# Towards Interpreting BERT for Reading Comprehension Based QA

EMNLP 2020

**Sahana Ramnath**, Preksha Nema, Deep Sahni, Mitesh M. Khapra

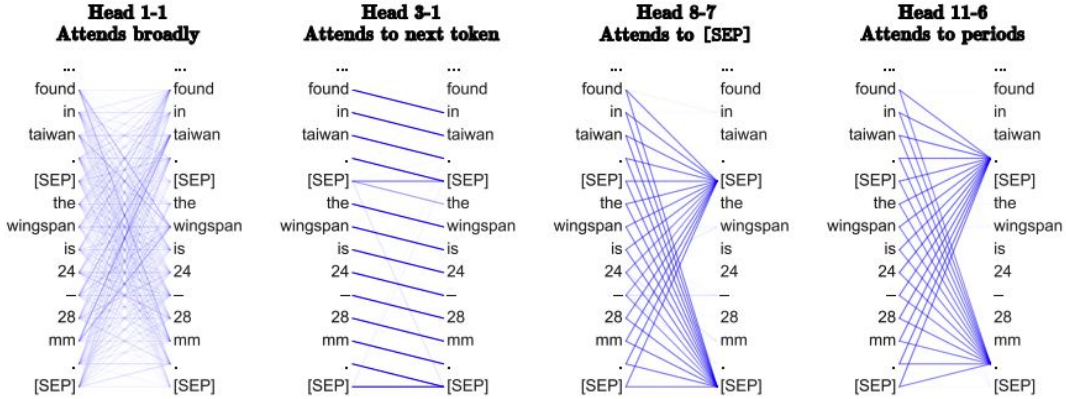
Robert Bosch Centre for Data Science and AI (RBC-DSAI)

Indian Institute of Technology Madras, Chennai, India

# Towards **Interpreting BERT** for Reading Comprehension Based QA

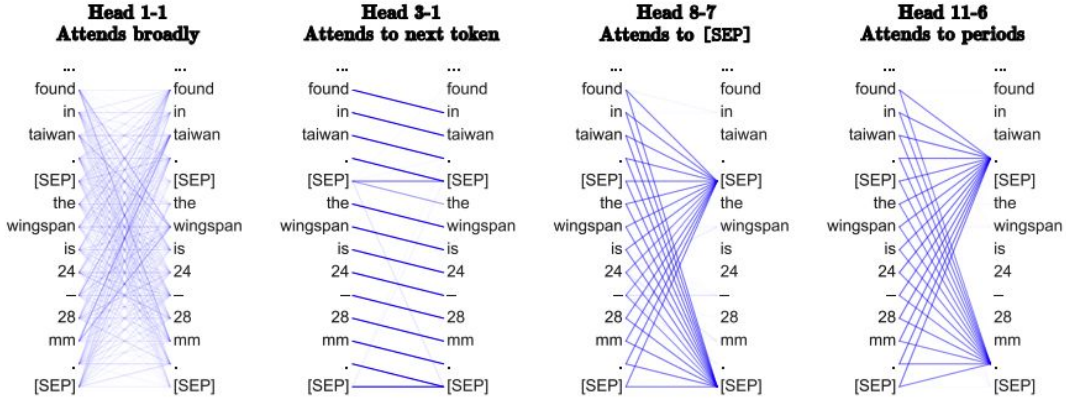
# Recent Works

- Syntactic and semantic roles played by BERT's layers
  [Tenney et al.,2019],[Peters et al.,2018]
- Linguistic phenomena in BERT's attention heads
  [Clark et al.,2019]



# Recent Works

- Syntactic and semantic roles played by BERT's layers [Tenney et al.,2019],[Peters et al.,2018]
- Linguistic phenomena in BERT's attention heads [Clark et al.,2019]




**Interpreting BERT has not been much explored for Complex tasks like QA.**

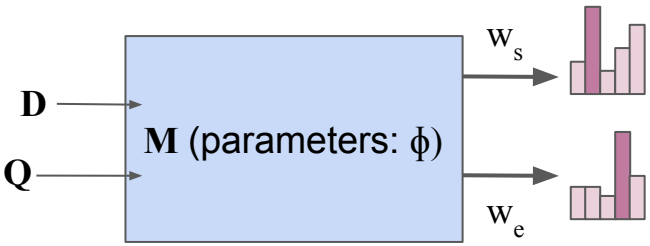
# Towards Interpreting BERT for **Reading Comprehension Based QA**

# RCQA Task

**Document (D):** The Panthers finished the regular season with a 15-1 record ... The Broncos finished the regular season with a 12-4 record.... They joined the Patriots, Dallas Cowboys, and Pittsburgh Steelers as one of **four teams** that have made eight appearances in the Super Bowl.

**Question (Q):** **How many teams have been in the super bowl eight times?**

# RCQA Task




$$p(w_s, w_e) = \text{argmax } M(w_s, w_e \mid \mathbf{D}, \mathbf{Q}, \phi)$$

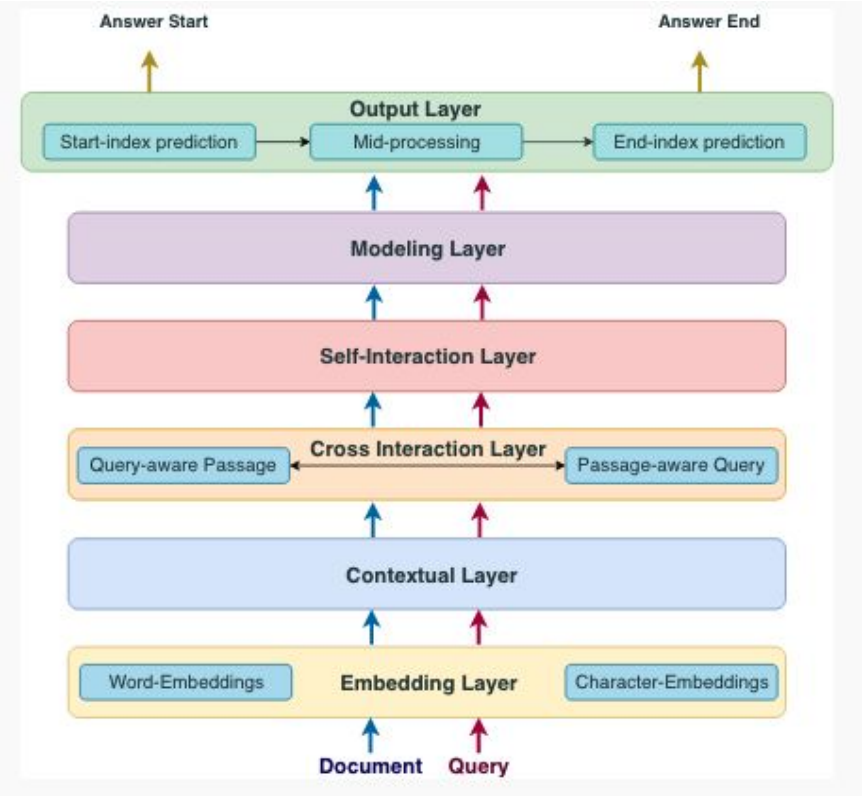
**Document (D):** The Panthers finished the regular season with a 15-1 record ... The Broncos finished the regular season with a 12-4 record.... They joined the Patriots, Dallas Cowboys, $w_s$ $w_e$ and Pittsburgh Steelers as one of **four teams** that have made eight appearances in the Super Bowl

