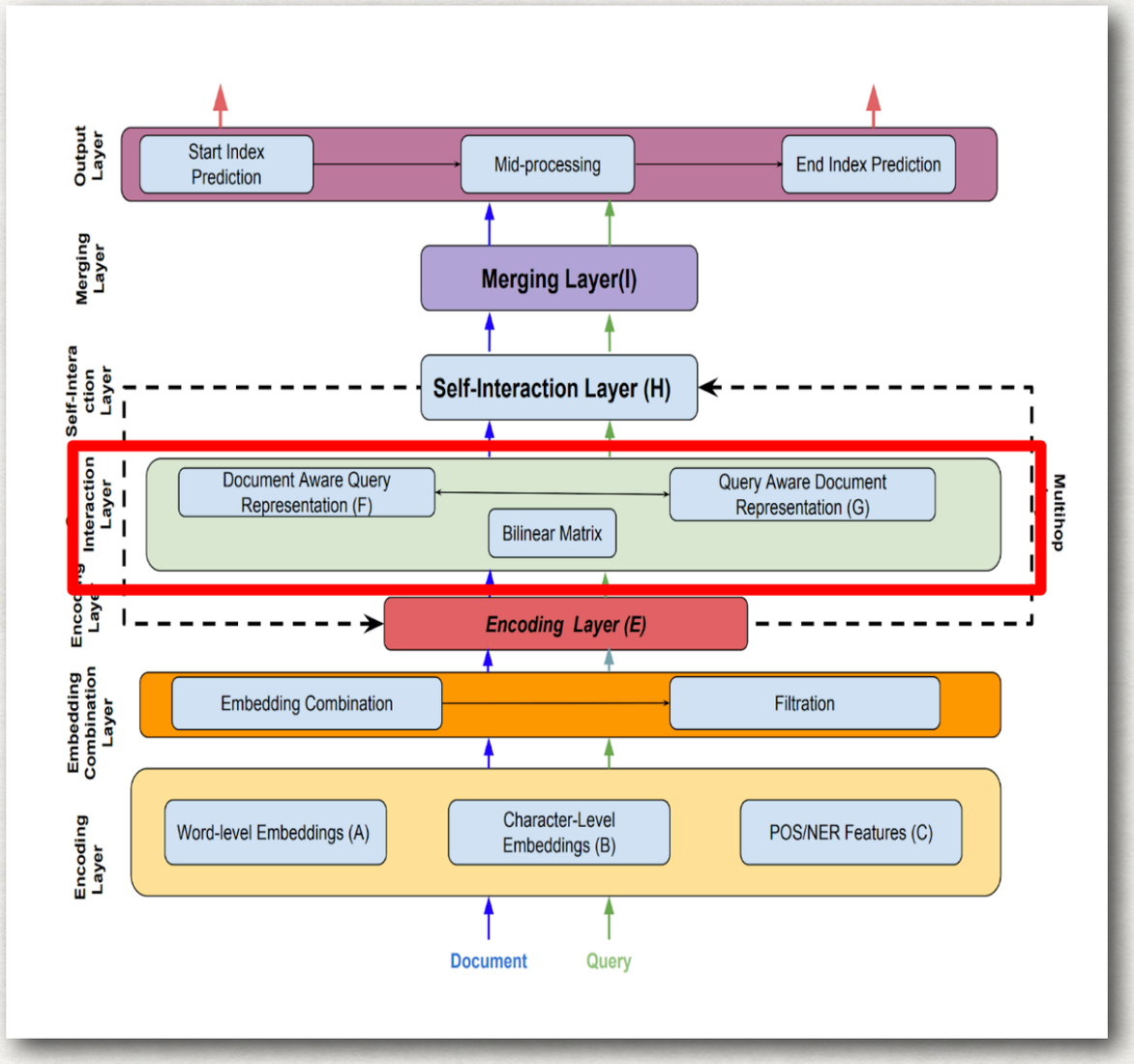
# CS7016: TOPICS IN DEEP LEARNING

# CROSS - ATTENTION

## BETWEEN DOCUMENT AND QUERY

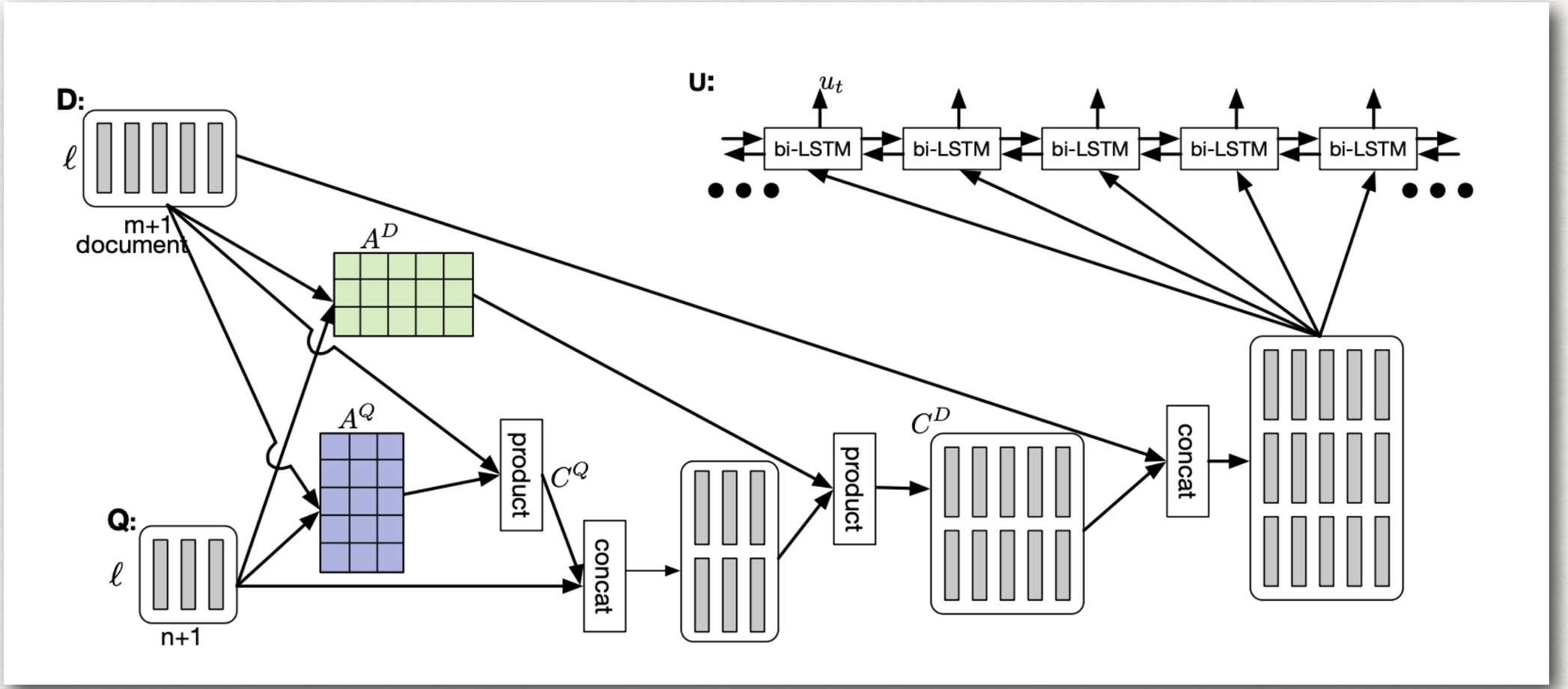
# GIVEN PARAMETERS

- $D$: Contextual document representation with $T$ (number of words = 10 for this material)
- $Q$: Contextual Query representation with $J$ ( number of query words = 6 for this material)
- Hidden size = 100



Let’s explore the cross-Interaction layer for Dynamic Co-Attention Networks

# DYNAMIC COATTENTION NETWORK




$A^D$ & $A^Q$ are dependant on Bilinear -Matrix $L$. Let's explore what $L$ is …

Document : “The boy in the red shirt went to the market”

Query : “Who went to the market ?”

$$D_{10\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| boy | 0.89 | 0.37 | ... | 0.04 | -0.84 |
| in | 0.08 | -0.22 | ... | 0.75 | -0.34 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| red | -0.30 | 0.50 | ... | 0.12 | -0.48 |
| shirt | 0.11 | 0.51 | ... | -0.21 | -0.13 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |

$$Q_{6\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| who | 0.26 | 0.44 | ... | 0.49 | -0.21 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |
| ? | 0.16 | 0.60 | ... | 0.18 | 0.36 |

