Machine Reading Comprehension with BIDAF

Peerapon Vateekul, Ph.D.

Credit:

- Theerit Lapchaicharoenkit
- Can Udomcharoenchaikit

Extractive Reading Comprehension with BIDAF [2016,2017]

https://arxiv.org/abs/1611.01603?fbclid=IwAR1f6eDp7LkaezaM0frHFFuB1LyJTfFxhHl16YOI2aamPaZGZHCghC9Eigg

Bidirectional Attention Flow for Machine Comprehension

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, Hannaneh Hajishirzi

(Submitted on 5 Nov 2016 (v1), last revised 21 Jun 2018 (this version, v6))

Machine comprehension (MC), answering a query about a given context paragraph, requires modeling complex interactions between the context and the query. Recently, attention mechanisms have been successfully extended to MC. Typically these methods use attention to focus on a small portion of the context and summarize it with a fixed-size vector, couple attentions temporally, and/or often form a uni-directional attention. In this paper we introduce the Bi-Directional Attention Flow (BIDAF) network, a multi-stage hierarchical process that represents the context at different levels of granularity and uses bi-directional attention flow mechanism to obtain a query-aware context representation without early summarization. Our experimental evaluations show that our model achieves the state-of-the-art results in Stanford Question Answering Dataset (SQuAD) and CNN/DailyMail cloze test.

Comments: Published as a conference paper at ICLR 2017
Subjects: **Computation and Language (cs.CL)**
Cite as: arXiv:1611.01603 [cs.CL]
(or arXiv:1611.01603v6 [cs.CL] for this version)

BIDAF

https://arxiv.org/abs/1611.01603

- Debut on SQuAD, 2016
- Machine Reading Comprehension Model from Allennlp
- The layers composes of
  1. Embedding layer (character + word)
  2. Contextual embedding layer (LSTM)
  3. Bidirectional attentional flow layer
  4. Modelling layer (LSTM)
  5. Prediction Layer
- Make use bi-directional attention flow to allow attention of all timesteps to flow into upper layer rather than summarizing into single vector.

Machine Reading Comprehension – Example

** In our example, answers can always be found in context

Context

In education, teachers facilitate student learning, often in a school or academy or perhaps in another environment such as outdoors. A teacher who teaches on an individual basis may be described as a tutor.

Question	Answer
What is the role of teacher in education?	Facilitate student learning
Where is another indoor location for a teacher other than a school ?	academy
What is the name for a teacher for just one person?	tutor

BIDAF – Input/output Overview

Start token: facilitate

End token: learning

Answer:
facilitate student learning

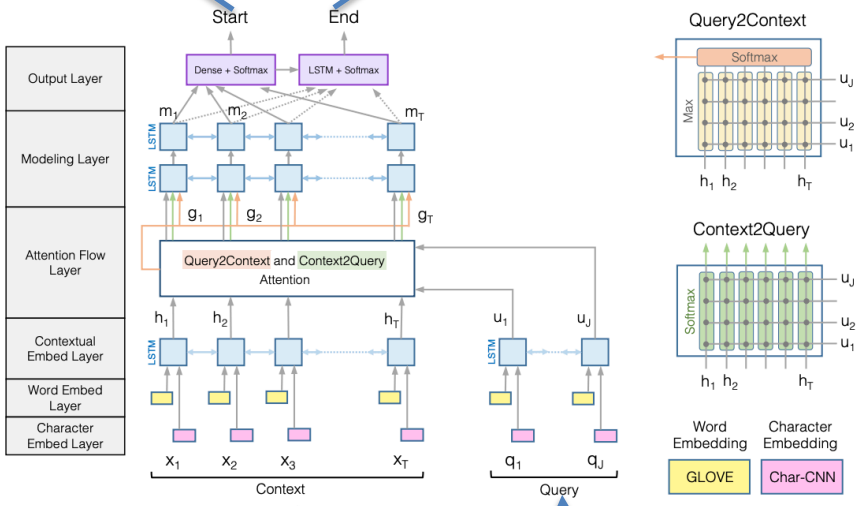


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

Context

Query

In education, teachers facilitate student learning, often in a school or academy or perhaps in another environment such as outdoors. A teacher who teaches on an individual basis may be described as a tutor.

What is the role of teacher in education?

BIDAF

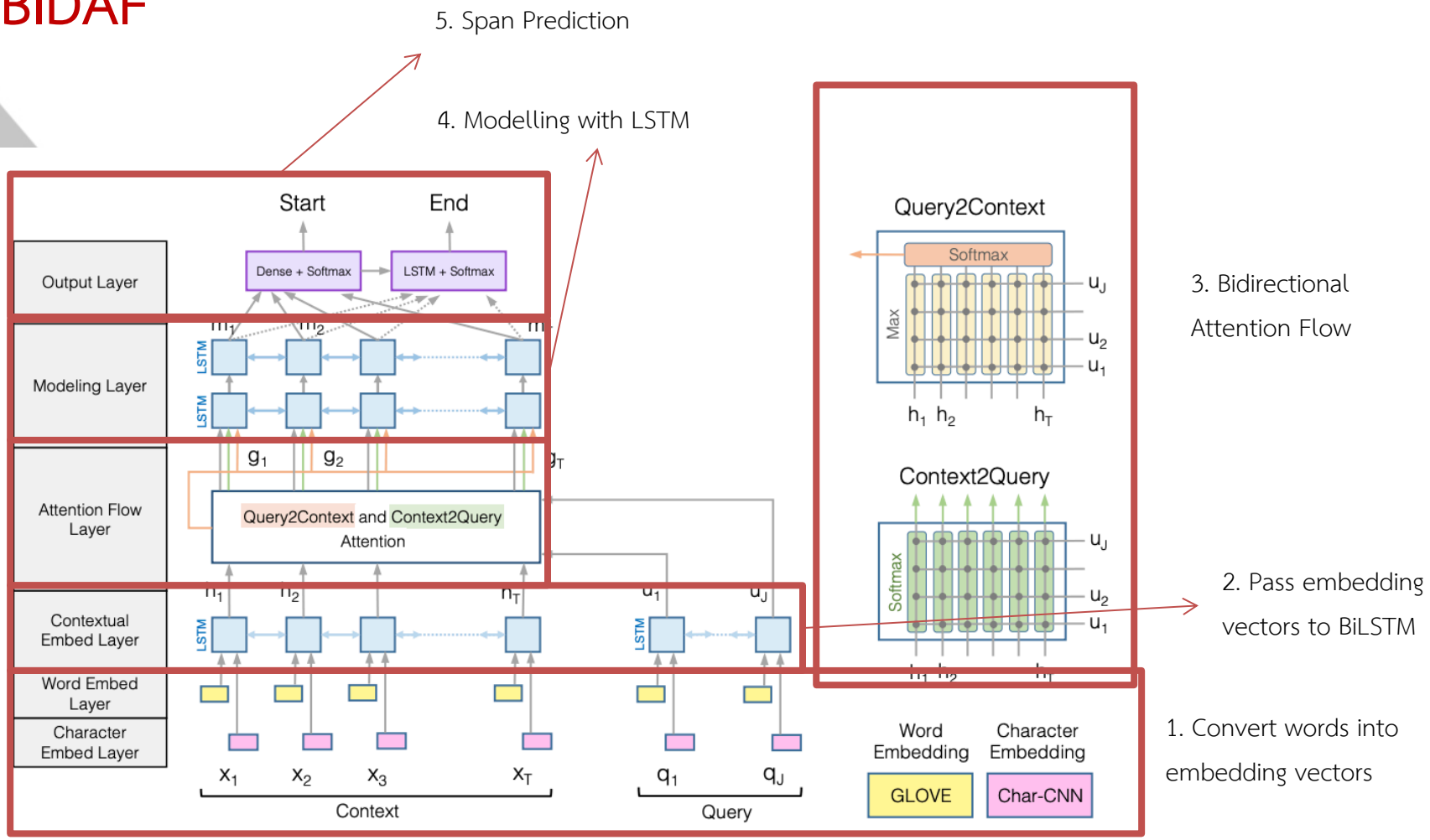


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

In this Presentation…

- We will walkthrough each layer of BIDAF to see how it works when performing machine reading comprehension task
- The steps/layers that we will cover
  - Input to the model
  - Embedding layer
  - Contextual embedding layer
  - Similarity matrix calculation
  - Attentional Flow layer
  - Modelling layer
  - Prediction layer

Input to the model

BIDAF - Input (1): Tokenize context and question

** Answer can be tokenized later during evaluation

Context

In education, teachers facilitate student learning, often in a school or academy or perhaps in another environment such as outdoors. A teacher who teaches on an individual basis may be described as a tutor.

Tokens

In	education	,	teachers	facilitate
student	learning	,	often	in
a	school	or	academy	or
perhaps	in	another	environment	such
as	outdoors	.	....	...
....	as	a	tutor	.

Question

What is the role of teacher in education?

Tokens

What	is	the	role	.....	education	?

Where is another indoor location for a teacher other than a school ?

Where	is	another	indoor	.....	school	?

What is the name for a teacher for just one person?

What	is	the	name	.....	person	?

BIDAF - Input (2): Map ground truth positions

**The dataset normally provides start and end position as character index

***We will have to map these character index position into token positions as we normally put the sequence of tokens as input of the model.

Answer	Start Position Index	End Position Index
Facilitate student learning	23	50
academy	73	80
tutor	200	205

BIDAF - Input (3): Token positions and index positions

** Character position span is usually mapped together with token positions

Context

In education, teachers facilitate student learning, often in a school or academy or perhaps in another environment such as outdoors. A teacher who teaches on an individual basis may be described as a tutor.

Tokens

In	education	,	teachers	facilitate
student	learning	,	often	in
a	school	or	academy	or
perhaps	in	another	environment	such
as	outdoors	.	....	...
....	as	a	tutor	.

Token positions

0	1	2	3	4
5	6	7	8	9
10	11	12	13	14
15	16	17	18	19
20	21	22	....	...
....	34	35	36	37

Character span

(0, 2)	(3, 12)	(12, 13)	(14, 22)	(23, 33)
(34, 41)	(42, 50)	(50, 51)	(52, 57)	(58, 60)
(61, 62)	(63, 69)	(70, 72)	(73, 80)	(81, 83)
(84, 91)	(92, 94)	(95, 102)	(103, 114)	(115, 119)
(120, 122)	(123, 131)	(131, 132)	....	...
....	(195, 197)	(198, 199)	(200, 205)	(205, 206)

BIDAF - Input (4): Mark ground truth labels in context passage

** Now we can map the ground truth answer character span into token spans.

Answer	Character position	Token position
Facilitate student learning	23, 50	4, 6
academy	73, 80	13
tutor	200, 205	36

Tokens

In	education	,	teachers	facilitate
student	learning	,	often	in
a	school	or	academy	or
perhaps	in	another	environment	such
as	outdoors	.	....	...
....	as	a	tutor	.