**Question (Q):** **How many teams have been in the super bowl eight times?**

# Pre-BERT RCQA Models

Pre-BERT models intuitively modeled layers based on how an answer should be arrived at : BiDAF or DCN

[Seo et al.,2016],[Xiong et al.,2016]



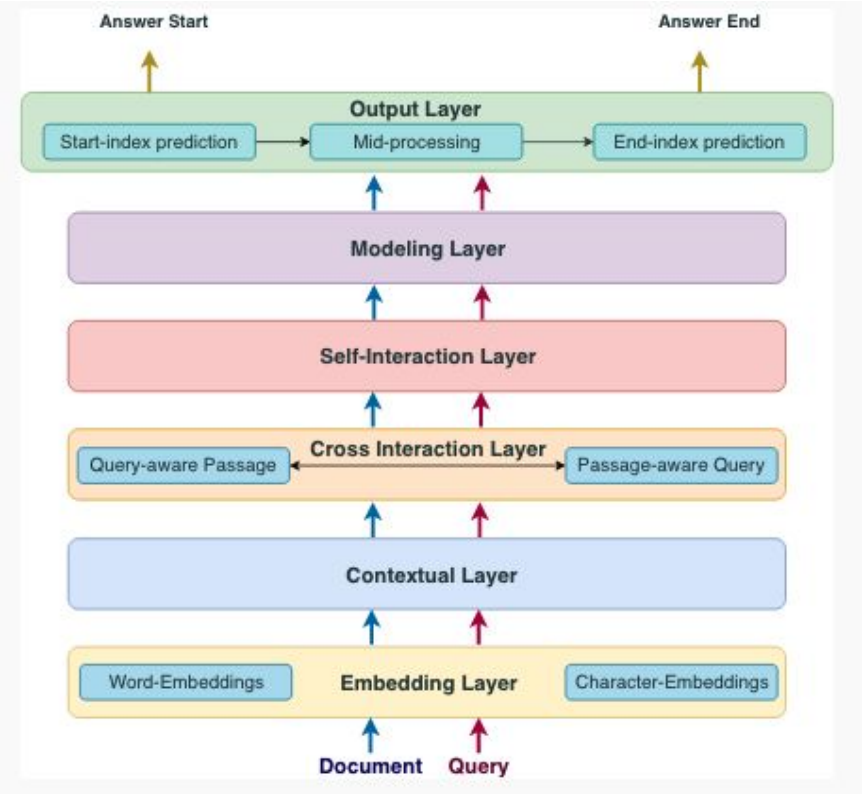
# Pre-BERT RCQA Models

**Contextual Layer:** Learns **contextual representations** for the passage and the question independently.

**Cross-Interaction Layer:** Attends to information in the passage specific to the question.

**Self-Interaction/Modelling Layer:** Refines the passage word representations.

**Output Layer:** Predicts the **answer**

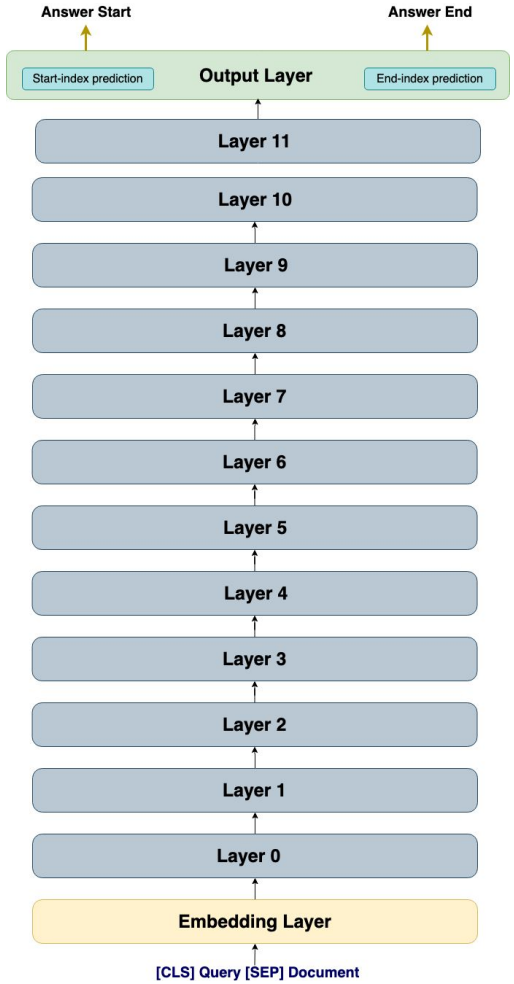


# BERT based RCQA Models

**BERT is a large 12/24 layer model, with all layers implementing the same transformer block function.**

**No pre-defined roles for layers in BERT**

**This combined with BERT's high non-linearity and number of parameters makes it *challenging* to analyse BERT for the task of RCQA.**



# Towards Interpreting BERT for Reading Comprehension Based QA

# This Work

We analyze each of **BERT's layers as one unit,** for RCQA.

- We first explicitly define mathematical functionalities for layers using the attribution method **Integrated Gradients**.

- We then analyze each layer's functionality with respect to the pre-defined roles modeled in earlier RCQA systems.

