Text classification

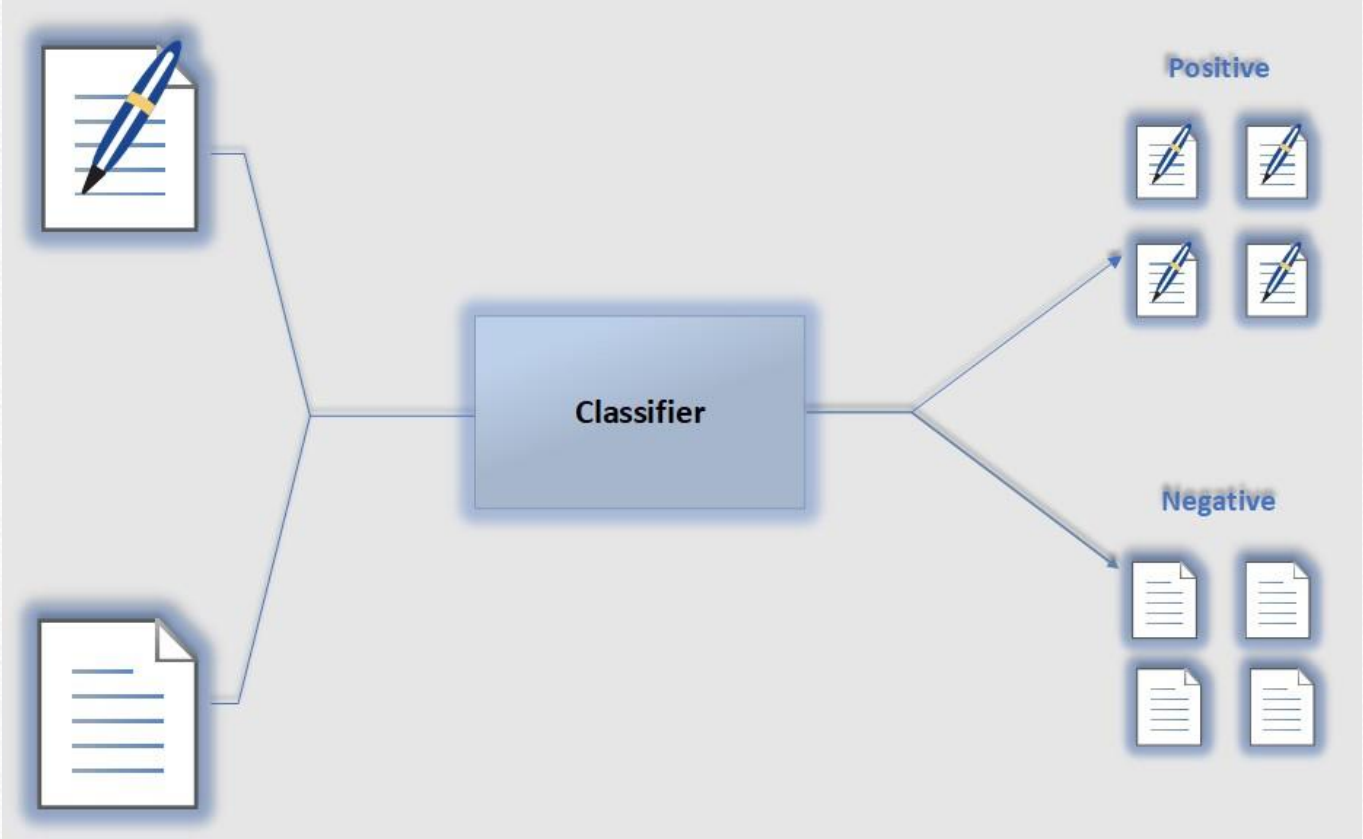
Peyman Naseri

ADS course presentation

Fall 01-02

TeIAS | Tehran Institute for Advanced Studies

Sentiment Analysis ☺/☹



خبرگزاری مهر

واکنش سردار قاآنی به مساله حجاب

صحبت‌های سردار قاآنی فرمانده سپاه قدس را در خصوص مسئله حجاب را در این فیلم مشاهده می‌کنید. کد خبر .5674639

1 day ago

خبرگزاری مهر

آماده باش ۳۰۶ گروه عملیاتی شرکت توزیع برق در مواقع بحرانی

رشت- مدیرعامل شرکت توزیع برق گیلان با اشاره به هشدار سازمان هواشناسی مبنی بر ورود موج جدید بارشی و کاهش دما گفت: ۳۰۶ گروه عملیاتی شرکت توزیع...

19 mins ago

خبرآنلاین

مجری پیشین تلویزیون هم آزاد شد

ایسنا نوشت:شیرین صمدی - مجری پیشین تلویزیون - با انتشار ویدیویی در فضای مجازی از آزادی خود خبر داد. او اجرای برنامه‌هایی، چون «صبح به‌خیر...

5 hours ago

جامعه خبری تحلیلی الف

ماجرای خبر ترور قاضی صلواتی چیست؟

مدیرعامل خبرگزاری قوه قضاییه ترور قاضی صلواتی را تکذیب کرد.

12 hours ago

BBC

تایید خبر کناره‌گیری معاون رابرت مالی از هیات مذاکره‌کننده با ایران

ند پرایس، سخنگوی وزارت امور خارجه آمریکا، گزارش‌های مربوط به کناره‌گیری جرت بلانک، معاون راب مالی در هیات مذاکره کننده هسته‌ای ایالات متحده...