Token positions

0	1	2	3	4
5	6	7	8	9
10	11	12	13	14
15	16	17	18	19
20	21	22	....	...
....	34	35	36	37

Character span

(0, 2)	(3, 12)	(12, 13)	(14, 22)	(23, 33)
(34, 41)	(42, 50)	(50, 51)	(52, 57)	(58, 60)
(61, 62)	(63, 69)	(70, 72)	(73, 80)	(81, 83)
(84, 91)	(92, 94)	(95, 102)	(103, 114)	(115, 119)
(120, 122)	(123, 131)	(131, 132)	....	...
....	(195, 197)	(198, 199)	(200, 205)	(205, 206)

BIDAF - Input (5): Input as tensors

- Tensors that will be used in the model are question and context tokens.
- Most library will other metadata are also mostly passed into the model as well.
- 1 set of learning instance will mainly consist of context tokens, question tokens, span start/end and other metadata

1 learning instance

Tokens (context) →
Tensor of [Batch size * context token length (padded)]

In	education	,	teachers	facilitate
student	learning	,	often	in
a	school	or	academy	or
perhaps	in	another	environment	such
as	outdoors	.	....	...
....	as	a	tutor	.

Tokens (question) →
Tensor of [Batch size * question token length (padded)]

What	is	the	role	.....	education	?

Other metadata

Character position

Token position

Token positions

Character span

Embedding Layer

Embedding Layer

- Map context/question tokens tensors to appropriate word vectors
  - Tokens are mapped to token_id before they are actually mapped to word vectors
- Hidden dimension depends on word vector and character embedding dimension size
- Concatnate word and character embedding together
- Input
  - Question : Batch size * length of questions tokens (padded) (J)
  - Passage : Batch size * length of context tokens (padded) (T)
- Output
  - Question (Denote as Q) : Batch size * length of questions (padded) (J) * Hidden dim (d1)
  - Passage (Denote as X) : Batch size * length of context (padded) (T) * Hidden dim (d1)

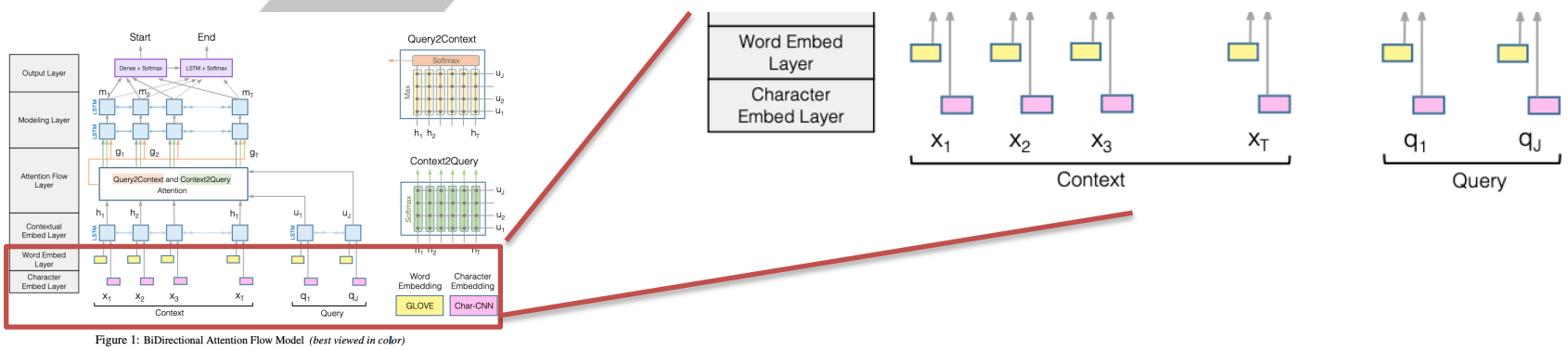
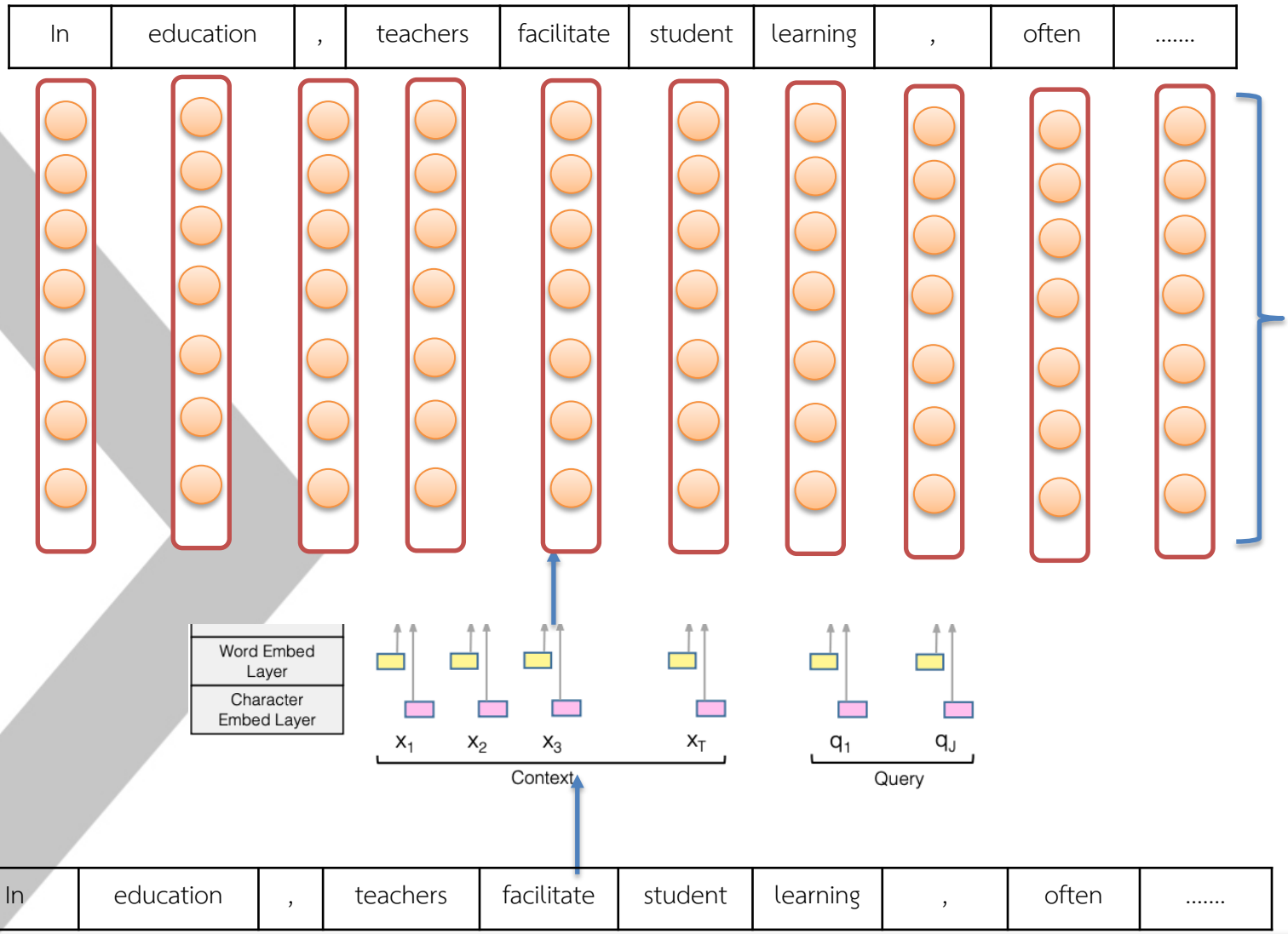


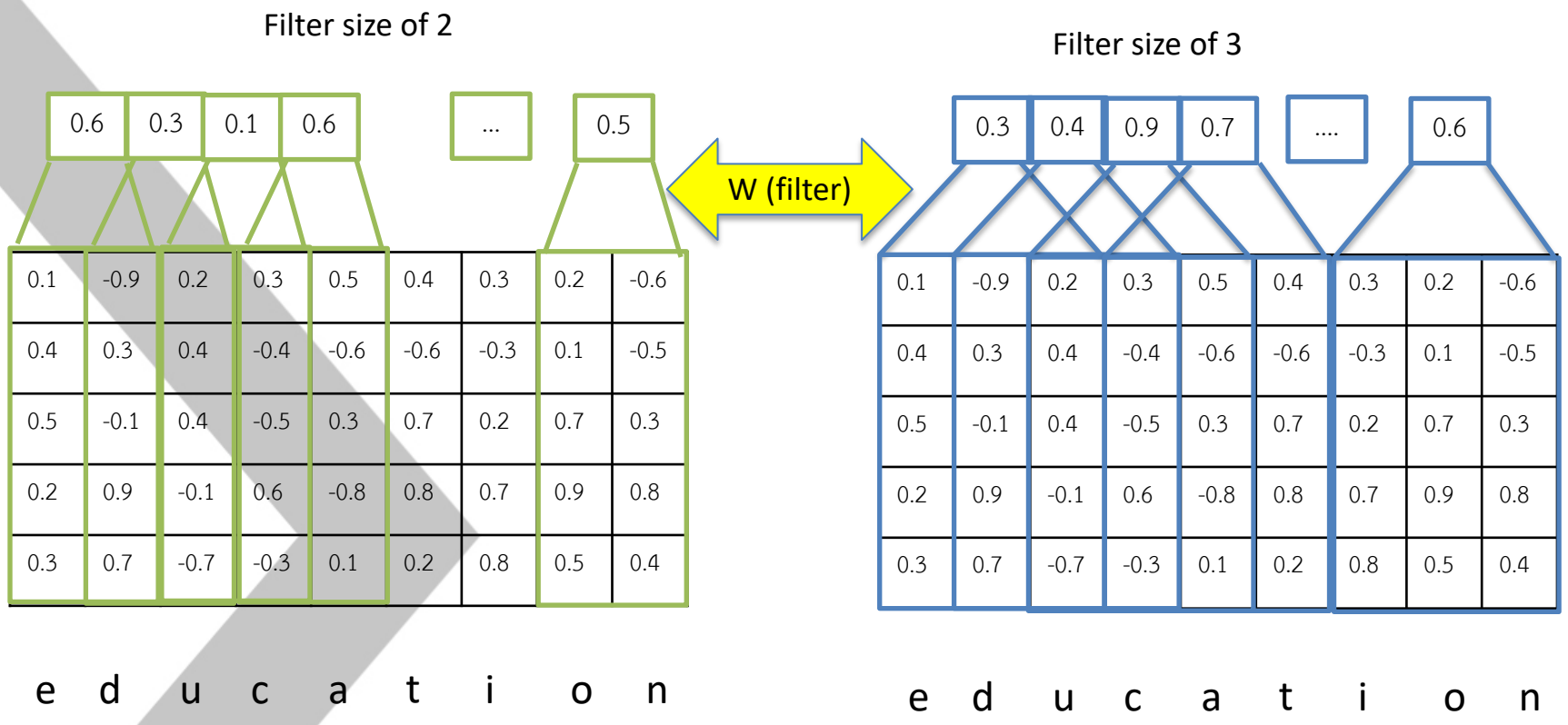
Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