# BILINEAR MATRIX L :

$$D_{10\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| boy | 0.89 | 0.37 | ... | 0.04 | -0.84 |
| in | 0.08 | -0.22 | ... | 0.75 | -0.34 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| red | -0.30 | 0.50 | ... | 0.12 | -0.48 |
| shirt | 0.11 | 0.51 | ... | -0.21 | -0.13 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |

$$Q_{6\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| who | 0.26 | 0.44 | ... | 0.49 | -0.21 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |
| ? | 0.16 | 0.60 | ... | 0.18 | 0.36 |

$$L_{10\times6} = Q^T D =$$

| Words | Who | Went | To | The | Market | ? |
|---|---|---|---|---|---|---|
| the | | | | | | |
| boy | | | | | | |
| in | | | | | | |
| the | | | | | | |
| red | | | | | | |
| shirt | | | | | | |
| went | | | | | | |
| to | | | | | | |
| the | | | | | | |
| market | | | | | | |

# BILINEAR MATRIX L :

$$D_{10\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| boy | 0.89 | 0.37 | ... | 0.04 | -0.84 |
| in | 0.08 | -0.22 | ... | 0.75 | -0.34 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| red | -0.30 | 0.50 | ... | 0.12 | -0.48 |
| shirt | 0.11 | 0.51 | ... | -0.21 | -0.13 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |

$$Q_{6\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| who | 0.26 | 0.44 | ... | 0.49 | -0.21 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |
| ? | 0.16 | 0.60 | ... | 0.18 | 0.36 |

$$L_{10\times6} = Q^T D =$$

| Words | Who | Went | To | The | Market | ? |
|---|---|---|---|---|---|---|
| the | **22.89** | | | | | |
| boy | | | | | | |
| in | | | | | | |
| the | | | | | | |
| red | | | | | | |
| shirt | | | | | | |
| went | | | | | | |
| to | | | | | | |
| the | | | | | | |
| market | | | | | | |

# BILINEAR MATRIX L:

$$D_{10\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| boy | 0.89 | 0.37 | ... | 0.04 | -0.84 |
| in | 0.08 | -0.22 | ... | 0.75 | -0.34 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| red | -0.30 | 0.50 | ... | 0.12 | -0.48 |
| shirt | 0.11 | 0.51 | ... | -0.21 | -0.13 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |

$$Q_{6\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| who | 0.26 | 0.44 | ... | 0.49 | -0.21 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |
| ? | 0.16 | 0.60 | ... | 0.18 | 0.36 |

$$L_{10\times6} = Q^T D =$$

| Words | Who | Went | To | The | Market | ? |
|---|---|---|---|---|---|---|
| the | 22.89 | **19.22** | | | | |
| boy | | | | | | |
| in | | | | | | |
| the | | | | | | |
| red | | | | | | |
| shirt | | | | | | |
| went | | | | | | |
| to | | | | | | |
| the | | | | | | |
| market | | | | | | |

# BILINEAR MATRIX L:

$$D_{10\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| boy | 0.89 | 0.37 | ... | 0.04 | -0.84 |
| in | 0.08 | -0.22 | ... | 0.75 | -0.34 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| red | -0.30 | 0.50 | ... | 0.12 | -0.48 |
| shirt | 0.11 | 0.51 | ... | -0.21 | -0.13 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |

$$Q_{6\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| who | 0.26 | 0.44 | ... | 0.49 | -0.21 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |
| ? | 0.16 | 0.60 | ... | 0.18 | 0.36 |

$$L_{10\times6} = Q^T D =$$

| Words | Who | Went | To | The | Market | ? |
|---|---|---|---|---|---|---|
| the | 22.89 | 19.22 | **28.03** | | | |
| boy | | | | | | |
| in | | | | | | |
| the | | | | | | |
| red | | | | | | |
| shirt | | | | | | |
| went | | | | | | |
| to | | | | | | |
| the | | | | | | |
| market | | | | | | |

# BILINEAR MATRIX L:

$$D_{10\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| boy | 0.89 | 0.37 | ... | 0.04 | -0.84 |
| in | 0.08 | -0.22 | ... | 0.75 | -0.34 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| red | -0.30 | 0.50 | ... | 0.12 | -0.48 |
| shirt | 0.11 | 0.51 | ... | -0.21 | -0.13 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |

$$Q_{6\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| who | 0.26 | 0.44 | ... | 0.49 | -0.21 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |
| ? | 0.16 | 0.60 | ... | 0.18 | 0.36 |

$$L_{10\times6} = Q^T D =$$

| Words | Who | Went | To | The | Market | ? |
|---|---|---|---|---|---|---|
| the | 22.89 | 19.22 | 28.03 | 33.85 | 21.81 | 18.17 |
| boy | **19.3** | | | | | |
| in | | | | | | |
| the | | | | | | |
| red | | | | | | |
| shirt | | | | | | |
| went | | | | | | |
| to | | | | | | |
| the | | | | | | |
| market | | | | | | |

# BILINEAR MATRIX L:

$$D_{10\times100} =$$

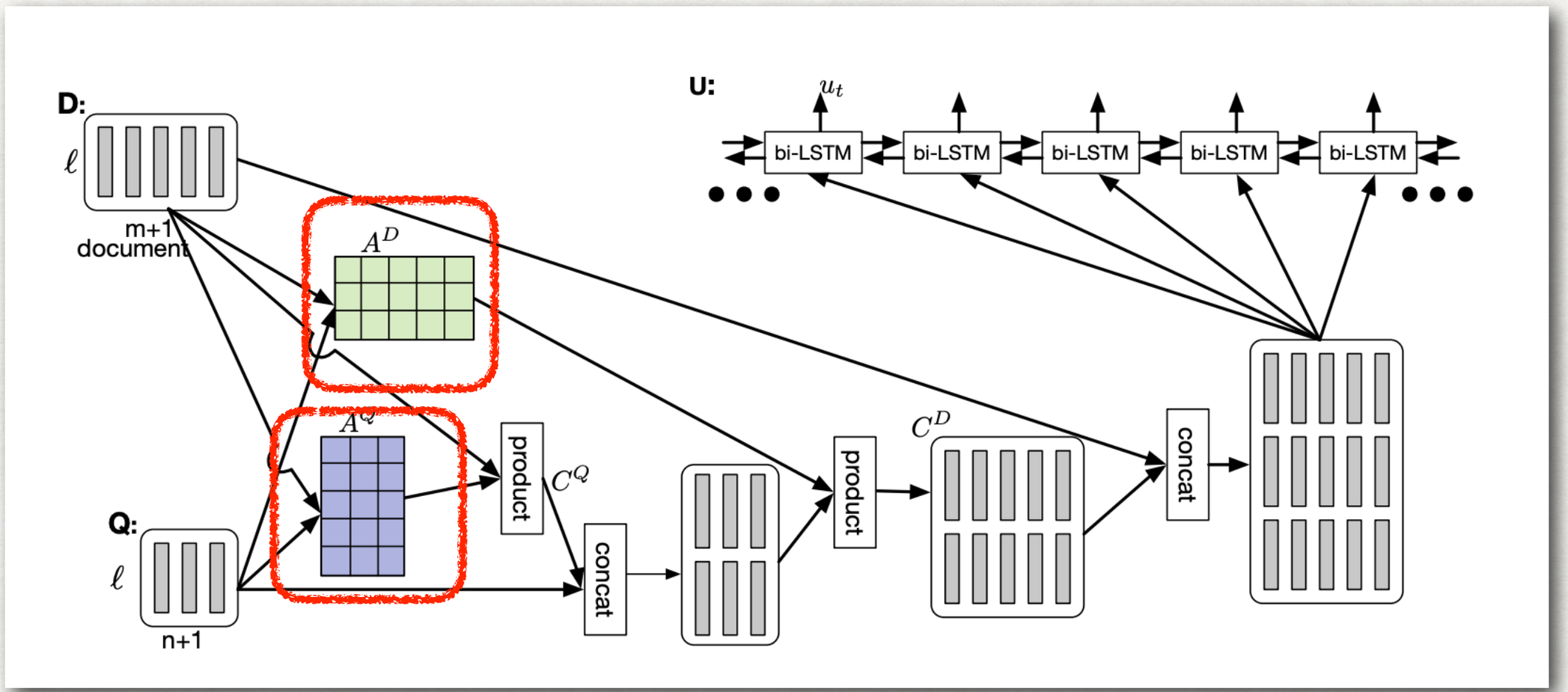
| Words | | | | | |
|---|---|---|---|---|---|
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| boy | 0.89 | 0.37 | ... | 0.04 | -0.84 |
| in | 0.08 | -0.22 | ... | 0.75 | -0.34 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| red | -0.30 | 0.50 | ... | 0.12 | -0.48 |
| shirt | 0.11 | 0.51 | ... | -0.21 | -0.13 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |

$$Q_{6\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| who | 0.26 | 0.44 | ... | 0.49 | -0.21 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |
| ? | 0.16 | 0.60 | ... | 0.18 | 0.36 |

$$L_{10\times6} = Q^T D =$$

| Words | Who | Went | To | The | Market | ? |
|---|---|---|---|---|---|---|
| the | 22.89 | 19.22 | 28.03 | 33.85 | 21.81 | 18.17 |
| boy | 19.30 | 13.98 | 14.37 | 14.08 | 9.49 | 16.19 |
| in | 23.61 | 20.62 | 27.74 | 28.07 | 23.09 | 18.99 |
| the | 22.89 | 19.22 | 28.03 | 33.88 | 21.81 | 18.17 |
| red | 16.14 | 14.41 | 16.35 | 20.77 | 12.23 | 15.07 |
| shirt | 10.04 | 9.86 | 8.45 | 8.84 | 6.48 | 12.22 |
| went | 21.68 | 25.68 | 20.95 | 19.22 | 13.77 | 16.88 |
| to | 26.33 | 20.95 | 41.61 | 28.03 | 21.69 | 20.45 |
| the | 22.89 | 19.22 | 28.03 | 33.88 | 21.81 | 18.17 |
| market | 15.87 | 13.77 | 21.69 | 21.81 | 41.34 | **16.97** |


D:
$\ell$
m+1
document
$A^D$
U:
$u_t$
bi-LSTM
bi-LSTM
bi-LSTM
bi-LSTM
bi-LSTM
$A^Q$
product
$C^Q$
concat
product
$C^D$
concat
Q:
$\ell$
n+1

# DYNAMIC COATTENTION NETWORK

$$L_{10\times6} =$$

| Words | Who | Went | To | The | Market | ? |
|---|---|---|---|---|---|---|
| the | 22.89 | 19.22 | 28.03 | 33.85 | 21.81 | 18.17 |
| boy | 19.30 | 13.98 | 14.37 | 14.08 | 9.49 | 16.19 |
| in | 23.61 | 20.62 | 27.74 | 28.07 | 23.09 | 18.99 |
| the | 22.89 | 19.22 | 28.03 | 33.88 | 21.81 | 18.17 |
| red | 16.14 | 14.41 | 16.35 | 20.77 | 12.23 | 15.07 |
| shirt | 10.04 | 9.86 | 8.45 | 8.84 | 6.48 | 12.22 |
| went | 21.68 | 25.68 | 20.95 | 19.22 | 13.77 | 16.88 |
| to | 26.33 | 20.95 | 41.61 | 28.03 | 21.69 | 20.45 |
| the | 22.89 | 19.22 | 28.03 | 33.88 | 21.81 | 18.17 |
| market | 15.87 | 13.77 | 21.69 | 21.81 | 41.34 | 16.97 |

softmax along the column:
*"which query word is more important for **each** document word"*

$$A^{D}_{10\times6} =$$

| Words | Who | Went | To | The | Markett | ? |
|---|---|---|---|---|---|---|
| the | 0.0 | 0.0 | 0.01 | 0.99 | 0.0 | 0.0 |
| boy | 0.94 | 0.004 | 0.006 | 0.005 | 0.0 | 0.045 |
| in | 0.067 | 0.0 | 0.41 | 0.57 | 0.036 | 0.0 |
| the | 0.0 | 0.0 | 0.01 | 0.99 | 0.0 | 0.0 |
| red | 0.0 | 0.0 | 0.12 | 0.97 | 0.0 | 0.028 |
| shirt | 0.09 | 0.07 | 0.01 | 0.02 | 0.002 | 0.78 |
| went | 0.0 | 0.97 | 0.008 | 0.015 | 0.0 | 0.0 |
| to | 0.0 | 0.0 | 0.99 | 0..0 | 0.0 | 0.0 |
| the | 0.0 | 0.0 | 0.0 | 0.99 | 0.0 | 0.0 |
| market | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |

# DYNAMIC COATTENTION NETWORK

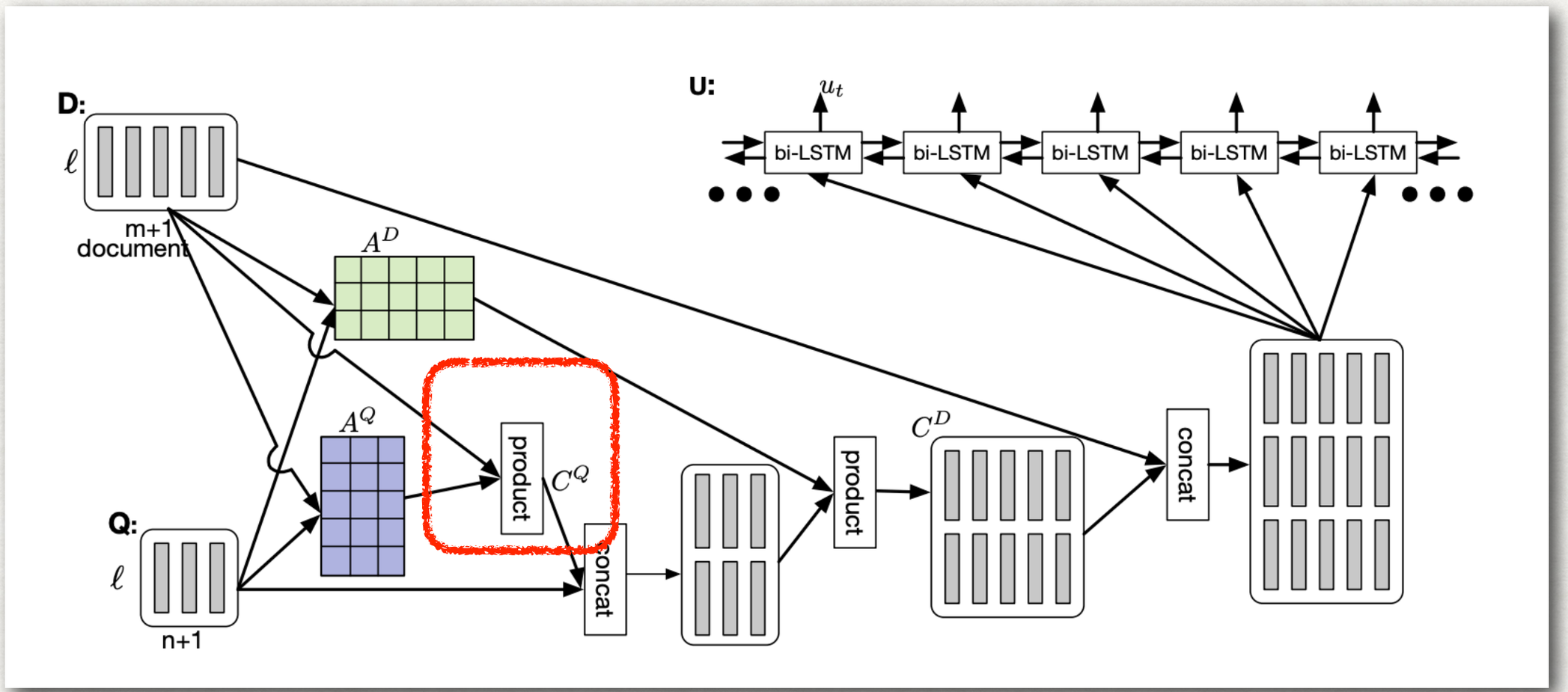
$L =$

| Words | Who | Went | To | The | Market | ? |
|---|---|---|---|---|---|---|
| the | 22.89 | 19.22 | 28.03 | 33.85 | 21.81 | 18.17 |
| boy | 19.30 | 13.98 | 14.37 | 14.08 | 9.49 | 16.19 |
| in | 23.61 | 20.62 | 27.74 | 28.07 | 23.09 | 18.99 |
| the | 22.89 | 19.22 | 28.03 | 33.88 | 21.81 | 18.17 |
| red | 16.14 | 14.41 | 16.35 | 20.77 | 12.23 | 15.07 |
| shirt | 10.04 | 9.86 | 8.45 | 8.84 | 6.48 | 12.22 |
| went | 21.68 | 25.68 | 20.95 | 19.22 | 13.77 | 16.88 |
| to | 26.33 | 20.95 | 41.61 | 28.03 | 21.69 | 20.45 |
| the | 22.89 | 19.22 | 28.03 | 33.88 | 21.81 | 18.17 |
| market | 15.87 | 13.77 | 21.69 | 21.81 | 41.34 | 16.97 |

softmax along the row:
*"which document word is more important for **each** query word"*