# Experimental Setup: Datasets

**SQuAD 1.1**

- 90K/10K train/dev samples
- 100-300 words passage
- Natural language question
- Answer span in passage itself

**DuoRC - SelfRC dataset**

- 60K/13K train/dev samples
- 500 words on average passage
- Natural language question
- Answer span in passage itself

# Layer Level Functionality - Importance Scores

We use the layer functionality to obtain a **distribution across passage words** that represents how important these words are for the answer prediction at that layer.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ... | 0.001 | 0.01 | 0.0 | 0.001 | 0.003 | 0.005 | ... | 0.61 | 0.31 | ... |
| ... | They | joined | the | Patriots | Dallas | Cowboys | ... | **four** | **teams** | ... |

**Layer-11**

.
.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ... | 0.1 | 0.2 | 0.0 | 0.07 | 0.10 | 0.15 | ... | 0.33 | 0.33 | ... |
| ... | They | joined | the | Patriots | Dallas | Cowboys | ... | **four** | **teams** | ... |

**Layer-k**

# Obtaining importance scores:

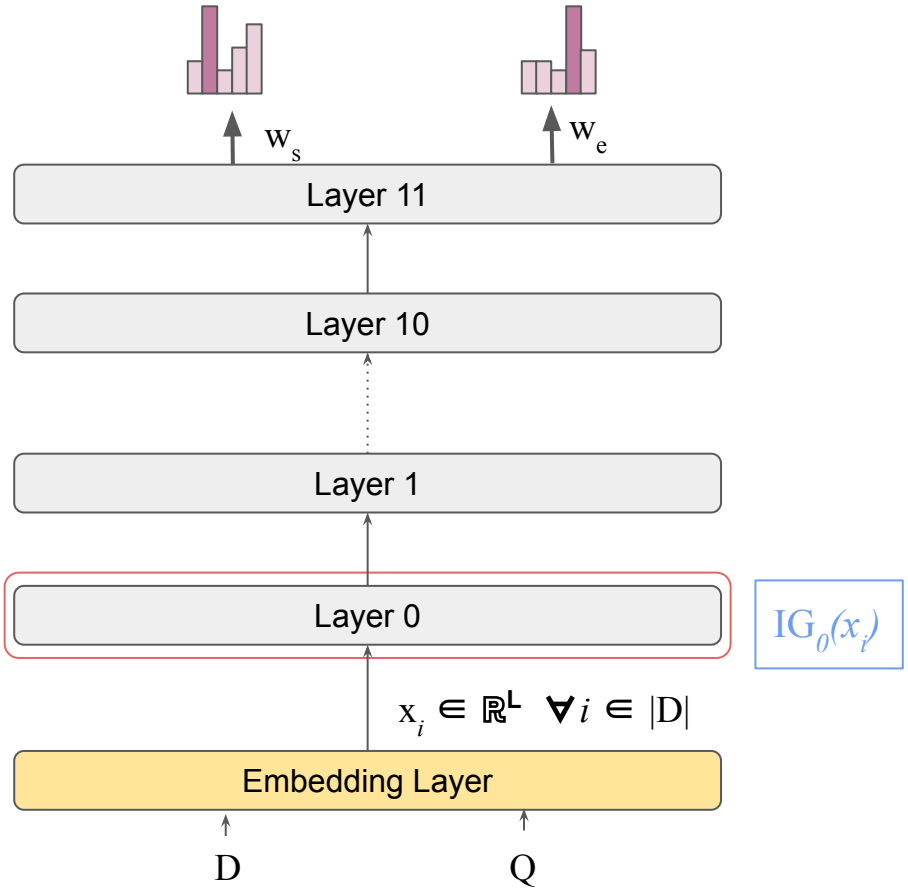
Many techniques to attribute a deep network's predictions to its input features - LIME, **Integrated Gradients**, DeepLift, LRP,etc.

The Integrated Gradients for a passage word $w_i$, represented as $x_i \in \mathbb{R}^L$ is computed as follows:

If a word w gets a **higher IG** score, it means that it is **more important** to the final prediction with respect to that layer.

$$IG(x_i) = \int_{\alpha=0}^{1} \frac{\partial M(\tilde{x} + \alpha(x_i - x_{baseline}))}{\partial x_i} \, d\alpha$$

# Obtaining importance scores for each Layer:



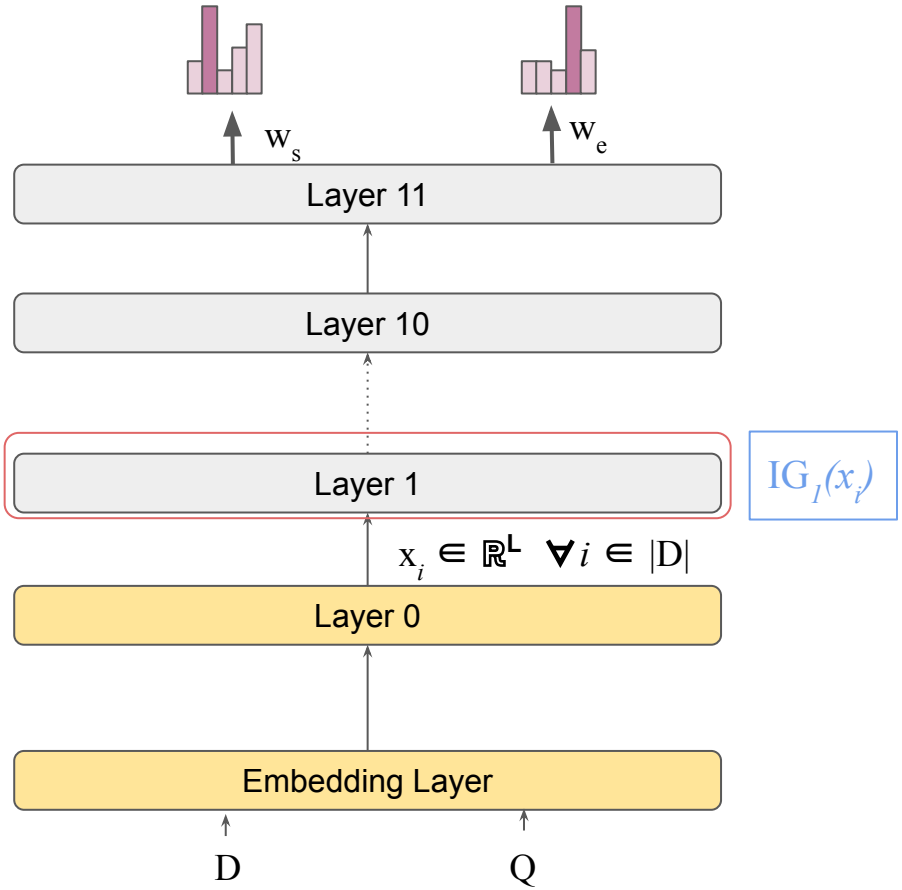

$IG_0(x_i)$

The Integrated Gradient for a passage word $w_i$, represented as $x_i \in \mathbb{R}^L$ is computed as follows:

A **higher IG score** highlights **the importance** of word $w_i$ in the final prediction.

$$IG(x_i) = \int_{\alpha=0}^{1} \frac{\partial M(\tilde{x} + \alpha(x_i - x_{baseline}))}{\partial x_i} \, d\alpha$$

# Obtaining importance scores across Layers:




The Integrated Gradient for a passage word $w_i$, represented as $x_i \in \mathbb{R}^L$ is computed as follows:

A **higher IG score** highlights **the importance** of word $w_i$ in the final prediction.

$$IG(x_i) = \int_{\alpha=0}^{1} \frac{\partial M(\tilde{x} + \alpha(x_i - x_{baseline}))}{\partial x_i} \, d\alpha$$

# Importance Distribution - Qualitative Example

**Question:** **Why was Polonia relegated from the country's top flight in 2013 ?**
**Answer:** disastrous financial situation

| Initial Layers | |
|---|---|
| **Layer 0** | Polonia was relegated from the country's top flight in 2013 because of their disastrous financial situation. They are now playing in the 4th league … |
| **Layer 1** | Polonia was relegated from the country's top flight in 2013 because of their disastrous financial situation. They are now playing in the 4th league … |

| Final Layers | |
|---|---|
| **Layer 10** | Polonia was relegated from the country's top flight in 2013 because of their disastrous financial situation. They are now playing in the 4th league … |
| **Layer 11** | Polonia was relegated from the country's top flight in 2013 because of their disastrous financial situation. They are now playing in the 4th league … |

# What makes a layer unique?

We compare the head and tail of each layer's distribution using **Jensen-Shannon Divergence**, to understand what makes each layer unique.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ... | 0.001 | 0.01 | 0.0 | 0.001 | 0.003 | 0.005 | ... | 0.61 | 0.31 | ... |
| ... | They | joined | the | Patriots | Dallas | Cowboys | ... | **four** | **teams** | ... |

.
.

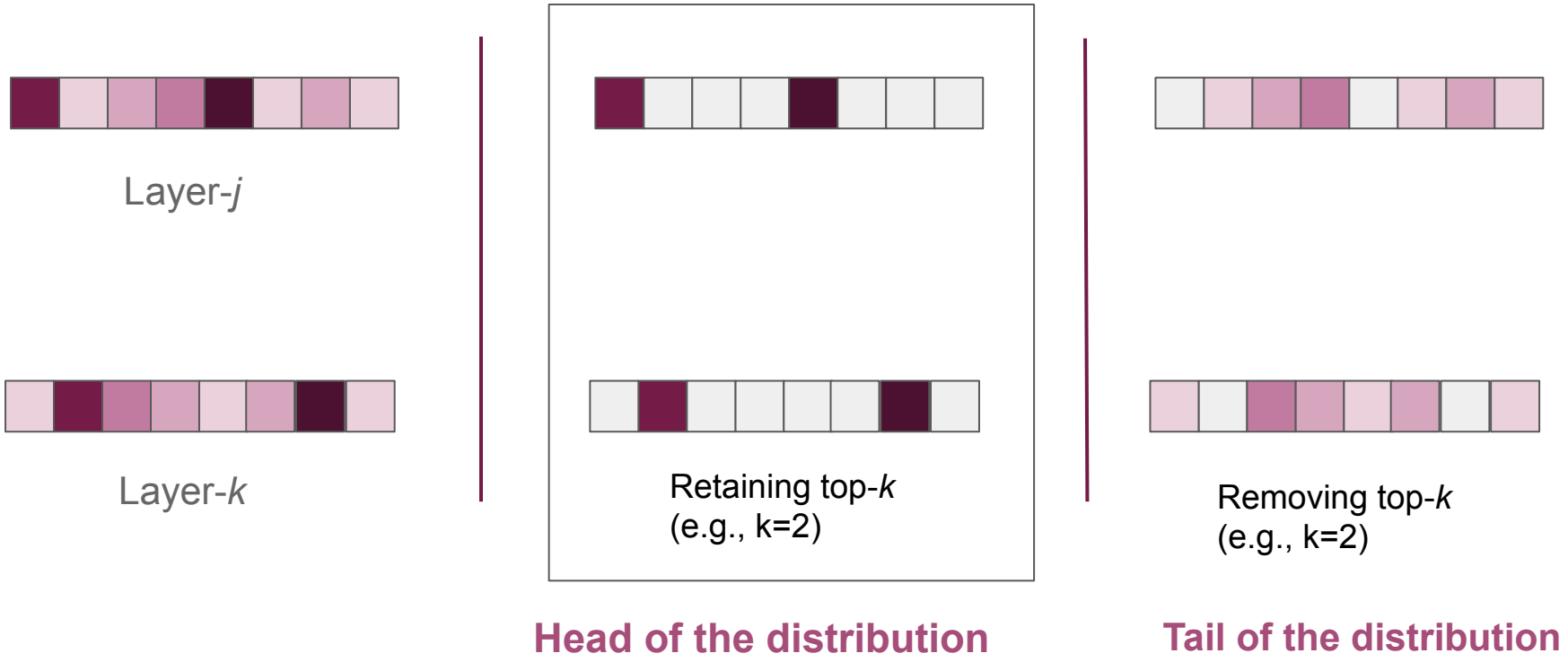
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ... | 0.1 | 0.2 | 0.0 | 0.07 | 0.10 | 0.15 | ... | 0.33 | 0.33 | ... |
| ... | They | joined | the | Patriots | Dallas | Cowboys | ... | **four** | **teams** | ... |