(Word) Embedding Layer – visualization



(Character) Embedding Layer – Visualization (1)

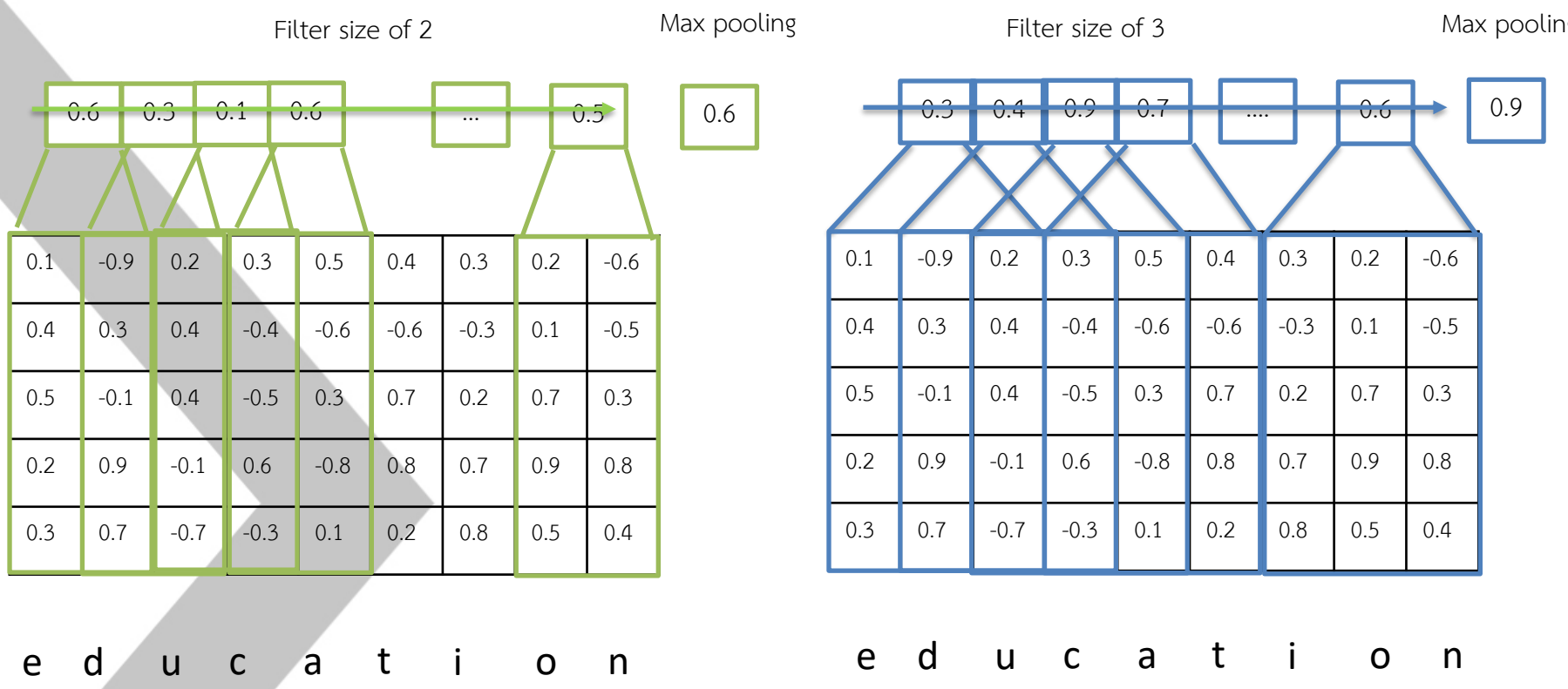
** Do a convolutional filter over each alphabet



https://towardsdatascience.com/the-definitive-guide-to-bidaf-part-2-word-embedding-character-embedding-and-contextual-c151fc4f05bb

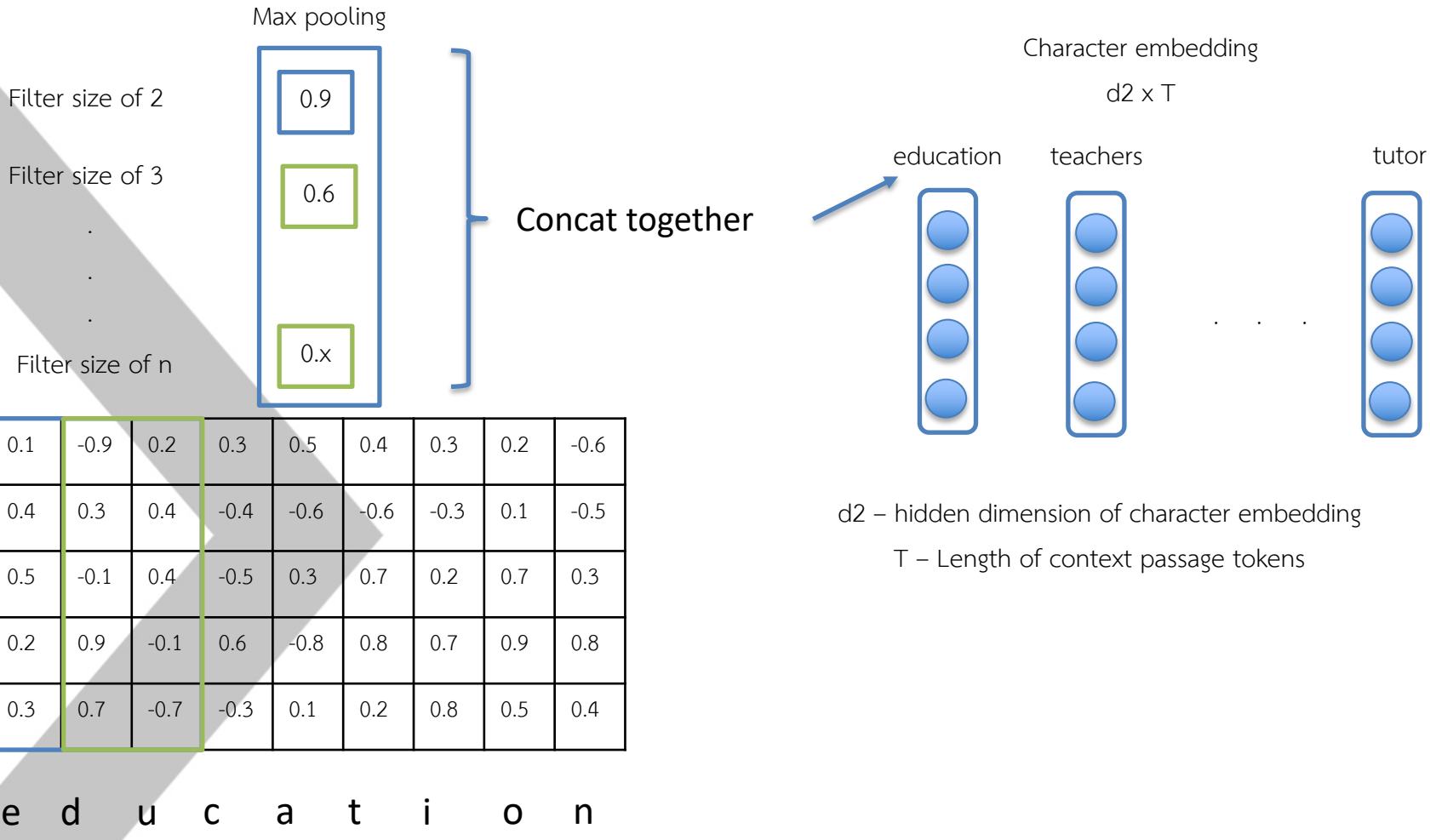
(Character) Embedding Layer – Visualization (2)

** Perform max pooling to select single scalar value from each filter pass



(Character) Embedding Layer – Visualization (3)

** Perform max pooling to select single scalar value from each filter pass

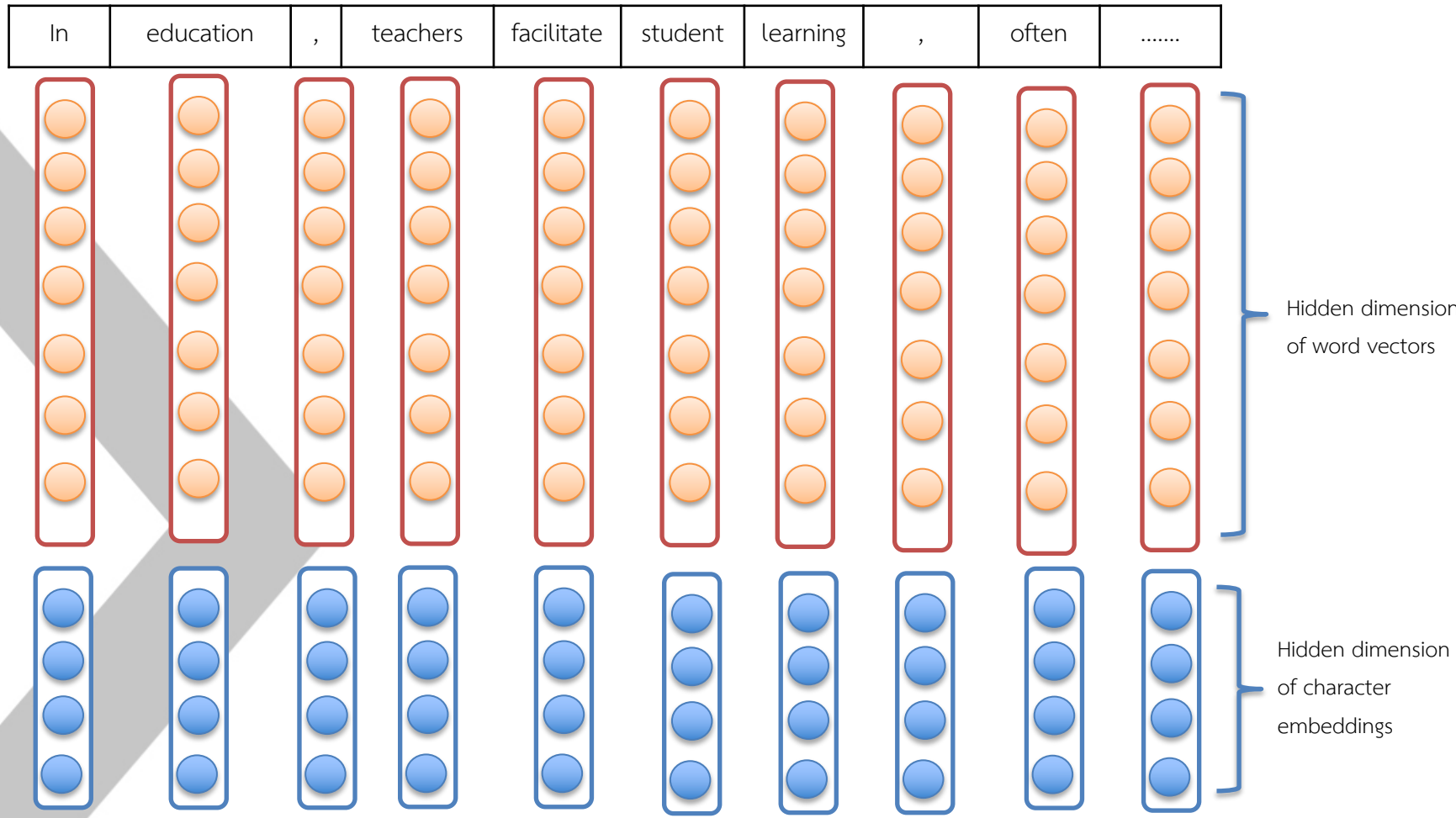


d2 – hidden dimension of character embedding

T – Length of context passage tokens

https://towardsdatascience.com/the-definitive-guide-to-bidaf-part-2-word-embedding-character-embedding-and-contextual-c151fc4f05bb