$A^{Q}_{10\times6} =$

| Words | Who | Went | To | The | Markett | ? |
|---|---|---|---|---|---|---|
| the | 0.028 | 0.0 | 0.0 | 0.33 | 0.0 | 0.06 |
| boy | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.008 |
| in | 0.056 | 0.0 | 0.0 | 0.0 | 0.0 | 0.14 |
| the | 0.028 | 0.015 | 0.0 | 0.33 | 0.0 | 0.06 |
| red | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.028 |
| shirt | 0.0 | 0.0 | 0.01 | 0.0 | 0.0 | 0.0 |
| went | 0.008 | 0.98 | 0.0 | 0.0 | 0.0 | 0.17 |
| to | 0.85 | 0.008 | 0.99 | 0..0 | 0.0 | 0.69 |
| the | 0.027 | 0.001 | 0.0 | 0.33 | 0.0 | 0.63 |
| market | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.19 |


D:
$\ell$
m+1
document
$A^D$
U:
$u_t$
bi-LSTM
bi-LSTM
bi-LSTM
bi-LSTM
bi-LSTM
$A^Q$
product
$C^Q$
Q:
$\ell$
n+1
concat
product
$C^D$
concat

$$C^Q = (A^Q)^T D$$

$$D_{10\times100} =$$

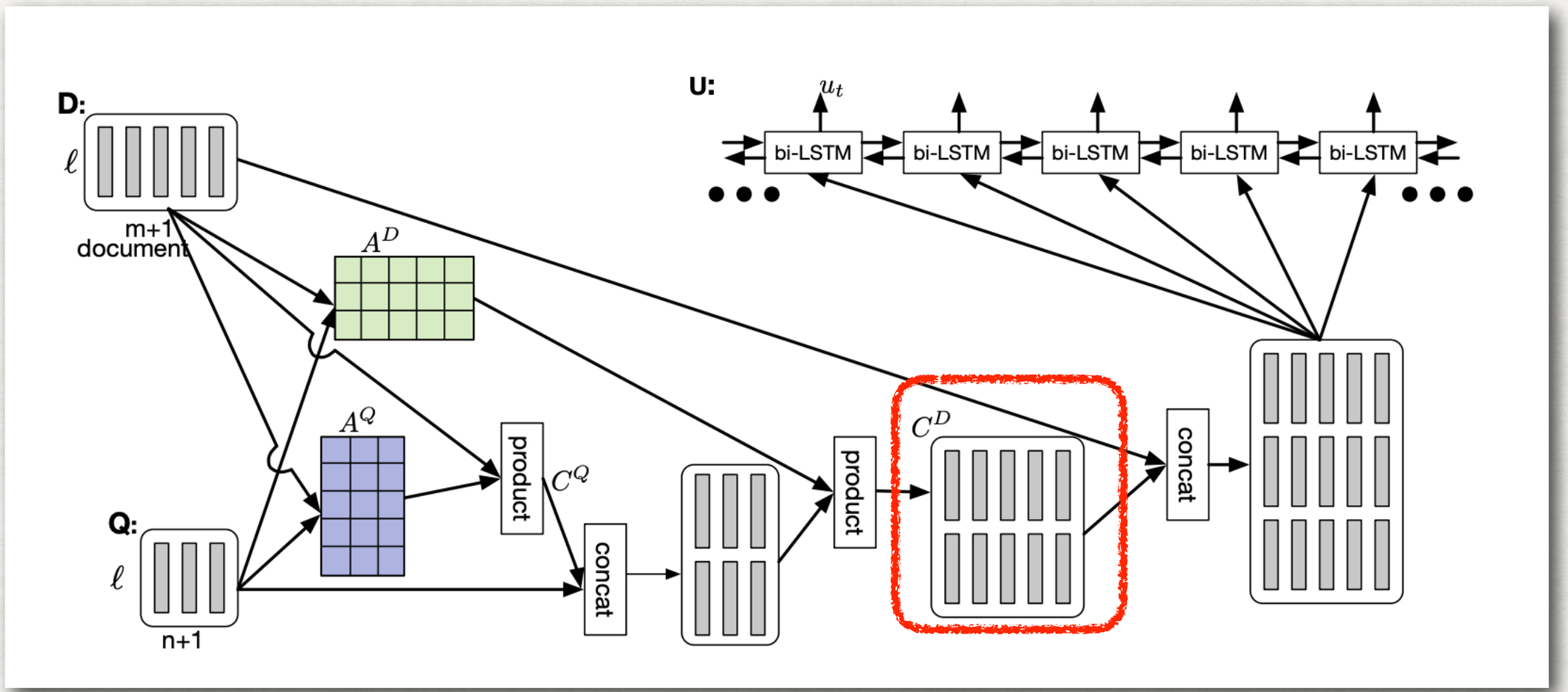
| Words | | | | | |
|---|---|---|---|---|---|
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| boy | 0.89 | 0.37 | ... | 0.04 | -0.84 |
| in | 0.08 | -0.22 | ... | 0.75 | -0.34 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| red | -0.30 | 0.50 | ... | 0.12 | -0.48 |
| shirt | 0.11 | 0.51 | ... | -0.21 | -0.13 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |

$$A^Q_{10\times6} =$$

| Words | Who | Went | To | The | Markett | ? |
|---|---|---|---|---|---|---|
| the | 0.028 | 0.0 | 0.0 | 0.33 | 0.0 | 0.06 |
| boy | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.008 |
| in | 0.056 | 0.0 | 0.0 | 0.0 | 0.0 | 0.14 |
| the | 0.028 | 0.015 | 0.0 | 0.33 | 0.0 | 0.06 |
| red | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.028 |
| shirt | 0.0 | 0.0 | 0.01 | 0.0 | 0.0 | 0.0 |
| went | 0.008 | 0.98 | 0.0 | 0.0 | 0.0 | 0.17 |
| to | 0.85 | 0.008 | 0.99 | 0..0 | 0.0 | 0.69 |
| the | 0.027 | 0.001 | 0.0 | 0.33 | 0.0 | 0.63 |
| market | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.19 |

$$C^Q_{6\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| who | -0.15 | 0.009 | ... | 0.52 | -0.13 |
| went | 0.61 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |
| ? | -0.08 | -0.04 | ... | 0.58 | -0.09 |


D:
$\ell$
m+1
document
$A^D$
U:
$u_t$
bi-LSTM
bi-LSTM
bi-LSTM
bi-LSTM
bi-LSTM
$A^Q$
product
$C^Q$
concat
Q:
$\ell$
n+1
product
$C^D$
concat

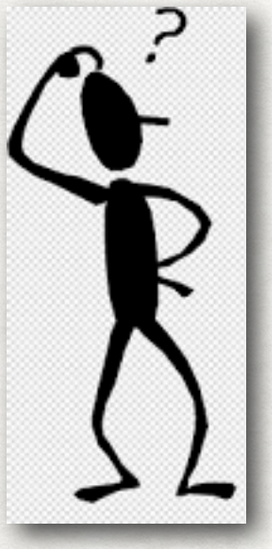
$$C^D = A^D[Q; C^Q]$$

$$C^D = [A^D Q; A^D C^Q]$$

$$A^D C^Q = [A^D (A^Q)^T] D$$

$A^D Q$ : *"Generates **10** query representations. One for each document word"*

$A^D C^Q$ : *"Complex document representation, using bilinear matrix operation on softmax matrices"*