Layer-*j*

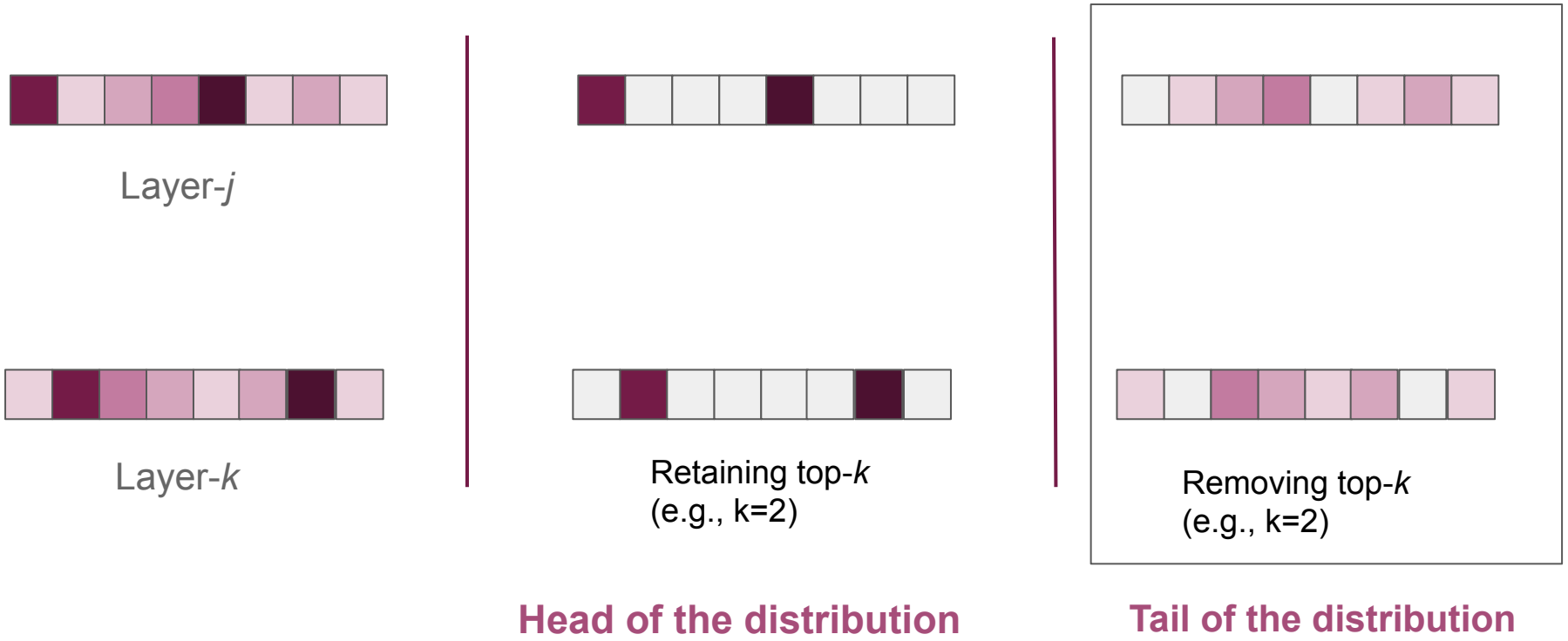
**JSD**

Layer-*k*

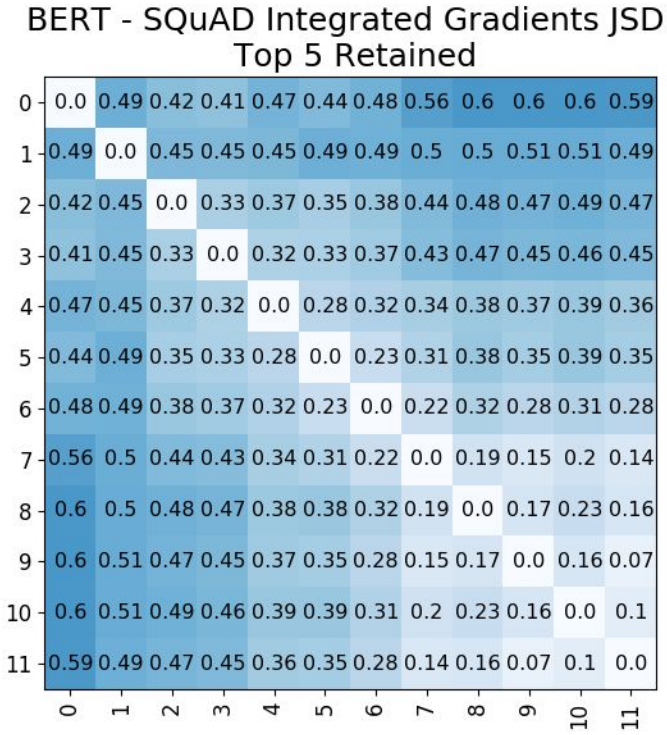
# Comparison between distributions of two layers:



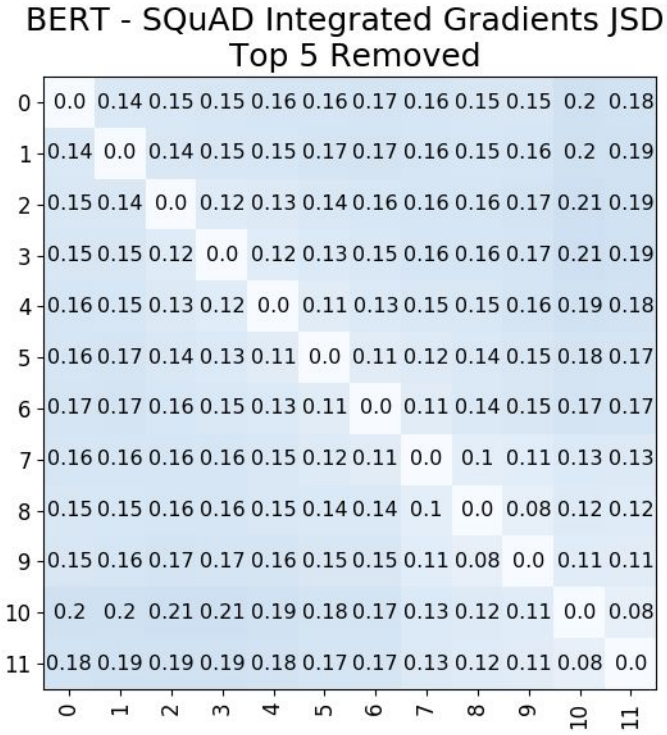
# Comparison between distributions of two layers:



# Comparison between distributions across two layers

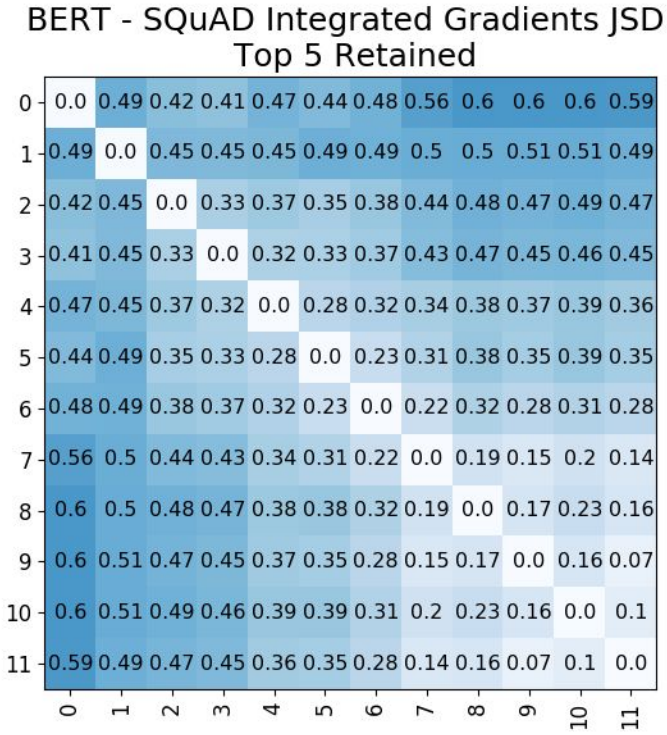


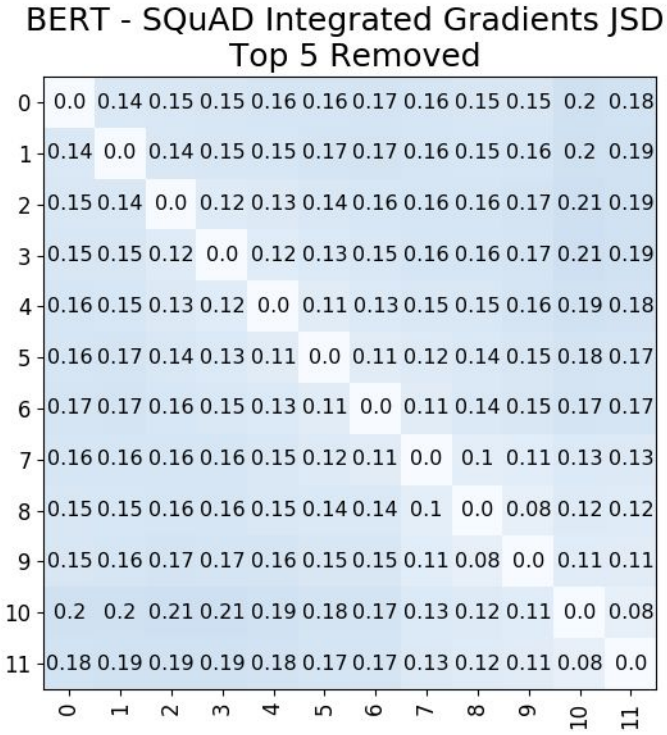

Higher values when top-5 words are retained (min 0.07/max 0.6)




Lower values when top-5 words are removed (min 0.09/max 0.21)

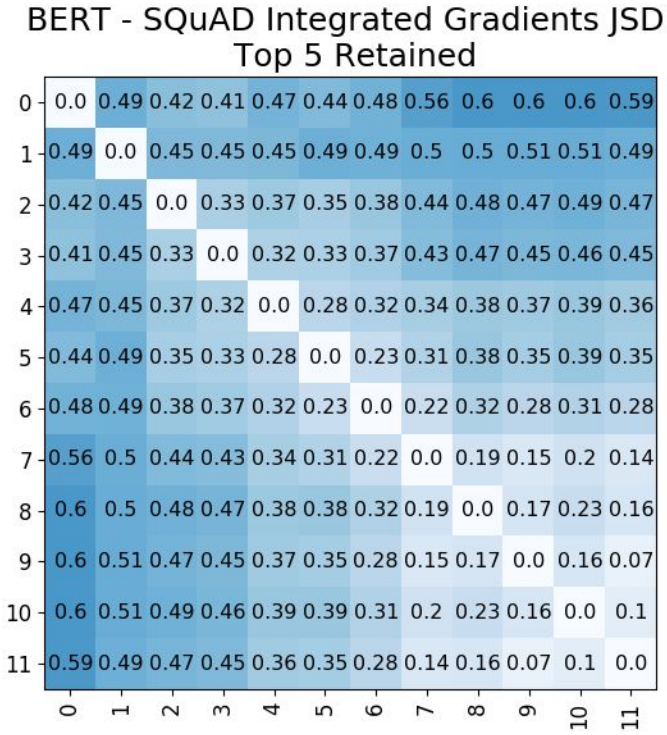
# Comparison between distributions across two layers

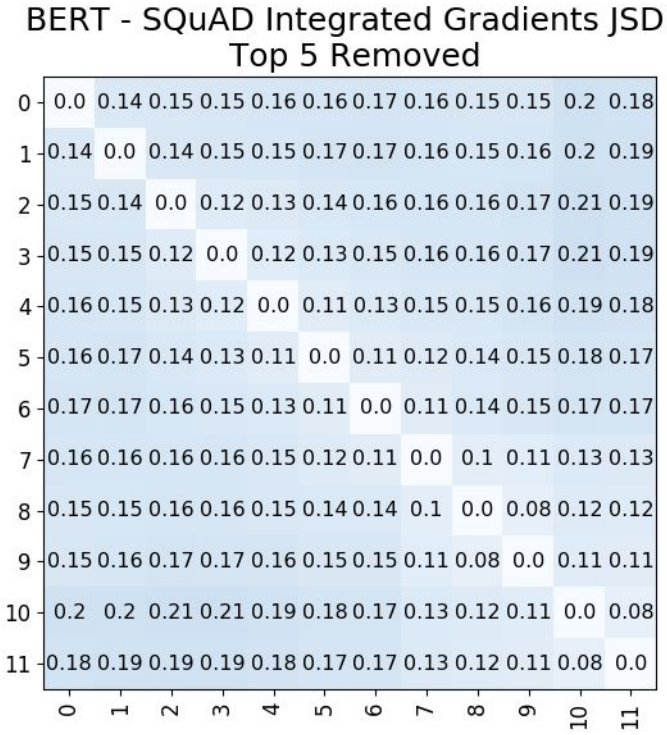



A layer's functionality is reflected by the **head (top-k important words)**.

# Comparison between distributions across two layers




Hence, for semantic analysis, we take the **top-5** words highlighted by a layer to **represent** that layer.