Embedding Layer – Combine word and character as output



Contextual Embedding Layer

Context Embedding ($H_{:t}$, $U_{:t}$)

- RNN layers that captures contextual information from sequence of word vectors
- Input (from embedding layer)
  - Question (Denote as Q): Batch size * biggest length of questions (J) * Hidden dim (d1 + d2)
  - Passage (Denote as X) : Batch size * biggest length of context (T) * Hidden dim (d1 + d2)
- Output
  - Question (Denote as $U_{:t}$) : Batch size * biggest length of questions (J) * 2Encoded Dim
  - Passage (Denote as $H_{:t}$) : Batch size * biggest length of context (T) * 2Encoded Dim
  - ***Times two the encoding dimension because of bidirectional LSTM**

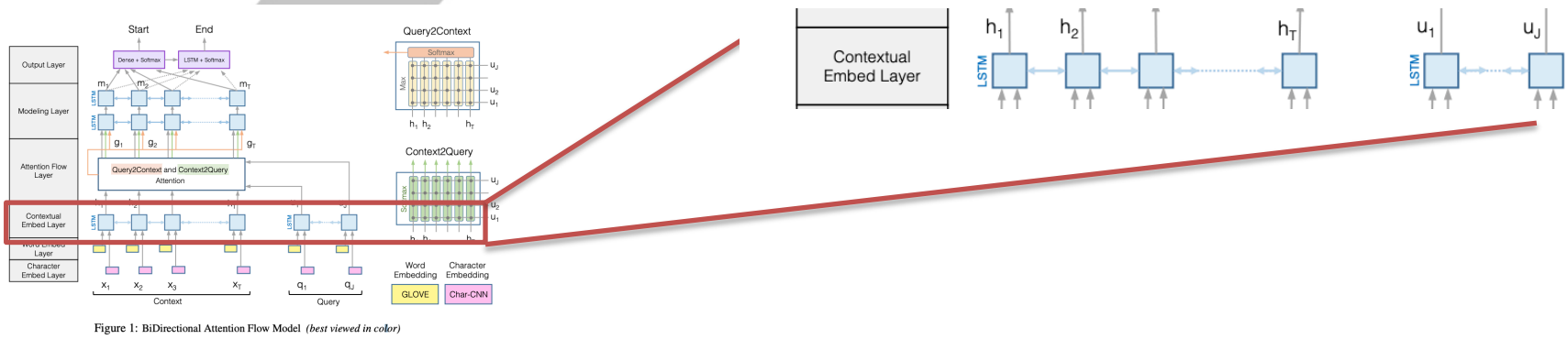
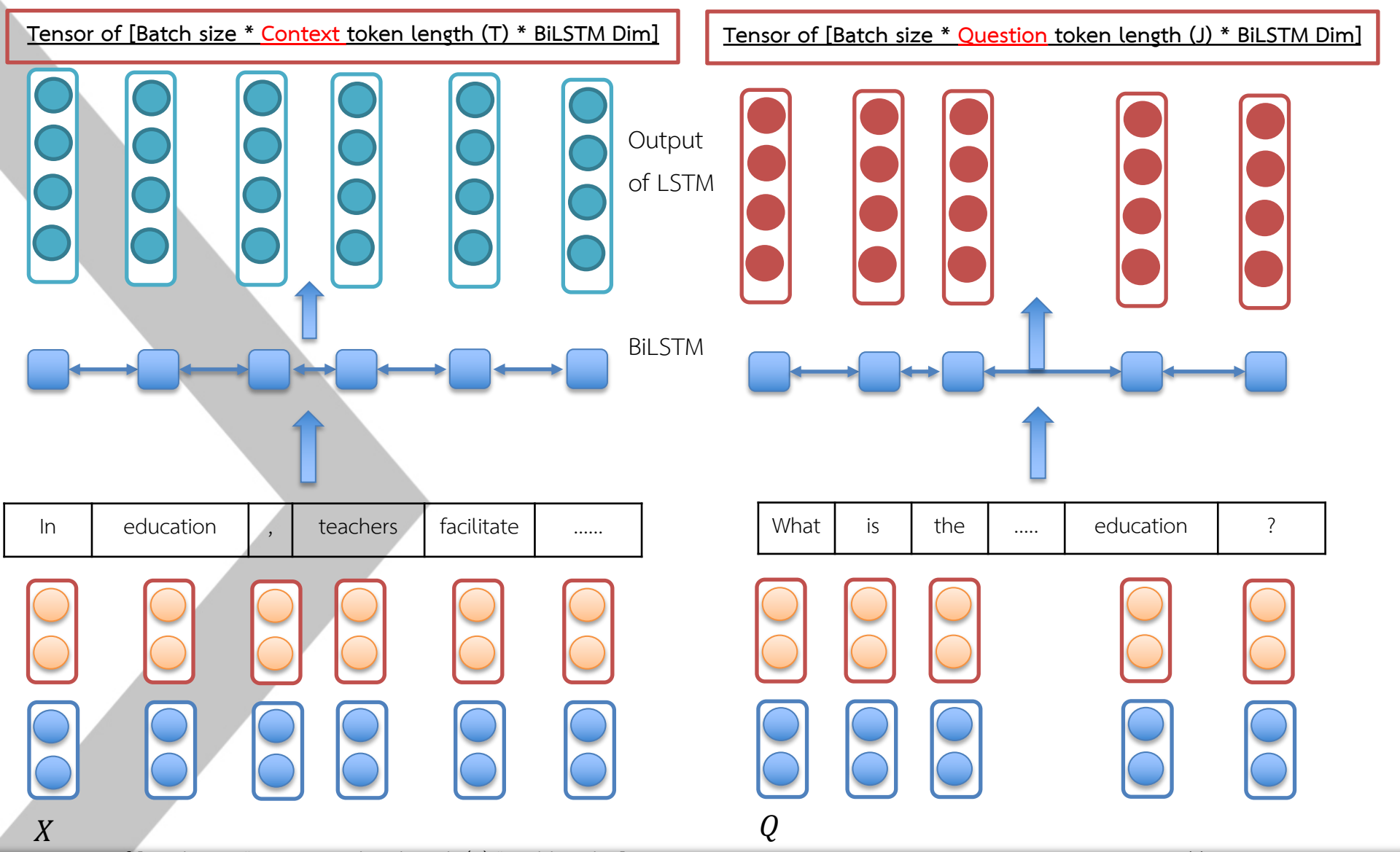


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

Context Embedding

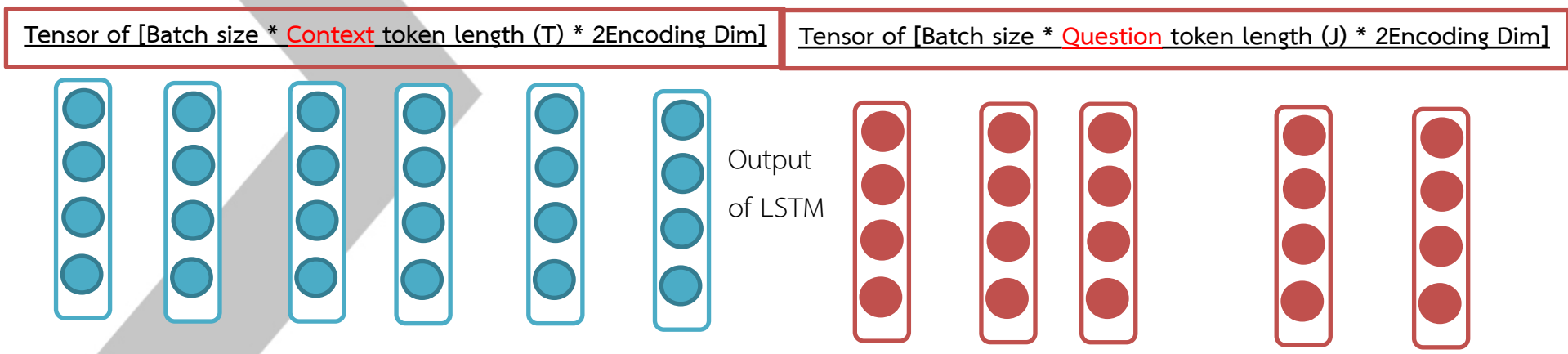


Context Embedding

- The output of BiLSTM in contextual embedding layer can be viewed as hidden representations of tokens in context and questions that captures contextual information by letting the BiLSTM read and process the sequence of tokens found.

Output of BiLSTM from context side: $H_{:t}$

Output of BiLSTM from question side: $U_{:t}$



Similarity Matrix Calculation

Similarity Matrix Calculation

- In BIDAF, the author propose to calculate similarity matrix by the following equation
  - a is context representation vector ($H_{:t}$)
  - b is question representation vector ($U_{:t}$)

$$\alpha(\mathbf{a}, \mathbf{b}) = w^T \times [\mathbf{a} ; \mathbf{b} ; \mathbf{a} \circ \mathbf{b}],$$ where...