$C^D$ : ***"is the co-representation of the question and the document"***

$$A^D Q =$$

$$Q_{6\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| who | 0.26 | 0.44 | ... | 0.49 | -0.21 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |
| ? | 0.16 | 0.60 | ... | 0.18 | 0.36 |

$$A^D_{10\times6} =$$

| Words | Who | Went | To | The | Markett | ? |
|---|---|---|---|---|---|---|
| the | 0.0 | 0.0 | 0.01 | 0.99 | 0.0 | 0.0 |
| boy | 0.94 | 0.004 | 0.006 | 0.005 | 0.0 | 0.045 |
| in | 0.067 | 0.0 | 0.41 | 0.57 | 0.036 | 0.0 |
| the | 0.0 | 0.0 | 0.01 | 0.99 | 0.0 | 0.0 |
| red | 0.0 | 0.0 | 0.12 | 0.97 | 0.0 | 0.028 |
| shirt | 0.09 | 0.07 | 0.01 | 0.02 | 0.002 | 0.78 |
| went | 0.0 | 0.97 | 0.008 | 0.015 | 0.0 | 0.0 |
| to | 0.0 | 0.0 | 0.99 | 0..0 | 0.0 | 0.0 |
| the | 0.0 | 0.0 | 0.0 | 0.99 | 0.0 | 0.0 |
| market | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |

$$A^D Q_{10\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| boy | 0.25 | 0.44 | ... | 0.48 | -0.17 |
| in | -0.09 | -0.16 | ... | 0.67 | 0.09 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| red | -0.30 | -0.23 | ... | 0.81 | 0.26 |
| shirt | 0.19 | 0.50 | ... | 0.24 | 0.25 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |

"It computes 10 ($T$) weighted-document representations, in query space"

$$A^D C^Q =$$

$$C^Q_{6\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| who | -0.15 | 0.009 | ... | 0.52 | -0.13 |
| went | 0.61 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |
| ? | -0.08 | -0.04 | ... | 0.58 | -0.09 |

$$A^D_{10\times6} =$$

| Words | Who | Went | To | The | Markett | ? |
|---|---|---|---|---|---|---|
| the | 0.0 | 0.0 | 0.01 | 0.99 | 0.0 | 0.0 |
| boy | 0.94 | 0.004 | 0.006 | 0.005 | 0.0 | 0.045 |
| in | 0.067 | 0.0 | 0.41 | 0.57 | 0.036 | 0.0 |
| the | 0.0 | 0.0 | 0.01 | 0.99 | 0.0 | 0.0 |
| red | 0.0 | 0.0 | 0.12 | 0.97 | 0.0 | 0.028 |
| shirt | 0.09 | 0.07 | 0.01 | 0.02 | 0.002 | 0.78 |
| went | 0.0 | 0.97 | 0.008 | 0.015 | 0.0 | 0.0 |
| to | 0.0 | 0.0 | 0.99 | 0..0 | 0.0 | 0.0 |
| the | 0.0 | 0.0 | 0.0 | 0.99 | 0.0 | 0.0 |
| market | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |

$$A^D C^Q_{10\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| boy | -0.14 | 0.05 | ... | 0.52 | -0.13 |
| in | -0.09 | -0.12 | ... | 0.67 | 0.09 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| red | -0.04 | -0.23 | ... | 0.81 | 0.26 |
| shirt | -0.04 | -0.04 | ... | 0.55 | -0.09 |
| went | 0.58 | -0.12 | ... | 0.26 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |

$$C^D = [A^D Q; A^D C^Q]$$

$$C^D_{10\times200} =$$

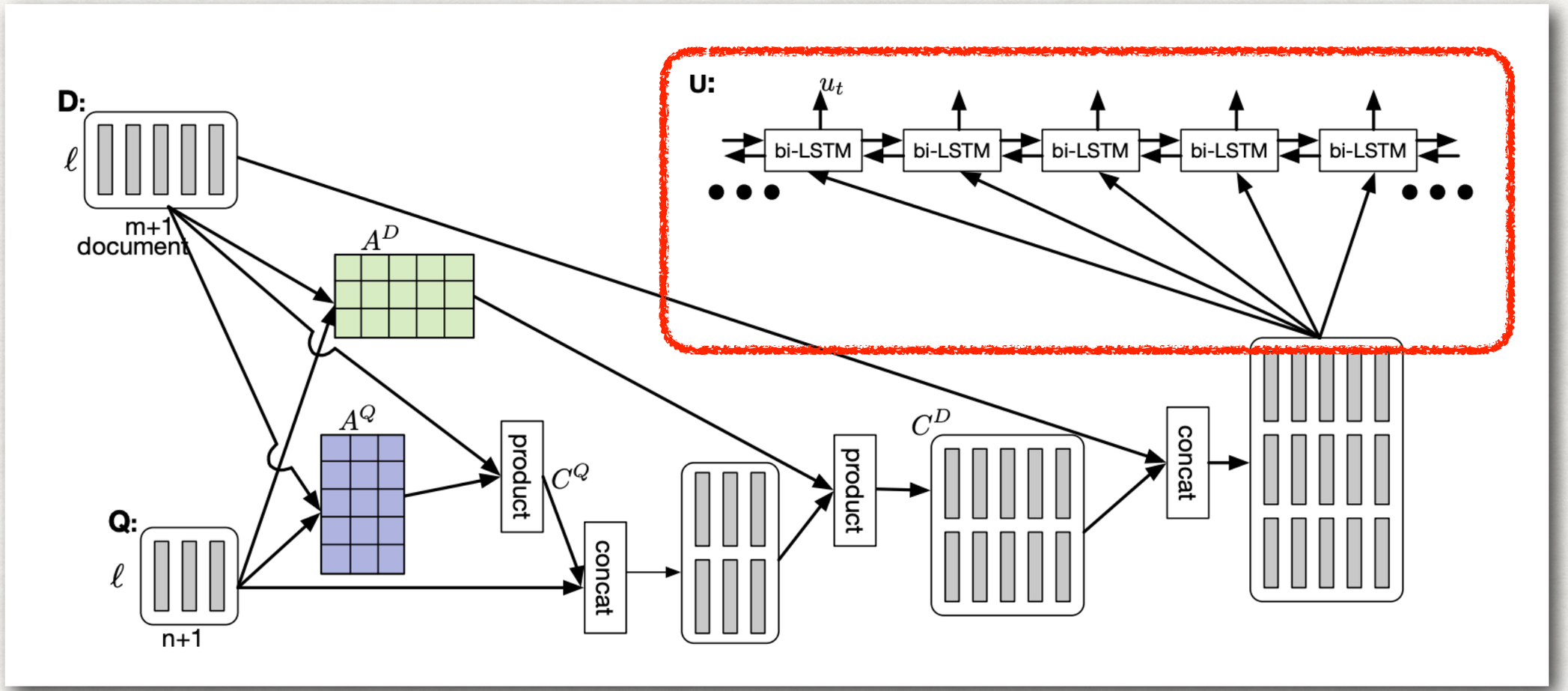
| Words | | | | | |
|---|---|---|---|---|---|
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| boy | 0.25 | 0.44 | ... | 0.48 | -0.17 |
| in | -0.09 | -0.16 | ... | 0.67 | 0.09 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| red | -0.30 | -0.23 | ... | 0.81 | 0.26 |
| shirt | 0.19 | 0.50 | ... | 0.24 | 0.25 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |

| Words | | | | | |
|---|---|---|---|---|---|
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| boy | -0.14 | 0.05 | ... | 0.52 | -0.13 |
| in | -0.09 | -0.12 | ... | 0.67 | 0.09 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| red | -0.04 | -0.23 | ... | 0.81 | 0.26 |
| shirt | -0.04 | -0.04 | ... | 0.55 | -0.09 |
| went | 0.58 | -0.12 | ... | 0.26 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |

$$A^D Q_{10\times100}$$

$$A^D C^Q_{10\times100}$$

obviously!