# Results and Discussions

# Probing Layers : QA Functionality

We analyze the top-5 important words of each layer, to see which of them focus on the question, the context around the answer and the answer span.

**Document (D):** The Panthers finished the regular season with a 15-1 record ... The Broncos finished the regular season with a 12-4 record.... They joined the Patriots, Dallas Cowboys, and Pittsburgh Steelers as one of **four teams** that have made eight appearances in the Super Bowl

**Question (Q):** **How many teams have been in the super bowl eight times?**

# Probing Layers : QA Functionality

We analyze the top-5 important words of each layer, to see which of them focus on the question, the context around the answer and the answer span.

**Answer Words**
Words in the answer span in the passage

**Document (D):** The Panthers finished the regular season with a 15-1 record ... The Broncos finished the regular season with a 12-4 record.... They joined the Patriots, Dallas Cowboys, and Pittsburgh Steelers as one of **four teams** that have made eight appearances in the Super Bowl

**Question (Q):** **How many teams have been in the super bowl eight times?**

# Probing Layers : QA Functionality

We analyze the top-5 important words of each layer, to see which of them focus on the question, the context around the answer and the answer span.

**Answer Words**
Words in the answer span in the passage

**Query Words**
Words in the question that appear in the passage (except stopwords)

**Document (D):** The Panthers finished the regular season with a 15-1 record ... The Broncos finished the regular season with a 12-4 record.... They joined the Patriots, Dallas Cowboys, and Pittsburgh Steelers as one of **four teams** that have made **eight** appearances in the **Super Bowl.**

**Question (Q):** **How many teams have been in the super bowl eight times?**

# Probing Layers : QA Functionality

We analyze the top-5 important words of each layer, to see which of them focus on the question, the context around the answer and the answer span.

**Answer Words**
Words in the answer span in the passage

**Query Words**
Words in the question that appear in the passage (except stopwords)

**Supporting Words**
Words surrounding the answer within a window size of 5

**Document (D):** The Panthers finished the regular season with a 15-1 record ... The Broncos finished the regular season with a 12-4 record.... They joined the Patriots, Dallas Cowboys, and *Pittsburgh Steelers as one of* four teams *that have made eight appearances* in the Super Bowl.

**Question (Q):** **How many teams have been in the super bowl eight times?**

# Probing Layers : QA Functionality

| Layer | % answer span | % Q-words | % Support Words |
|---|---|---|---|
| L0 | 26.99 | | |
| L1 | 26.09 | | |
| L2 | 29.9 | | |
| L3 | 30.44 | | |
| L4 | 30.06 | | |
| L5 | 30.75 | | |
| L6 | 31.25 | | |
| L7 | 32.37 | | |
| L8 | 30.78 | | |
| L9 | 34.58 | | |
| L10 | 34.31 | | |
| L11 | 34.63 | | |

## Observations

- Later layers focus on enhancing and verifying the model's prediction.

# Probing Layers : QA Functionality

| Layer | % answer span | % Q-words | % Support Words |
|---|---|---|---|
| L0 | 26.99 | 22.94 | |
| L1 | 26.09 | 24.35 | |
| L2 | 29.9 | 22.41 | |
| L3 | 30.44 | 19.55 | |
| L4 | 30.06 | 18.33 | |
| L5 | 30.75 | 14.71 | |
| L6 | 31.25 | 15.33 | |
| L7 | 32.37 | 12.29 | |
| L8 | 30.78 | 18.91 | |
| L9 | 34.58 | 10.21 | |
| L10 | 34.31 | 10.56 | |
| L11 | 34.63 | 12.0 | |

## Observations

- Later layers focus on enhancing and verifying the model's prediction.
- Initial layers focus on connecting the query and passage.

# Probing Layers : QA Functionality

| Layer | % answer span | % Q-words | % Support Words |
|---|---|---|---|
| L0 | 26.99 | 22.94 | 9.45 |
| L1 | 26.09 | 24.35 | 9.43 |
| L2 | 29.9 | 22.41 | 11.65 |
| L3 | 30.44 | 19.55 | 11.13 |
| L4 | 30.06 | 18.33 | 11.23 |
| L5 | 30.75 | 14.71 | 11.57 |
| L6 | 31.25 | 15.33 | 11.94 |
| L7 | 32.37 | 12.29 | 12.32 |
| L8 | 30.78 | 18.91 | 12.07 |
| L9 | 34.58 | 10.21 | 13.41 |
| L10 | 34.31 | 10.56 | 13.39 |
| L11 | 34.63 | 12.0 | 13.74 |

## Observations

- Later layers focus on enhancing and verifying the model's prediction.
- Initial layers focus on connecting the query and passage.
- Contextual role increases from the initial to the final layers.

# Visualizing Word Representations

**Document (D):** The Panthers finished the regular season with a 15-1 record ... The Broncos finished the regular season with a 12-4 record.... *They joined the Patriots, Dallas Cowboys, and Pittsburgh Steelers as one of four teams that have made* **eight** *appearances in the Super Bowl*

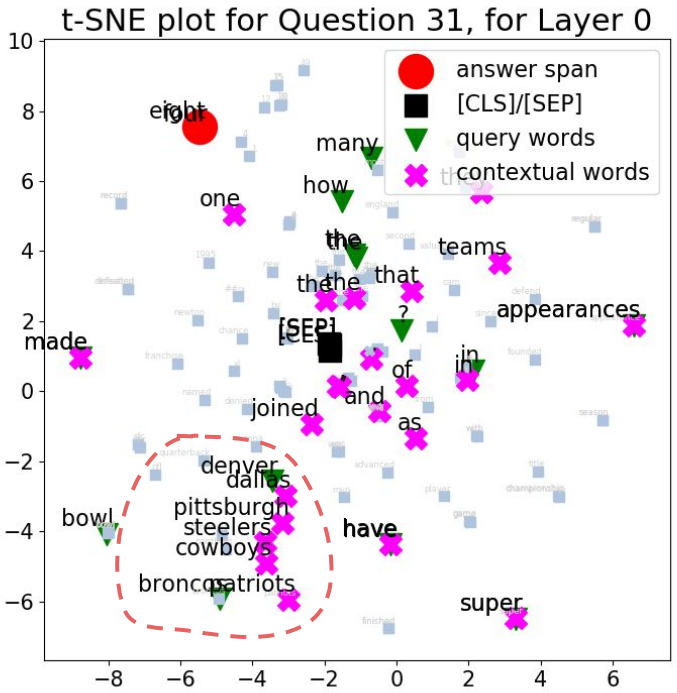
**Question (Q):** **How many appearances have the Broncos made in the super bowl?**

