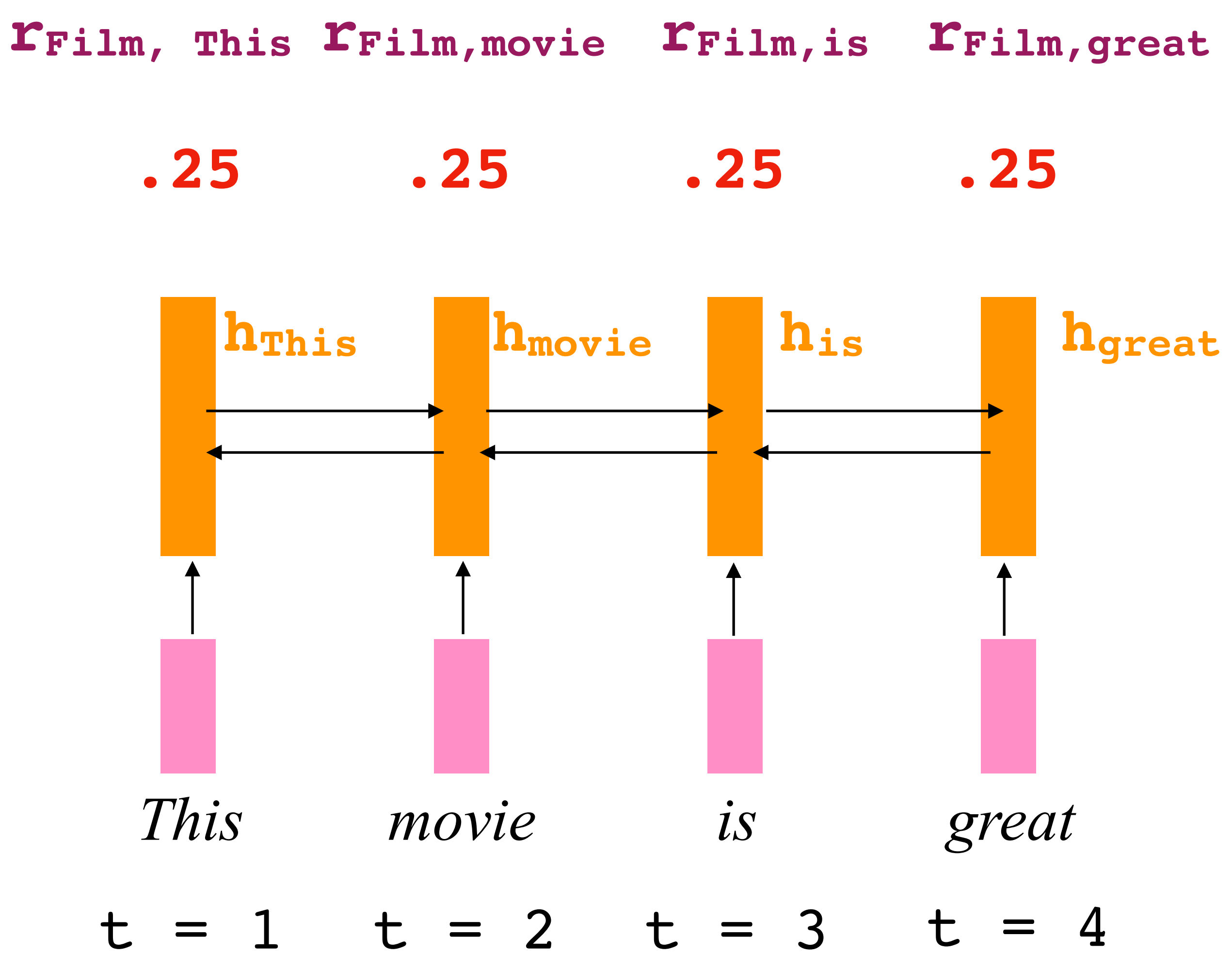


To produce the second output token

Intuition of Learning r, “Dictionary"

Gradient update to maximize:

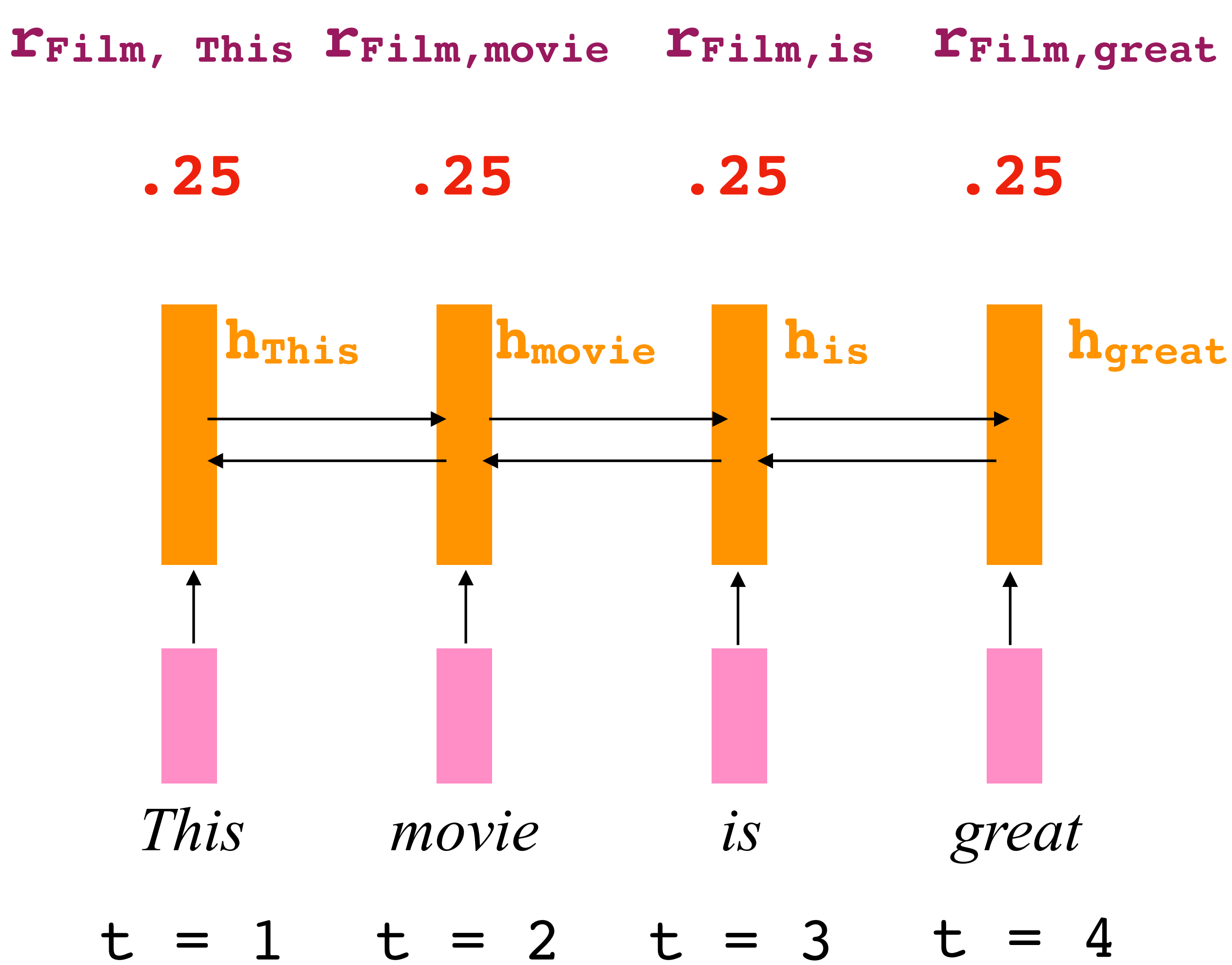
Likelihood(l = 2) = .25($r_{Film, This}$ + $r_{Film,movie}$ + $r_{Film,is}$ + $r_{Film,great}$)



Intuition of Learning r, "Dictionary"

Gradient update to maximize: Likelihood(l = 2) = $\Sigma_t\, r_{Film,t}$

t ∈ {This, movie, is, great }



Intuition of Learning r, "Dictionary"

This movie is great
=> Dieser Film ist großartig

Gradient update to maximize:

Loss(sentence) $\sum_{l} \sum_{t} r_{l,t}$

$t \in$ **{This, movie, is, great }**

$l \in$ **{Dieser, Film, ist, großartig}**

How many times $r_{l,t}$ appears in the sum.

r	This	Movie	Is	Great
Dieser	1	1	1	1
Film	1	1	1	1
Ist	1	1	1	1
großartig	1	1	1	1

Looks like there is no reason
for the model to learn word-to-word correspondence

Intuition of Learning r, “Dictionary"

Gradient update to maximize: $\Sigma_{\mathtt{l}} \Sigma_t r_{\mathtt{l},t}$
$t \in$ {This, movie, is, great }
$\mathtt{l} \in$ {Dieser, Film, ist, großartig}

This movie is great
=> Dieser Film ist großartig

Gradient update to maximize: $\Sigma_{\mathtt{l}} \Sigma_t r_{\mathtt{l},t}$
$t \in$ {This, movie, is, bad }
$\mathtt{l} \in$ {Dieser, Film, ist, schlecht}

This movie is bad
=> Dieser Film ist schlecht

add

r	This	Movie	Is	Great	Bad
Dieser	2	2	2	1	1
Film	2	2	2	1	1
Ist	2	2	2	1	1
großartig	1	1	1	1	0
schlecht	1	1	1	0	1

Can recover r (word-to-word correspondence) from “co-occurrence”.

Testing the “co-occurrence” Intuition

- Hypothesis: If we remove the co-occurrence statistics, **r** cannot be learned, and hence attention **a** fails to learn.

- Experiments on a “sequence copying task”.

  - Setting 1: the model learns the copying task from a distribution of permutations of [1, 40].

    ~50% fails

  - Setting 2: the model to learns from a distribution of length 40 array, each token is a uniform i.i.d. sample from [1, 40].

    Always successful

```
input:  3 2 1 0 ; output:  3' 2' 1' 0'
input:  0 2 1 3 ; output:  0' 2' 1' 3'
input:  3 2 0 1 ; output:  3' 2' 0' 1'
input:  0 2 1 3 ; output:  0' 2' 1' 3'
input:  3 2 0 1 ; output:  3' 2' 0' 1'
```

```
input:  0 3 3 1 2 ; output:  2' 1' 1' 3' 0'
input:  1 2 2 3 3 ; output:  3' 0' 0' 1' 1'
input:  2 2 1 2 3 ; output:  0' 0' 3' 0' 1'
input:  3 2 1 0 0 ; output:  1' 0' 3' 2' 2’
input:  3 0 2 1 2 ; output:  1' 2' 0' 3' 0'
```

Takeaways …

- Interpretable attention might not be necessary to achieve high accuracy (e.g. in text classification).
- Attention is shaped by training dynamics.
- Open the blackbox of training to understand neural networks.

References

- Jain, Sarthak, and Byron C. Wallace. "Attention is not explanation." *arXiv preprint arXiv:1902.10186* (2019).
- Wiegreffe, Sarah, and Yuval Pinter. "Attention is not not explanation." arXiv preprint arXiv:1908.04626 (2019).
- Snell, Charlie, et al. “Understanding Attention Training via Output Relevance” OpenReview Preprint (2020)