D:
$\ell$
m+1
document
$A^D$
U:
$u_t$
bi-LSTM
bi-LSTM
bi-LSTM
bi-LSTM
bi-LSTM
$A^Q$
product
$C^Q$
concat
product
$C^D$
concat
Q:
$\ell$
n+1

# BIDAF : CROSS-INTERACTION LAYER

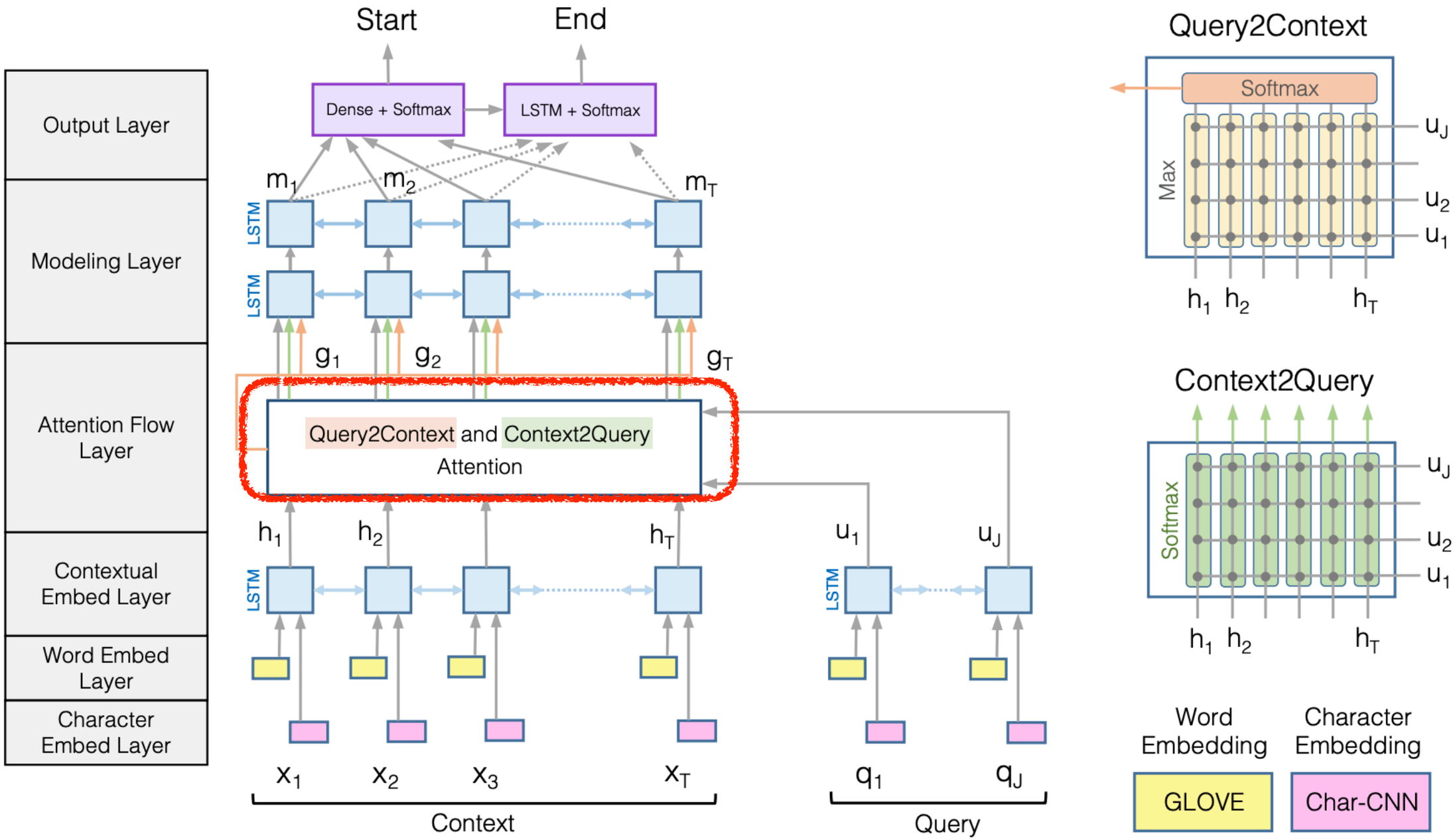



Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

- Similar to bilinear matrix, BiDAF also computes word-word interaction between document and query words
- However the formulation is slightly different:

$$L = f(D, Q)$$

Bilinear Matrix :

$$L_{ij} = d_i \cdot q_j$$

BiDAF :

$$L_{ij} = w^T[d_i; q_j; d_i \odot q_j]$$

- It gives the model more freedom, it can decide how much of query and document can be taken to capture the interaction.
- $w^T$ is a learnable parameter.

# BILINEAR MATRIX L :

$$Q_{6\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| who | 0.26 | 0.44 | ... | 0.49 | -0.21 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |
| ? | 0.16 | 0.60 | ... | 0.18 | 0.36 |

$$w = tf.ones(6 \times 100)$$

$$T_{11} = [d_1; q_1; d_1 \odot q_1] =$$

| the | -0.03 | -0.24 | ... | 0.83 | 0.27 | who | 0.26 | 0.44 | ... | 0.49 | -0.21 | the-who | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

$$L_{11} = w^T T_{11} = 17.39$$

$$D_{10\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| boy | 0.89 | 0.37 | ... | 0.04 | -0.84 |
| in | 0.08 | -0.22 | ... | 0.75 | -0.34 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| red | -0.30 | 0.50 | ... | 0.12 | -0.48 |
| shirt | 0.11 | 0.51 | ... | -0.21 | -0.13 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |

$$L_{10\times6} =$$

| Words | Who | Went | To | The | Market | ? |
|---|---|---|---|---|---|---|
| the | **17.39** | | | | | |
| boy | | | | | | |
| in | | | | | | |
| the | | | | | | |
| red | | | | | | |
| shirt | | | | | | |
| went | | | | | | |
| to | | | | | | |
| the | | | | | | |
| market | | | | | | |

# BILINEAR MATRIX L :

$$Q_{6\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| who | 0.26 | 0.44 | ... | 0.49 | -0.21 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |
| ? | 0.16 | 0.60 | ... | 0.18 | 0.36 |

$$w = tf.ones(6 \times 100)$$

$$T_{12} = [d_1; q_2; d_1 \odot q_2] =$$

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 | Went | 0.26 | 0.44 | ... | 0.49 | -0.21 | The-went | | | ... |

$$L_{12} = w^T T_{12} = 15.32$$

$$D_{10\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| boy | 0.89 | 0.37 | ... | 0.04 | -0.84 |
| in | 0.08 | -0.22 | ... | 0.75 | -0.34 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| red | -0.30 | 0.50 | ... | 0.12 | -0.48 |
| shirt | 0.11 | 0.51 | ... | -0.21 | -0.13 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |

$$L_{10\times6} =$$

| Words | Who | Went | To | The | Market | ? |
|---|---|---|---|---|---|---|
| the | 17.39 | **15.32** | | | | |
| boy | | | | | | |
| in | | | | | | |
| the | | | | | | |
| red | | | | | | |
| shirt | | | | | | |
| went | | | | | | |
| to | | | | | | |
| the | | | | | | |
| market | | | | | | |

# BILINEAR MATRIX L :

$$Q_{6\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| who | 0.26 | 0.44 | ... | 0.49 | -0.21 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |
| ? | 0.16 | 0.60 | ... | 0.18 | 0.36 |

$$w = tf.ones(6 \times 100)$$

$$T_{13} = [d_1; q_3; d_1 \odot q_3] =$$

| the | -0.03 | -0.24 | ... | 0.83 | 0.27 | To | 0.26 | 0.44 | ... | 0.49 | -0.21 | The-to | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

$$L_{13} = w^T T_{13} = 17.26$$

$$D_{10\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| boy | 0.89 | 0.37 | ... | 0.04 | -0.84 |
| in | 0.08 | -0.22 | ... | 0.75 | -0.34 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| red | -0.30 | 0.50 | ... | 0.12 | -0.48 |
| shirt | 0.11 | 0.51 | ... | -0.21 | -0.13 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |

$$L_{10\times6} =$$

| Words | Who | Went | To | The | Market | ? |
|---|---|---|---|---|---|---|
| the | 17.39 | 15.32 | **17.26** | | | |
| boy | | | | | | |
| in | | | | | | |
| the | | | | | | |
| red | | | | | | |
| shirt | | | | | | |
| went | | | | | | |
| to | | | | | | |
| the | | | | | | |
| market | | | | | | |