- **a** and **b** are the two input vectors
- $w^T \in \mathbb{R}^{6d}$ is a trainable weight vector
- $\circ$ is element wise multiplication
- $[\ ;\ ]$ is vector concatenation across row

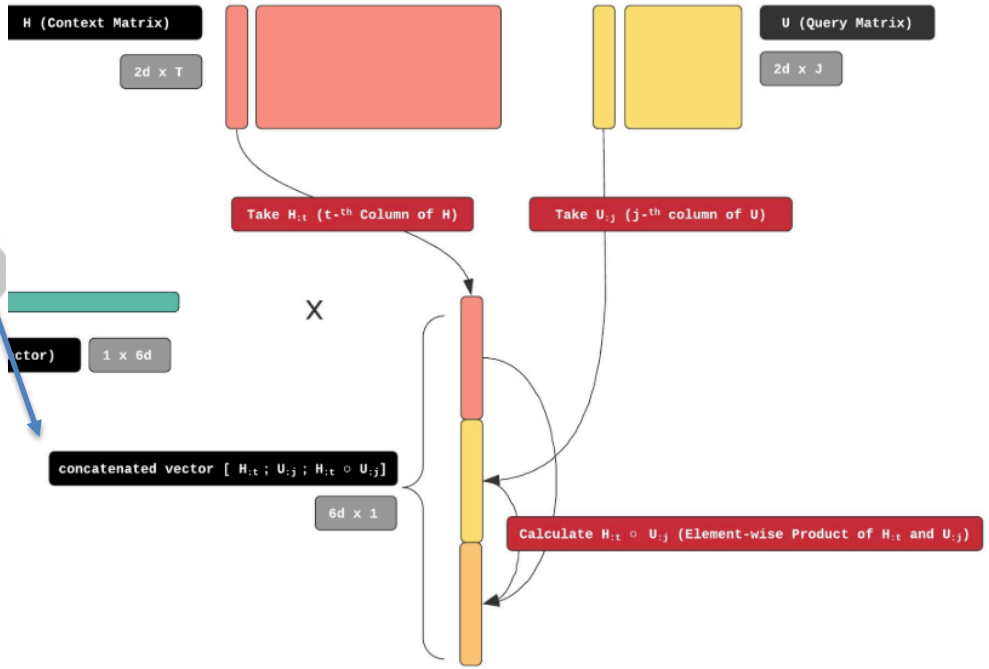
- Source: https://towardsdatascience.com/the-definitive-guide-to-bidaf-part-3-attention-92352bbdcb07

Similarity Matrix Calculation

$\alpha(\mathbf{a}, \mathbf{b}) = w^T \times [\mathbf{a} ; \mathbf{b} ; \mathbf{a} \circ \mathbf{b}]$, where...

- **a** and **b** are the two input vectors
- $w^T \in \mathbb{R}^{6d}$ is a trainable weight vector
- $\circ$ is element wise multiplication
- $[;]$ is vector concatenation across row

- Just pick vector element (represent tokens)
- 2d is the encoding dimension from contextual embedding layer
- T is number of context tokens
- J is number of question tokens
- H represents context matrix
- U represents question matrix



Step 6. Using the similarity function α, we combine context matrix **H** and query matrix **U** to form similarity matrix **S**.

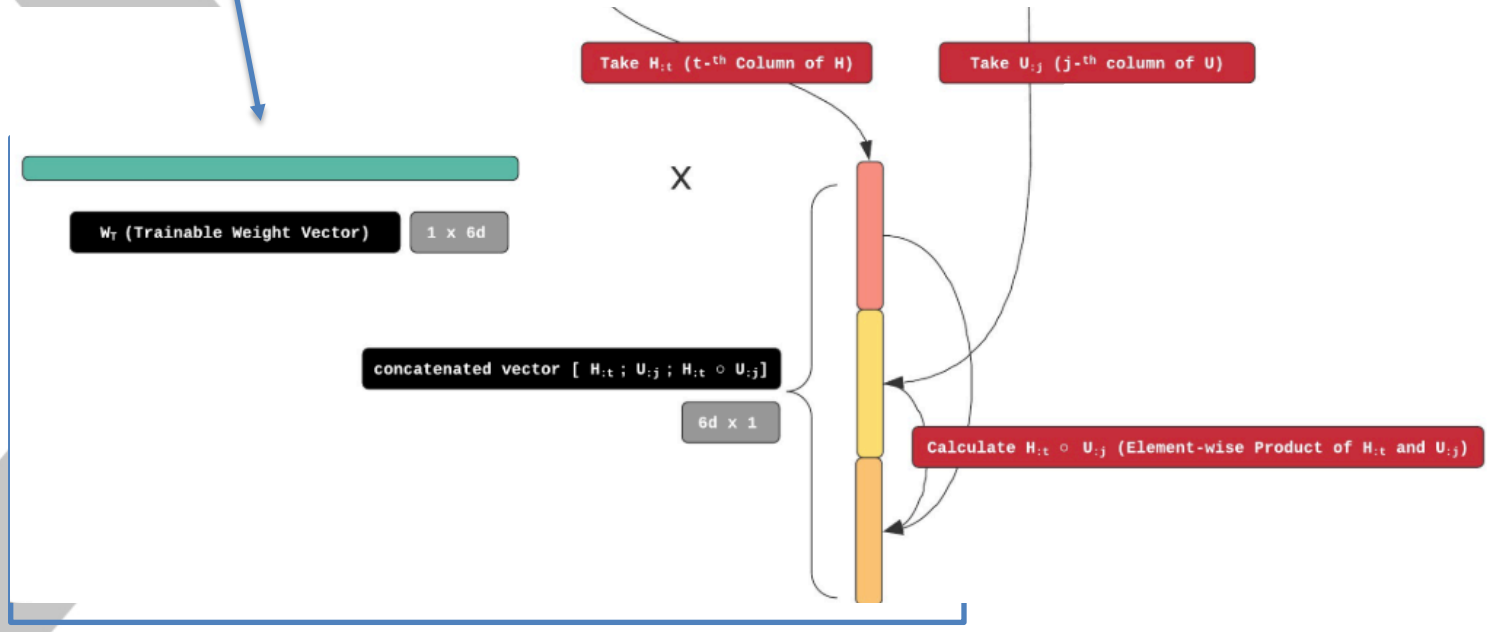
- Source: https://towardsdatascience.com/the-definitive-guide-to-bidaf-part-3-attention-92352bbdcb07

Similarity Matrix Calculation

$\alpha(\mathbf{a}, \mathbf{b}) = w^T \times [\mathbf{a} ; \mathbf{b} ; \mathbf{a} \circ \mathbf{b}]$, where...

- $\mathbf{a}$ and $\mathbf{b}$ are the two input vectors
- $w^T \in \mathbb{R}^{6d}$ is a trainable weight vector
- $\circ$ is element wise multiplication
- $[\ ;\ ]$ is vector concatenation across row

- Just pick vector element (represent tokens)
- 2d is the encoding dimension from contextual embedding layer
- T is number of context tokens
- J is number of question tokens
- H represents context matrix
- U represents question matrix



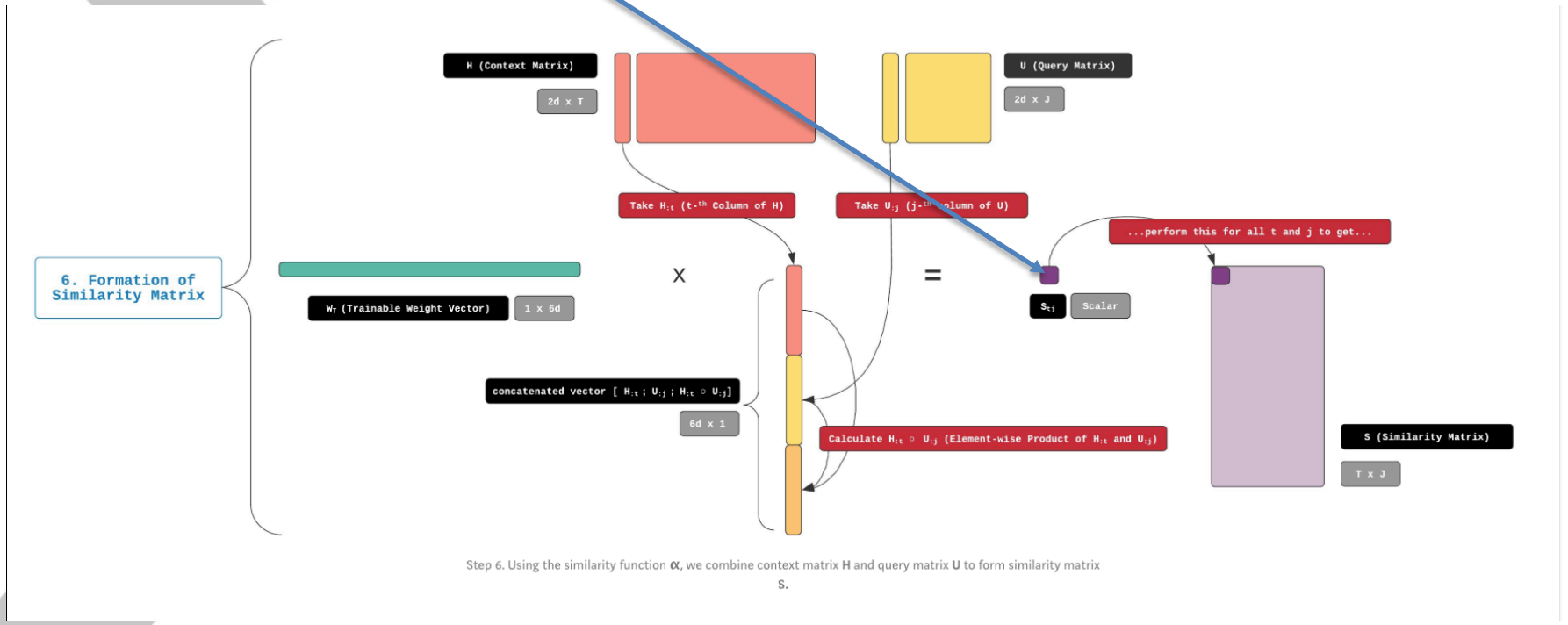
- Source: https://towardsdatascience.com/the-definitive-guide-to-bidaf-part-3-attention-92352bbdcb07

Similarity Matrix Calculation

$\alpha(\mathbf{a}, \mathbf{b}) = w^T \times [\mathbf{a} ; \mathbf{b} ; \mathbf{a} \circ \mathbf{b}]$, where...

- **a** and **b** are the two input vectors
- $w^T \in \mathbb{R}^{6d}$ is a trainable weight vector
- $\circ$ is element wise multiplication
- $[\ ;\ ]$ is vector concatenation across row

- Just pick vector element (represent tokens)
- 2d is the encoding dimension from contextual embedding layer
- T is number of context tokens
- J is number of question tokens
- H represents context matrix
- U represents question matrix



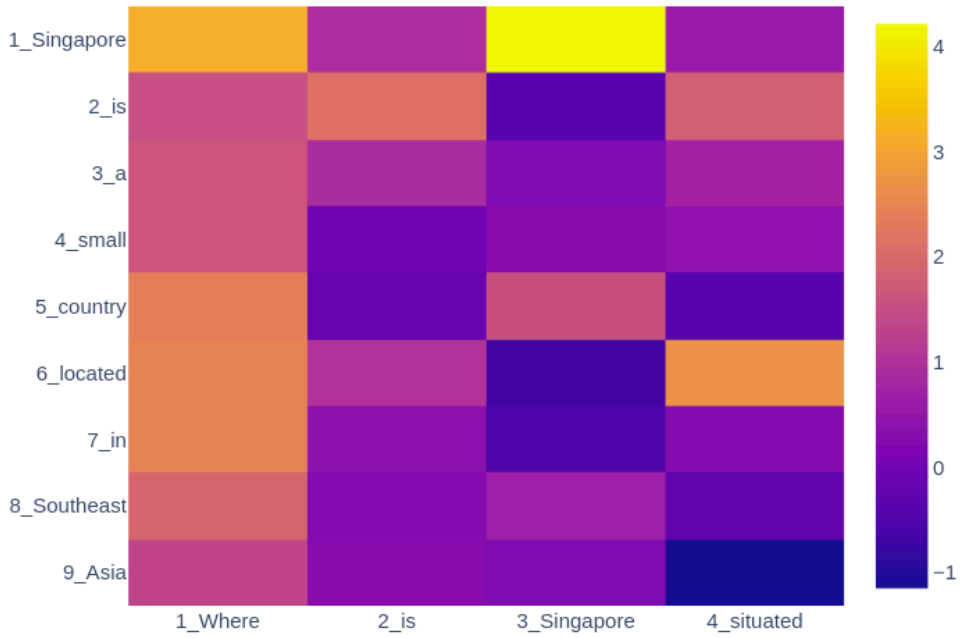
Step 6. Using the similarity function α, we combine context matrix **H** and query matrix **U** to form similarity matrix **S**.