We perform qualitative analysis of word embeddings using **t-SNE plots.**

# Visualizing Word Representations

**Layer 0:**

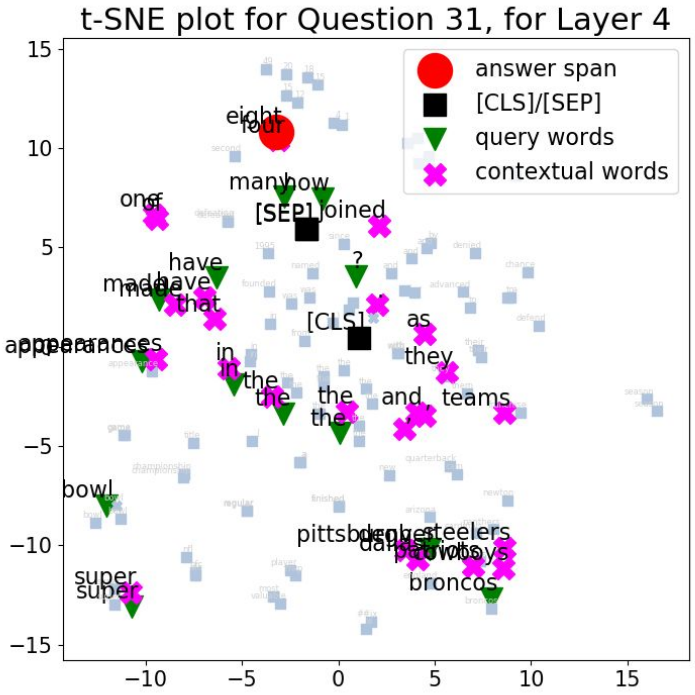
Similar words (e.g., team names, stop words) are close to each other.



# Visualizing Word Representations

**Intermediate Layers: such as Layer 4**

All the contextual, answer and question words intermingle.



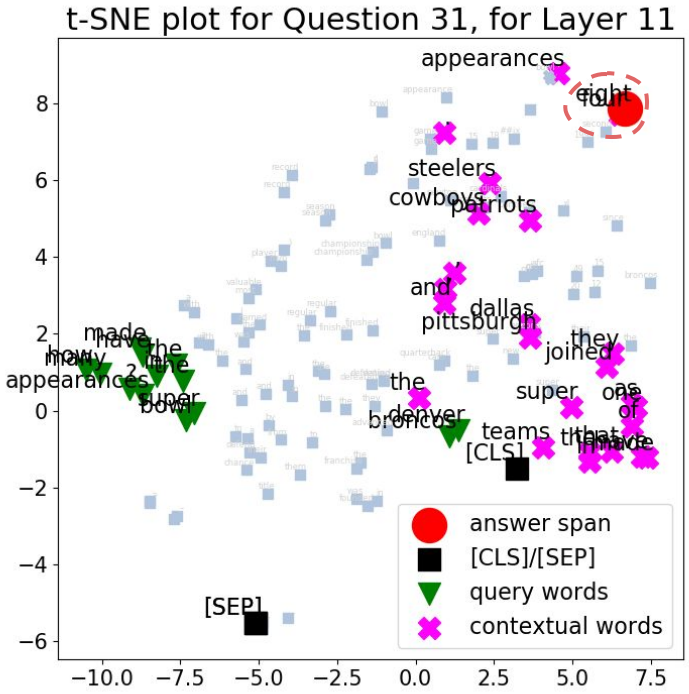
# Visualizing Word Representations

## Later Layers: Layer (9-11)

The answer **eight** segregates from other words.

The question words separate from the answer and the supporting words

However, numerical entity **four**, is very close to the answer, across all 12 layers.



# Analysis on Quantifier Questions

**Question Type**

- How much?
- How many?

**Confusing Words**

- Many numerical entities in the passage
- Could lead to the model getting confused

**Document (D):** The Panthers finished the regular season with a **15-1** record ... The Broncos finished the regular season with a **12-4** record.... They joined the Patriots, Dallas Cowboys, and Pittsburgh Steelers as one of **four** teams that have made **eight** appearances in the Super Bowl.

**Question (Q):** **How many appearances have the Broncos made in the super bowl?**

**Answer (A):** **eight**

# Analysis on Quantifier Questions

We measure the ratio of **confusing words ($N_q$)** marked important by each layer.

$$\text{ratio} = \frac{\text{card}(N_q \text{ in top-5})}{\text{card}(N_q \text{ in passage})}$$

# Analysis on Quantifier Questions

**Interestingly, we observe that this ratio increases as we go from layers 0 to 11.**

BERT, in its later layers, **distributes its importance over potentially confusing words**

| Layer | SQuAD | DuoRC |
|---|---|---|
| *Layer 0* | 5.6% | 12.9% |
| *Layer 10* | 17.7% | 21.6% |
| *Layer 11* | 15.5% | 22.6% |

# Analysis on Quantifier Questions

**Interestingly, we observe that this ratio increases as we go from layers 0 to 11.**

BERT, in its later layers, **distributes its importance over potentially confusing words**

| Layer | SQuAD | DuoRC |
|---|---|---|
| *Layer 0* | 5.6% | 12.9% |
| *Layer 10* | 17.7% | 21.6% |
| *Layer 11* | 15.5% | 22.6% |

*However, BERT still predicts the correct answer for such questions with a high confidence of ~85%*

# Analysis on Quantifier Questions

**Interestingly, we observe that this ratio increases as we go from layers 0 to 11.**

BERT, in its later layers, **distributes its importance over potentially confusing words**

| Layer | SQuAD | DuoRC |
|---|---|---|
| *Layer 0* | 5.6% | 12.9% |
| *Layer 10* | 17.7% | 21.6% |
| *Layer 11* | 15.5% | 22.6% |

*However, BERT still predicts the correct answer for such questions with a high confidence of ~85%*

*(**87.35%** accuracy for such questions for SQuAD, and **53.5%** in DuoRC)*

# Open-ended questions

1. If the focus on confusing words increases from the initial to later layers, how does BERT still have a high accuracy?
2. Why do the question word representations move away from contextual and answer representation in later layers?

**We hope that our work will help the research community interpret BERT for other complex tasks and explore the above open-ended questions.**

# THANK YOU