# BILINEAR MATRIX L :

$$Q_{6\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| who | 0.26 | 0.44 | ... | 0.49 | -0.21 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |
| ? | 0.16 | 0.60 | ... | 0.18 | 0.36 |

$$w = tf.ones(6 \times 100)$$

$$T_{106} = [d_{10}; q_6; d_{10} \odot q_6] =$$

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 | ? | 0.26 | 0.44 | ... | 0.49 | -0.21 | the-? | | | ... |

$$L_{106} = w^T T_{106} = 5.40$$

$$D_{10\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| boy | 0.89 | 0.37 | ... | 0.04 | -0.84 |
| in | 0.08 | -0.22 | ... | 0.75 | -0.34 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| red | -0.30 | 0.50 | ... | 0.12 | -0.48 |
| shirt | 0.11 | 0.51 | ... | -0.21 | -0.13 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |

$$L_{10\times6} =$$

| Words | Who | Went | To | The | Market | ? |
|---|---|---|---|---|---|---|
| the | 17.39 | 15.32 | 17.26 | 28.29 | 10.70 | 12.11 |
| boy | 22.21 | 18.49 | 12.101 | 16.90 | 6.80 | 18.55 |
| in | 20.78 | 19.38 | 19.64 | 25.14 | 14.56 | 15.60 |
| the | 17.39 | 15.32 | 17.26 | 28.29 | 10.70 | 12.11 |
| red | 14.12 | 13.98 | 9.05 | 18.65 | 4.60 | 12.49 |
| shirt | 8.85 | 10.26 | 1.99 | 7.55 | -0.32 | 10.46 |
| went | 17.87 | 23.46 | 11.86 | 15.32 | 4.34 | 12.50 |
| to | 15.65 | 11.86 | 25.66 | 17.26 | 5.40 | 9.22 |
| the | 17.39 | 15.32 | 17.26 | 28.29 | 10.70 | 12.11 |
| market | 4.86 | 4.34 | 5.40 | 10.70 | 24.71 | **5.40** |

# CONTEXT-TO-QUERY ATTENTION

$$L_{10\times6} =$$

| Words | Who | Went | To | The | Marke | ? |
|---|---|---|---|---|---|---|
| the | 17.39 | 15.32 | 17.26 | 28.29 | 10.70 | 12.11 |
| boy | 22.21 | 18.49 | 12.101 | 16.90 | 6.80 | 18.55 |
| in | 20.78 | 19.38 | 19.64 | 25.14 | 14.56 | 15.60 |
| the | 17.39 | 15.32 | 17.26 | 28.29 | 10.70 | 12.11 |
| red | 14.12 | 13.98 | 9.05 | 18.65 | 4.60 | 12.49 |
| shirt | 8.85 | 10.26 | 1.99 | 7.55 | -0.32 | 10.46 |
| went | 17.87 | 23.46 | 11.86 | 15.32 | 4.34 | 12.50 |
| to | 15.65 | 11.86 | 25.66 | 17.26 | 5.40 | 9.22 |
| the | 17.39 | 15.32 | 17.26 | 28.29 | 10.70 | 12.11 |
| market | 4.86 | 4.34 | 5.40 | 10.70 | 24.71 | 5.40 |

softmax along the column:
*"which query word is more important for **each** document word"*

$$A^{D}_{10\times6} =$$

| Words | Who | Went | To | The | Markett | ? |
|---|---|---|---|---|---|---|
| the | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| boy | 0.94 | 0.02 | 0.0 | 0.004 | 0.0 | 0.02 |
| in | 0.01 | 0.003 | 0.003 | 0.98 | 0.0 | 0.0 |
| the | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| red | 0.01 | 0.009 | 0.0 | 0.97 | 0.0 | 0.02 |
| shirt | 0.096 | 0.39 | 0.0 | 0.02 | 0.0 | 0.48 |
| went | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| to | 0.0 | 0.0 | 1.0 | 0..0 | 0.0 | 0.0 |
| the | 0.0 | 0.0 | 0.0 | 0.99 | 0.0 | 0.0 |
| market | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |

# CONTEXT-TO-QUERY ATTENTION

$$A^D Q =$$

$$Q_{6\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| who | 0.26 | 0.44 | ... | 0.49 | -0.21 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |
| ? | 0.16 | 0.60 | ... | 0.18 | 0.36 |

$$A^D_{10\times6} =$$

| Words | Who | Went | To | The | Marke | ? |
|---|---|---|---|---|---|---|
| the | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| boy | 0.94 | 0.02 | 0.0 | 0.004 | 0.0 | 0.02 |
| in | 0.01 | 0.003 | 0.003 | 0.98 | 0.0 | 0.0 |
| the | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| red | 0.01 | 0.009 | 0.0 | 0.97 | 0.0 | 0.02 |
| shirt | 0.096 | 0.39 | 0.0 | 0.02 | 0.0 | 0.48 |
| went | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| to | 0.0 | 0.0 | 1.0 | 0..0 | 0.0 | 0.0 |
| the | 0.0 | 0.0 | 0.0 | 0.99 | 0.0 | 0.0 |
| market | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |

$$(A^D Q)_{10\times100} =$$

| Words | | | | | |
|---|---|---|---|---|---|
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| boy | 0.25 | 0.44 | ... | 0.48 | -0.17 |
| in | -0.03 | -0.23 | ... | 0.82 | 0.26 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| red | -0.02 | -0.23 | ... | 0.81 | 0.26 |
| shirt | 0.34 | 0.27 | ... | 0.25 | 0.52 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |

"It computes document representations, in query space"