- Source: https://towardsdatascience.com/the-definitive-guide-to-bidaf-part-3-attention-92352bbdcb07

Similarity Matrix Calculation

In the end, we would have a matrix with the size of T (context length) x J (question length) that shows the similarity score between context hidden representation ($H_{:t}$) and question hidden representation ($U_{:t}$)

The range of score here is still arbitrary and will need to be applied with softmax later.

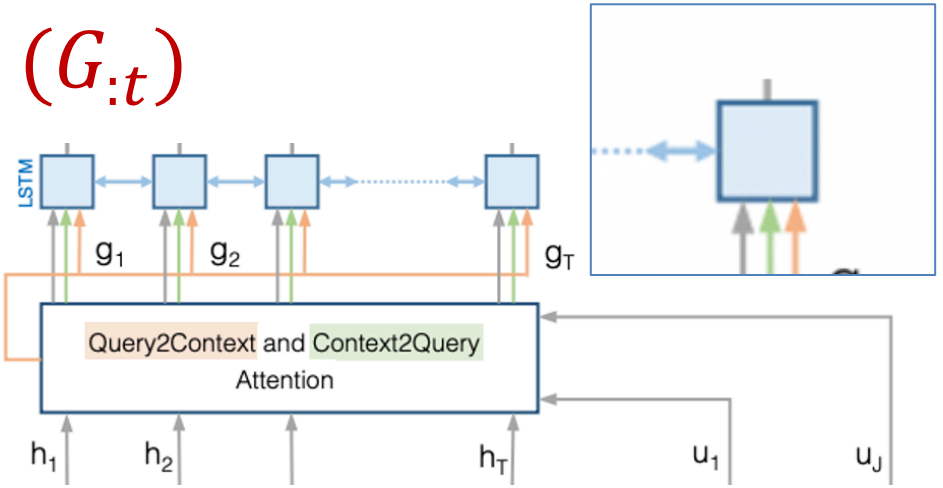


- Source: https://towardsdatascience.com/the-definitive-guide-to-bidaf-part-3-attention-92352bbdcb07

Attentional Flow Layer

Attentional Flow Layer ($G_{:t}$)

- There are 2 directions of attention:
  - Context-to-query attention $\widetilde{U}_{:t}$
  - Query-to-context attention $\widetilde{H}_{:t}$



- **Input** to this layer is the similarity matrix created from context representation vectors ($H_{:t}$) and question representation vectors ($U_{:t}$)

- **The output** of this layer is the output vector from two attention mechanisms ($\widetilde{U}_{:t}$ and $\widetilde{H}_{:t}$) fuse with context representation vectors ($H_{:t}$)
  - Output (Denote as $G_{:t}$) : Batch size * length of questions (padded) (J) * 8Encoding dim(d1)

Context-to-query-attention $\tilde{U}_{:t}$

- Each context tokens gauges importance of each question token.

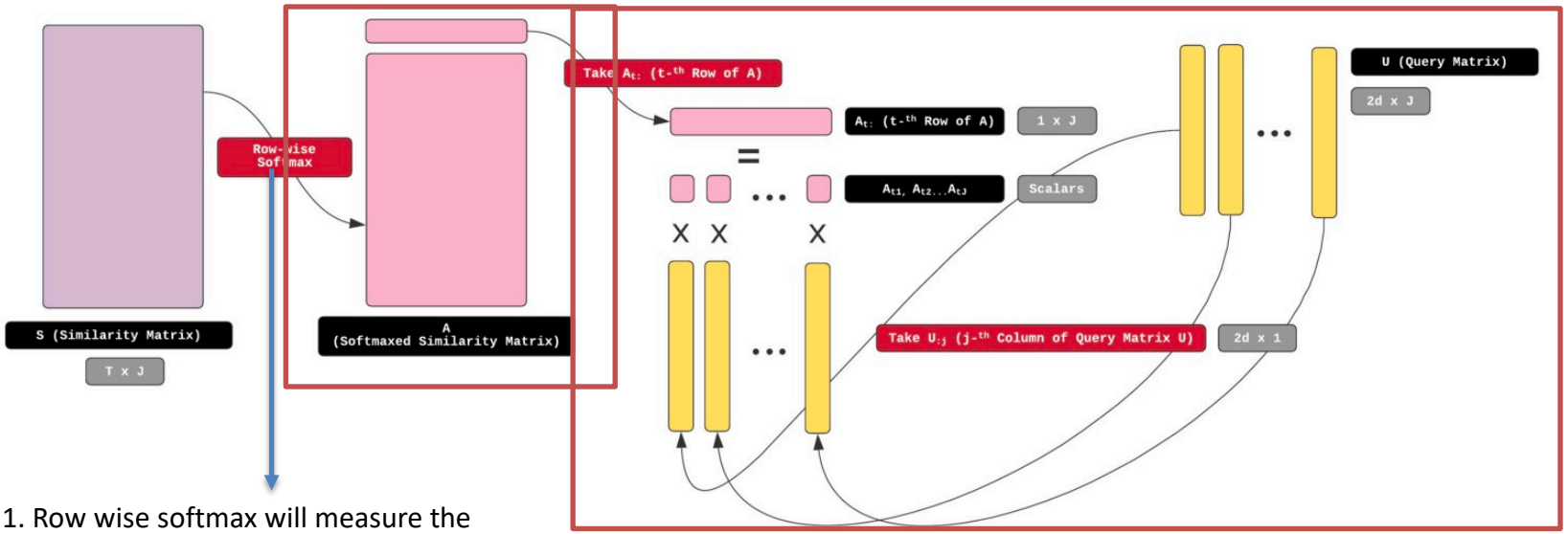
- Apply each row-wise similarity matrix (pick a row then apply softmax the repeat) → (step 1 in next slide)
- After obtaining the row-wise softmax weights, apply the weights with question vectors → (step 2 in next slide)

- Final product will be attended question vectors for all passage tokens.

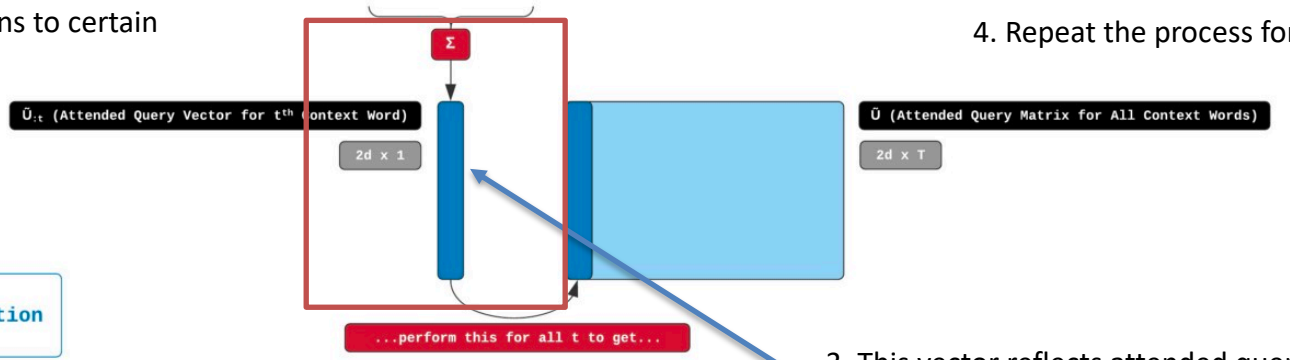
Context-to-Query Attention

** Row is context, column is question

2. Apply the weighed softmax layer to each question tokens in U (question matrix) then sum it together.



1. Row wise softmax will measure the importance of query tokens to certain context token



4. Repeat the process for all tokens

3. This vector reflects attended query vector for certain context tokens

7. Context-to-Query Attention

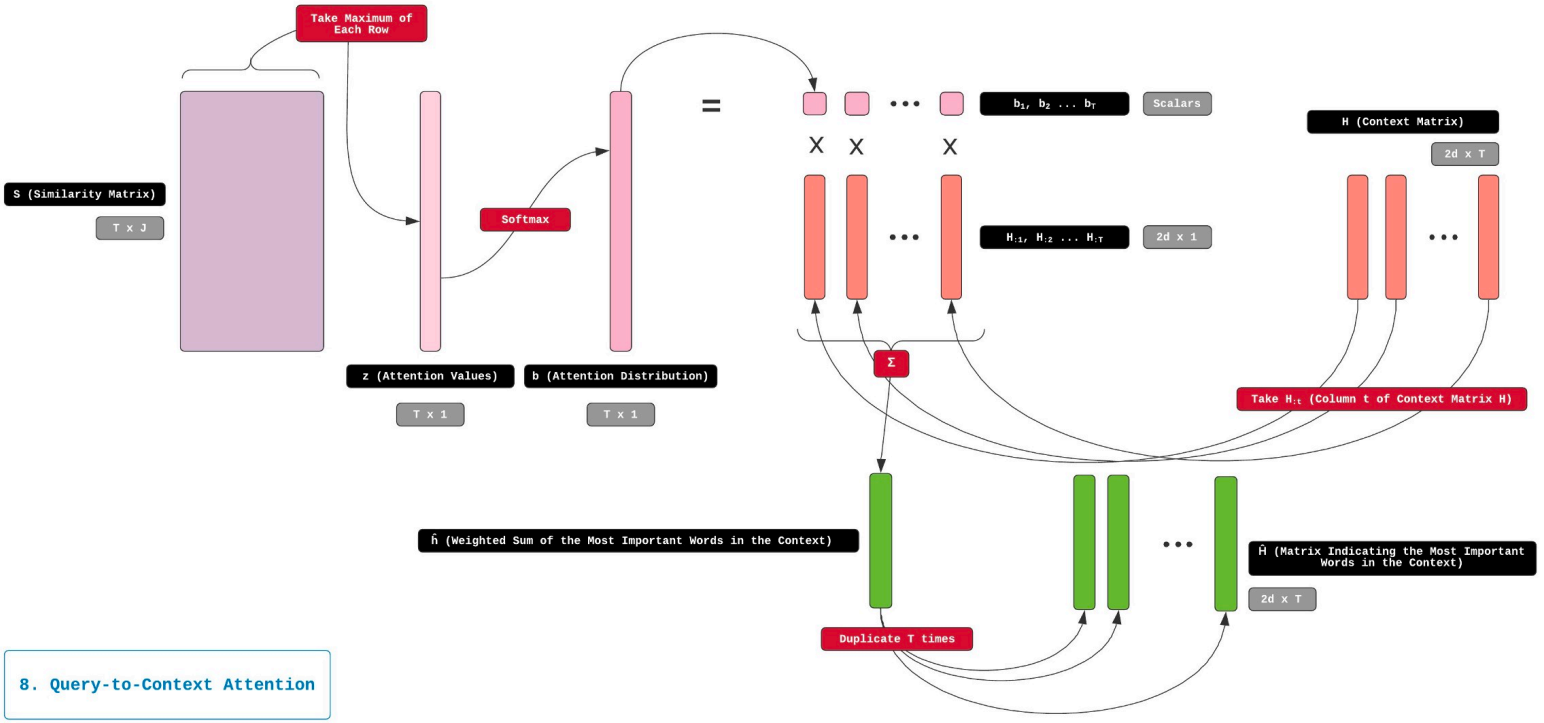
- Source: https://towardsdatascience.com/the-definitive-guide-to-bidaf-part-3-attention-92352bbdcb07

Query-to-context-attention $\widetilde{H}_{:t}$

- Reverse direction from context-to-query attention
- Take maximum of each row in similarity matrix then apply softmax
  - In context-to-query we apply softmax for each row of the similarity matrix.
- Final product will be attended context vectors (weighted sum of most important words in context tokens)

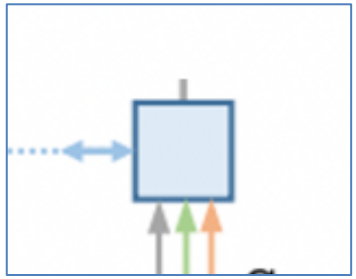
Query-to-Context Attention

- Similar to context to query attention but switch the direction
- Meaning: Question vectors evaluate which context words are the most important
- All Q2C vectors will be the same for all context tokens



- Source: https://towardsdatascience.com/the-definitive-guide-to-bidaf-part-3-attention-92352bbdcb07

Attentional Flow Layer – Output (1)

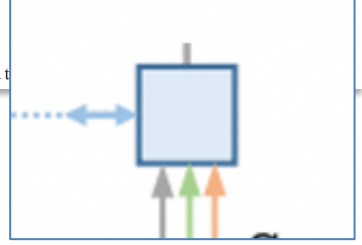


- Output
  - Batch size * length of context (T) * 8Encoding dim (8d)
  - In this layer, the output will have the length of the context passage.
  - The fuse function used is

$\beta(\mathbf{a}, \mathbf{b}, \mathbf{c}) = [\, \mathbf{a} ; \mathbf{b} ; \mathbf{a} \circ \mathbf{b} ; \mathbf{a} \circ \mathbf{c} \,]$, where...

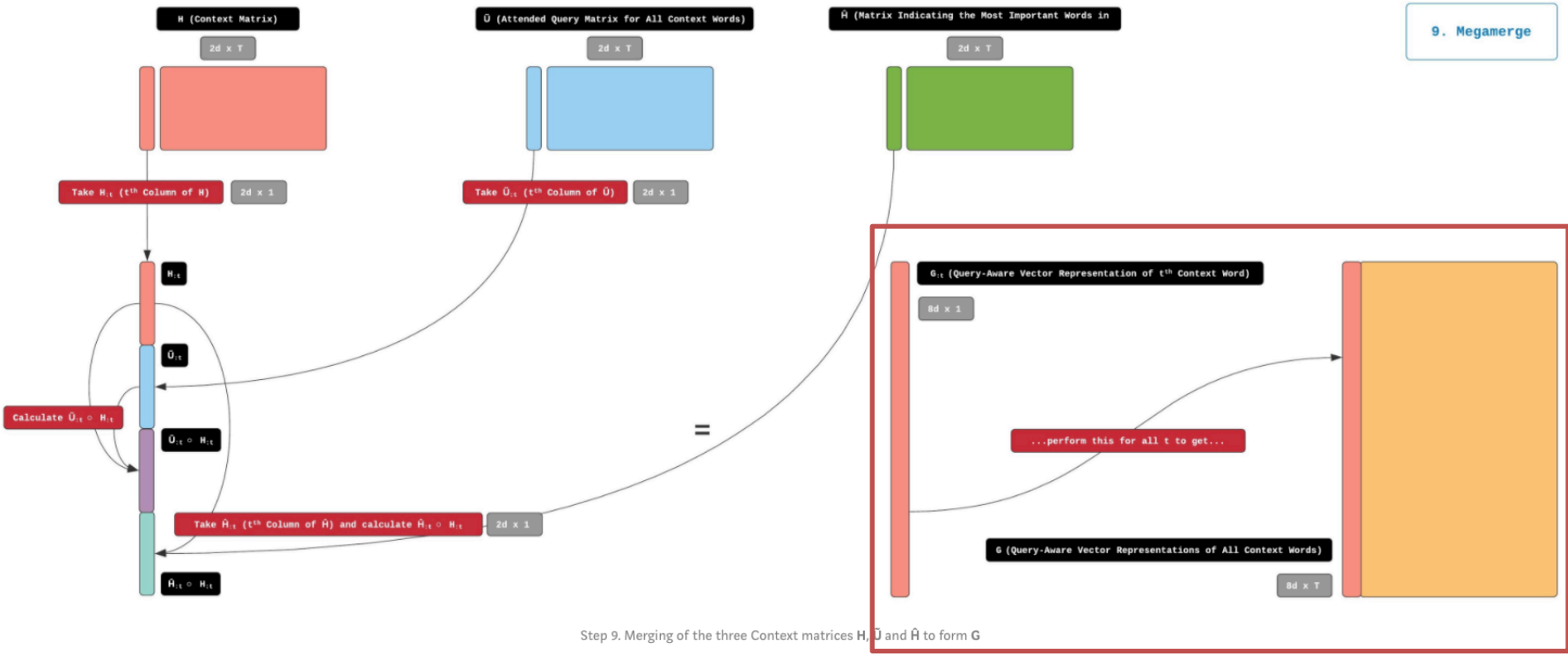
- **a, b** and **c** are the three input vectors
- $\circ$ is element wise multiplication
- $[\ ;\ ]$ is vector concatenation across row

  - a is $H_{:t}$, b is $\widetilde{U}_{:t}$, c is $\widetilde{H}_{:t}$ (each has 2d dimension, d = encoding dimension)



Attentional Flow Layer – Output (2)

- **The output** of this layer is the output vector from two attention mechanisms ($\widetilde{U}_{:t}$ and $\widetilde{H}_{:t}$) fuse with context representation vectors ($H_{:t}$)
- Output is denoted as $G_{:t}$



Step 9. Merging of the three Context matrices **H**, $\tilde{\mathbf{U}}$ and $\hat{\mathbf{H}}$ to form **G**

Tensor of [Batch size * Context token length (T) * 8Encoding Dim]

Modelling Layer

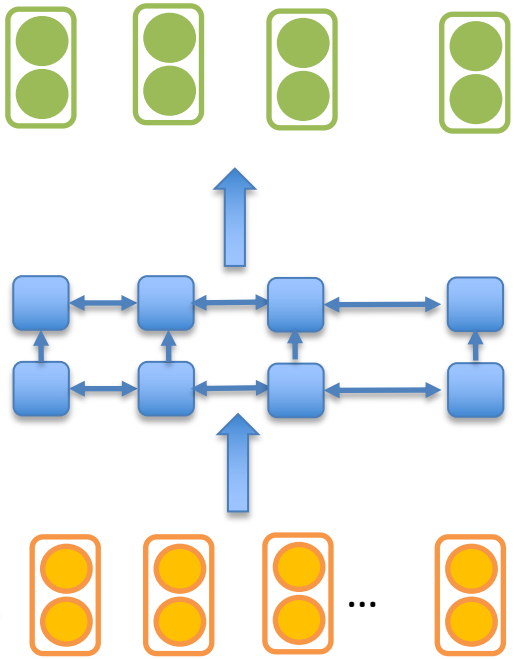
Modelling Layer $M_{:t}$

- This layer is almost identical to contextual embedding layer
  - Both are bidirectional LSTM (biLSTM)
  - In modelling layer, we utilize 2 layer of BiLSTMs

- The input to this layer can be viewed as the context representation vectors that are enhanced with bidirectional attention mechanism
  - The attention mechanism help capture information from question side
  - Context vectors that will be input to this layer now can be modelled to predict the position of token spans that could be the answer to the tokens.
- Input
  - Batch size * length of context (T) * 8Encoding dim (8d) (from previous layer)
- Output
  - Batch size * length of context (T) * 2Encoding dim (2d)

Modelling Layer (BiLSTM)

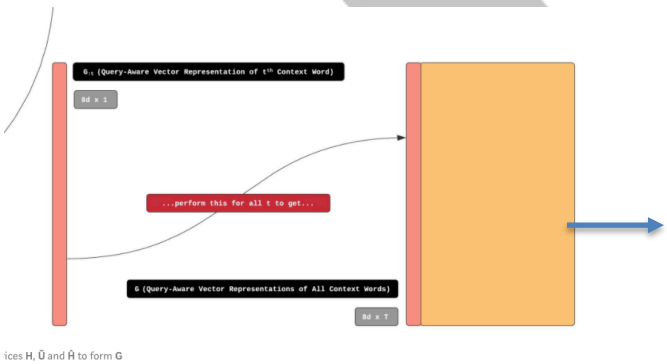
- The output of BiLSTM in is the enhanced context vectors in which each position encode information that can be used to answer the questions.