***Seems familiar ?***

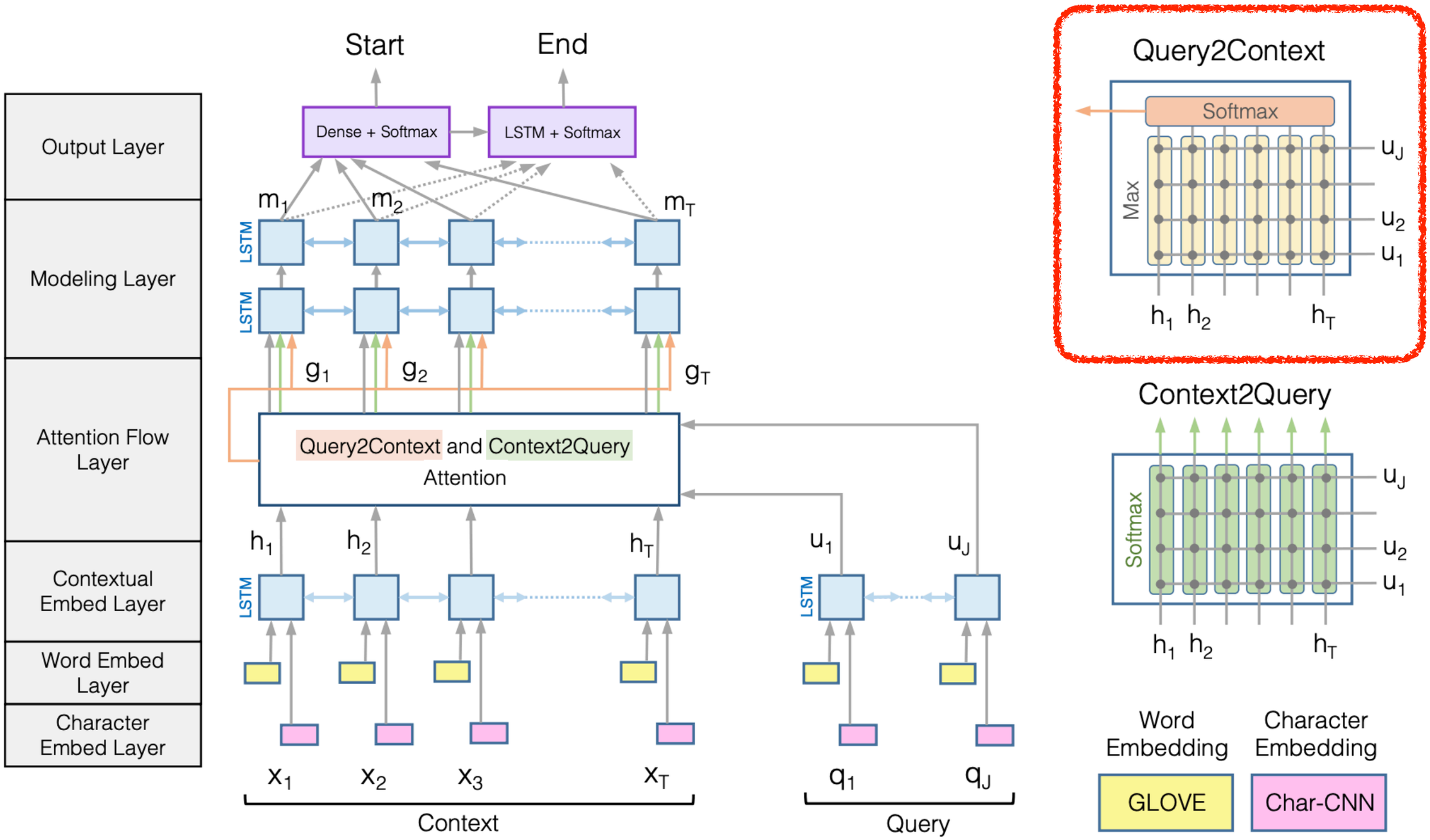



Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

# QUERY-TO-CONTEXT REPRESENTATION

- Slightly different than the co-attention matrix $C^D$ discussed in DCNs.
- **Highlights the context words having closest similarity to any of the query words.**

Closest Similarity to
Any of the query words

$$M_{10\times1} = max_{col}(L)$$

$$L_{10\times6} =$$

| Words | Who | Went | To | The | Marke | ? |
|---|---|---|---|---|---|---|
| the | 17.39 | 15.32 | 17.26 | 28.29 | 10.70 | 12.11 |
| boy | 22.21 | 18.49 | 12.101 | 16.90 | 6.80 | 18.55 |
| in | 20.78 | 19.38 | 19.64 | 25.14 | 14.56 | 15.60 |
| the | 17.39 | 15.32 | 17.26 | 28.29 | 10.70 | 12.11 |
| red | 14.12 | 13.98 | 9.05 | 18.65 | 4.60 | 12.49 |
| shirt | 8.85 | 10.26 | 1.99 | 7.55 | -0.32 | 10.46 |
| went | 17.87 | 23.46 | 11.86 | 15.32 | 4.34 | 12.50 |
| to | 15.65 | 11.86 | 25.66 | 17.26 | 5.40 | 9.22 |
| the | 17.39 | 15.32 | 17.26 | 28.29 | 10.70 | 12.11 |
| market | 4.86 | 4.34 | 5.40 | 10.70 | 24.71 | 5.40 |

$$M_{10\times1} =$$

| Words | |
|---|---|
| the | 28.29 |
| boy | 22.21 |
| in | 25.14 |
| the | 28.29 |
| red | 18.65 |
| shirt | 10.46 |
| went | 23.46 |
| to | 25.66 |
| the | 28.29 |
| marke | 24.71 |

$$\text{softmax}(M_{10\times1}) =$$

| Words | Who |
|---|---|
| the | 0.31 |
| boy | 0.0 |
| in | 0.01 |
| the | 0.31 |
| red | 0.0 |
| shirt | 0.002 |
| went | 0.02 |
| to | 0.0 |
| the | 0.31 |
| marke | 0.0 |

# QUERY-TO-CONTEXT REPRESENTATION

**Query Aware Document Representation** $\tilde{D} = \text{softmax}(M_{10\times1})^T D$

$D_{10\times100} =$

| Words | | | | | |
|---|---|---|---|---|---|
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| boy | 0.89 | 0.37 | ... | 0.04 | -0.84 |
| in | 0.08 | -0.22 | ... | 0.75 | -0.34 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| red | -0.30 | 0.50 | ... | 0.12 | -0.48 |
| shirt | 0.11 | 0.51 | ... | -0.21 | -0.13 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |

$\tilde{D} =$

| Words | -0.03 | -0.24 | ... | 0.81 | 0.25 |
|---|---|---|---|---|---|

$\text{softmax}(M_{10\times1}) =$

| Words | |
|---|---|
| the | 0.31 |
| boy | 0.0 |
| in | 0.01 |
| the | 0.31 |
| red | 0.0 |
| shirt | 0.002 |
| went | 0.02 |
| to | 0.0 |
| the | 0.31 |
| marke | 0.0 |

# QUERY-TO-CONTEXT REPRESENTATION

## How to merge context-to-query and query-to-context representations ?

$$C_i^D = [d_i; C_i^Q; d_i \odot C_i^Q; d_i \odot \tilde{D}] \quad \forall i \in \{1,10\}$$

**DOCUMENT REPRESENTATION**

$\mathbf{D} =$

| Words | | | | | |
|---|---|---|---|---|---|
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| boy | 0.89 | 0.37 | ... | 0.04 | -0.84 |
| in | 0.08 | -0.22 | ... | 0.75 | -0.34 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| red | -0.30 | 0.50 | ... | 0.12 | -0.48 |
| shirt | 0.11 | 0.51 | ... | -0.21 | -0.13 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |

**QUERY REPRESENTATION**

$\mathbf{Q} =$

| Words | | | | | |
|---|---|---|---|---|---|
| who | 0.26 | 0.44 | ... | 0.49 | -0.21 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |
| ? | 0.16 | 0.60 | ... | 0.18 | 0.36 |

### CONTEXT-TO-QUERY ATTENTION

$A^D Q =$

| Words | | | | | |
|---|---|---|---|---|---|
| who | 0.26 | 0.44 | ... | 0.49 | -0.21 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |
| ? | 0.16 | 0.60 | ... | 0.18 | 0.36 |

$Q_{6\times100} =$

$(A^D Q)_{10\times100} =$

| Words | | | | | |
|---|---|---|---|---|---|
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| boy | 0.25 | 0.44 | ... | 0.48 | -0.17 |
| in | -0.03 | -0.23 | ... | 0.82 | 0.26 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| red | -0.02 | -0.23 | ... | 0.81 | 0.26 |
| shirt | 0.34 | 0.27 | ... | 0.25 | 0.52 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |

$A^D_{10\times6} =$

| Words | Who | Went | To | The | Marke | ? |
|---|---|---|---|---|---|---|
| the | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| boy | 0.94 | 0.02 | 0.0 | 0.004 | 0.0 | 0.02 |
| in | 0.01 | 0.003 | 0.003 | 0.98 | 0.0 | 0.0 |
| the | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| red | 0.01 | 0.009 | 0.0 | 0.97 | 0.0 | 0.02 |
| shirt | 0.096 | 0.39 | 0.0 | 0.02 | 0.0 | 0.48 |
| went | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| to | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| the | 0.0 | 0.0 | 0.0 | 0.99 | 0.0 | 0.0 |
| market | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |

"It computes 10 ($T$) weighted-query representations, in light of every document word"
***Seems familiar ?***

### QUERY-TO-CONTEXT REPRESENTATION

**Query Aware Document Representation** $\tilde{D} = \text{softmax}(M_{10\times1})^T D$

| Words | | | | | |
|---|---|---|---|---|---|
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| boy | 0.89 | 0.37 | ... | 0.04 | -0.84 |
| in | 0.08 | -0.22 | ... | 0.75 | -0.34 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| red | -0.30 | 0.50 | ... | 0.12 | -0.48 |
| shirt | 0.11 | 0.51 | ... | -0.21 | -0.13 |
| went | 0.62 | -0.13 | ... | 0.25 | -0.27 |
| to | -0.18 | 0.05 | ... | 0.04 | -0.15 |
| the | -0.03 | -0.24 | ... | 0.83 | 0.27 |
| market | 0.39 | 0.23 | ... | 0.77 | 1.10 |

$D_{10\times100} =$

$\tilde{D} =$

| Words | -0.03 | -0.24 | ... | 0.81 | 0.25 |
|---|---|---|---|---|---|

$\text{softmax}(M_{10\times1}) =$

| Words | Who |
|---|---|
| the | 0.31 |
| boy | 0.0 |
| in | 0.01 |
| the | 0.31 |
| red | 0.0 |
| shirt | 0.002 |
| went | 0.02 |
| to | 0.0 |
| the | 0.31 |
| marke | 0.0 |

# REFERENCES:

- Dynamic Co-attention Networks
- Bidirectional Attention Flow