$M_{:t}$

Tensor of [Batch size * Context token length (T) * 2Encoding Dim]

2 layers BiLSTM



$G_{:t}$

Tensor of [Batch size * Context token length (T) * 8Encoding Dim]

Prediction Layer

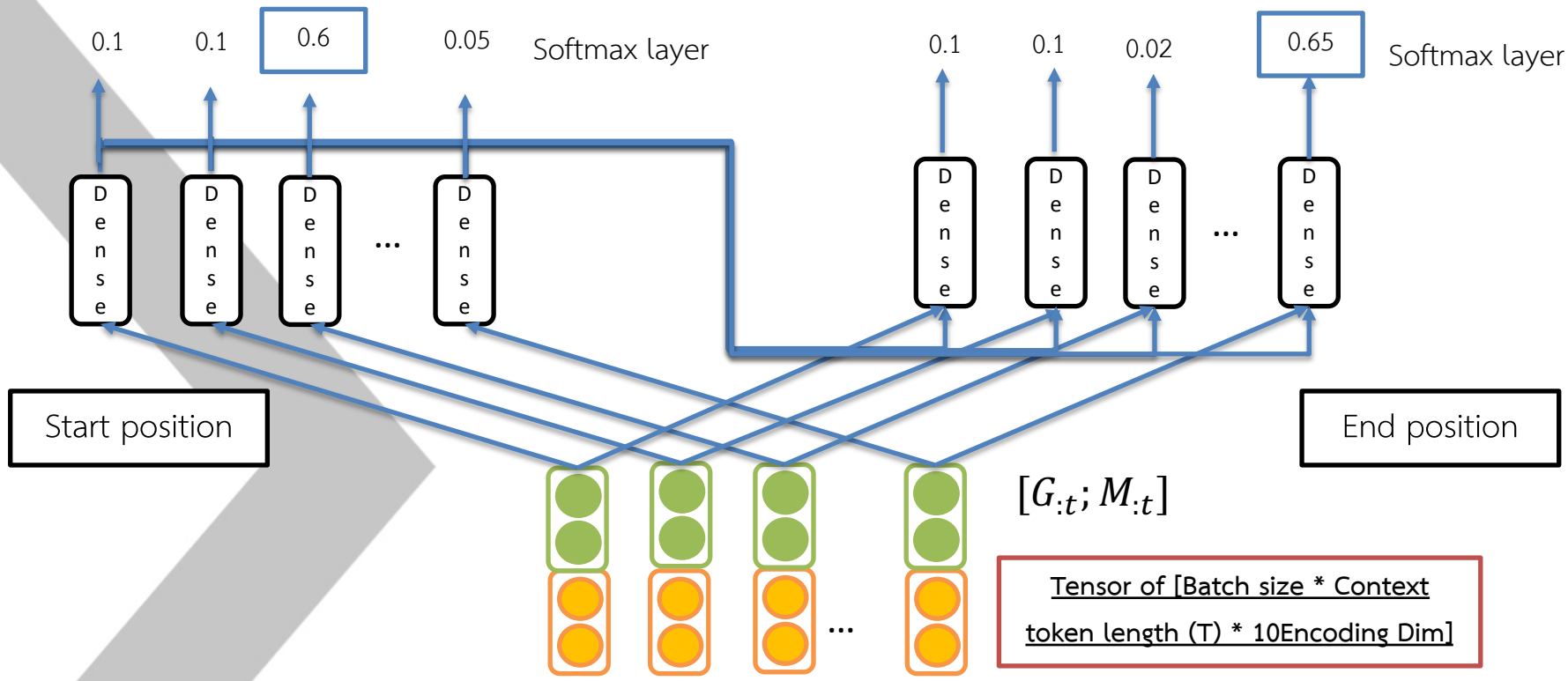
Prediction Layer

- Predict start position and end position by applying softmax to the hidden representations
- Takes input from attentional flow layer and modelling layer

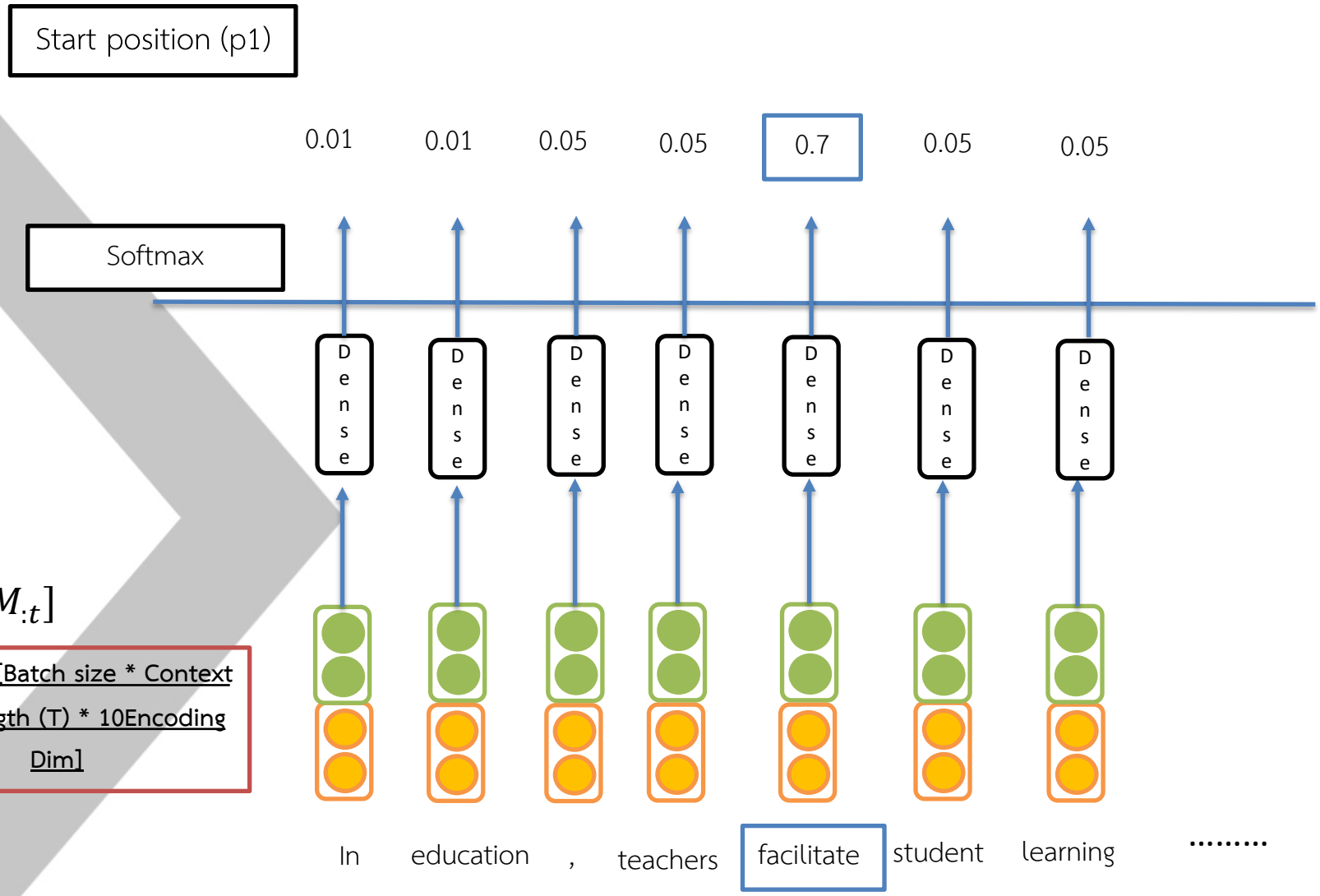
- Input
  - Batch size * length of context (T) * 10Encoding dim (10d)
- Output
  - Batch size * 2 (start and end position)

Prediction Layer

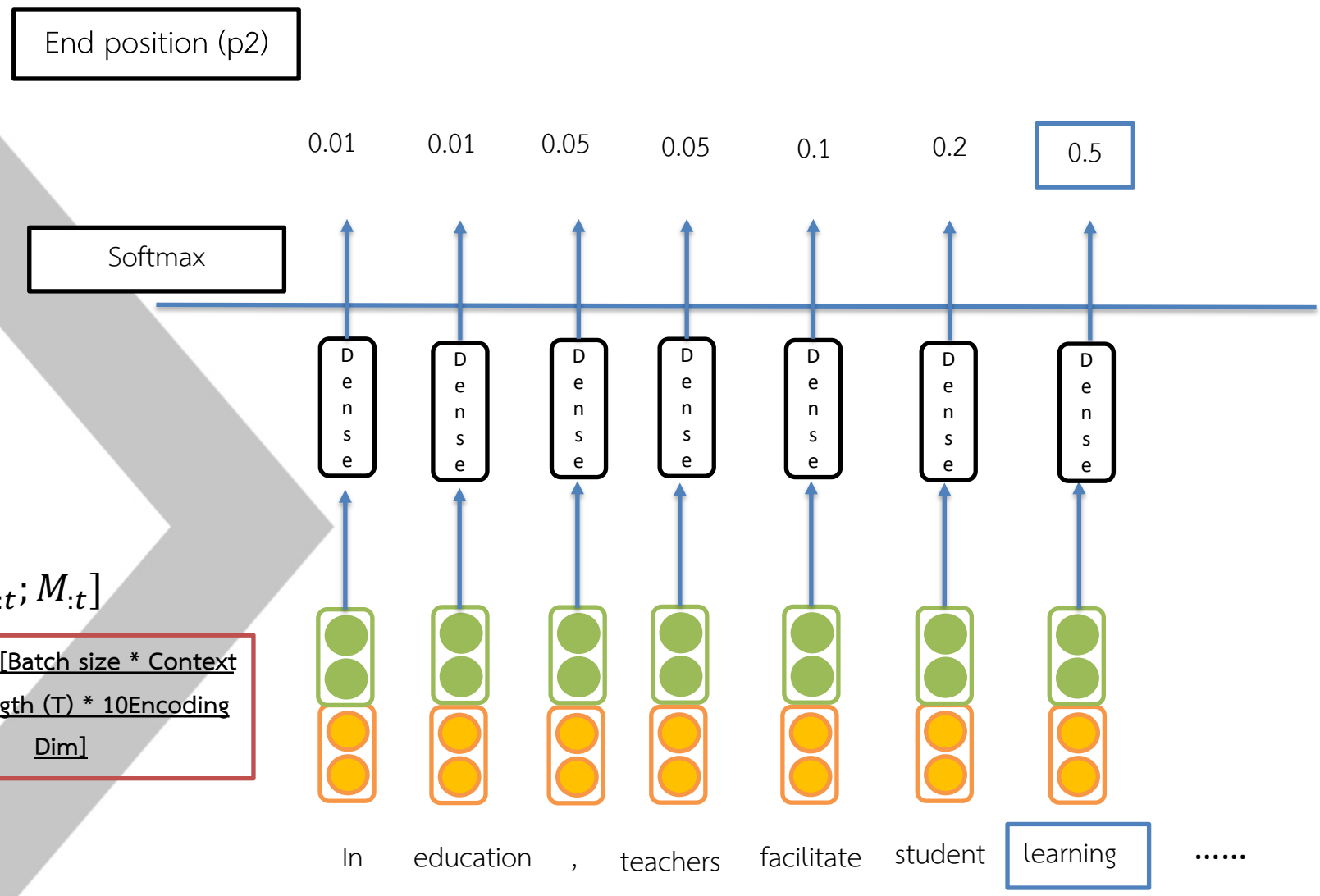
**Picks position with highest probability



Prediction Layer – Start Position Example



Prediction Layer – End Position Example



$[G_{:t}; M_{:t}]$

Tensor of [Batch size * Context token length (T) * 10Encoding Dim]

Prediction Layer – Final Answer

Start position (p1)

In education , teachers **facilitate** student learning .....

End position (p2)

In education , teachers facilitate student **learning** .....

Final Answer : facilitate student learning