1 day ago

خبرگزاری تسنیم

حمله سایبری گسترده به بانک مرکزی خنثی شد

12 hours ago

رادیو فردا

بانک مرکزی ایران هدف حمله سایبری قرار گرفت

9 hours ago

خبرگزاری فارس

حمله سایبری به بانک مرکزی، بله و روبیکا ناکام ماند

11 hours ago


Twit
Photo illustration

As Data Scientist ☺

- Data
  - Data Collection
  - Preprocessing
  - EDA
  - Feature Engineering (Text Representation)
- Models
  - Classic
  - Neural Network

Data Collection

A glimpse of the data...

خبر	برچسب
عصبانیت شدید مسی از لاپورتا	رئیس بارسلونا تهدید شد	غیرجریان‌ساز
حضور اقشار مختلف در مراسم اربعین امسال حاکی از اهمیت امام حسین برای مردم حتی پس از دو سال وقفه در مراسمات به دلیل بیماری کووید، می‌باشد	جریان‌ساز
مکزیک نخستین مورد آبله میمونی را تایید می‌کند، کارشناسان جهانی خواستار اقدامات بیشتر در برابر بیماری هستند	جریان‌ساز
تاکید وزیر آموزش و پرورش به لزوم تحول آموزشی در ایران با ورود متاورس و بلاکچین	جریان‌ساز
”پل طبیعت“ تهران دو شب متوالی خاموش می‌شود	غیرجریان‌ساز

Preprocessing

Text	فیض رُوح القدس ار باز مـــدد فرماید	دیـــگران هم بکنند آن چه مَسیحا می کرد
Normalization	فیض روح القدس اگر باز مدد فرماید	دیگران هم بکنند آن چه مسیح می کرد
Tokenization	فیض روح‌القدس اگر باز مدد فرماید	دیگران هم بکنند آن‌چه مسیح می‌کرد

Another Preprocessing Techniques

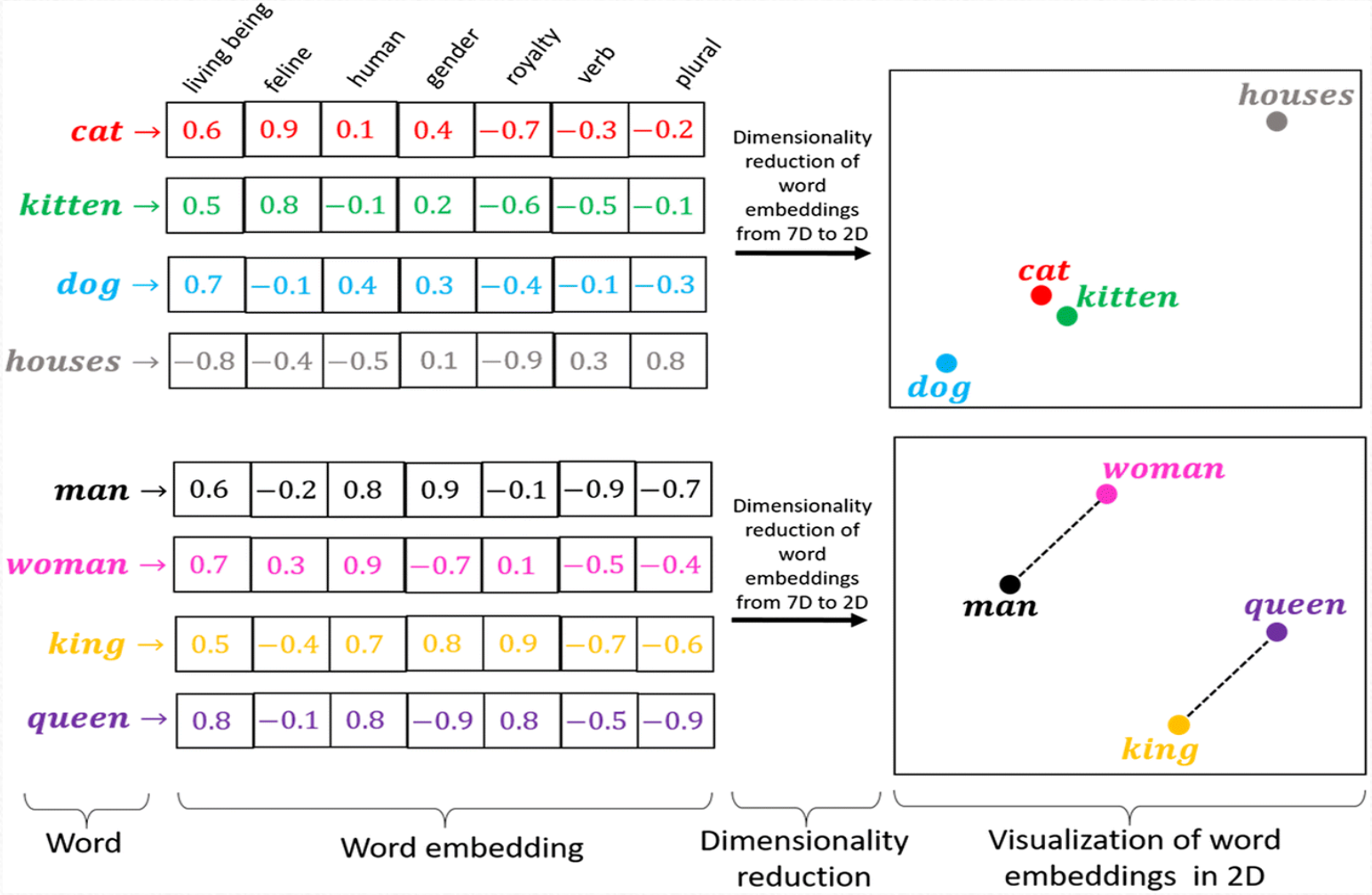
- Remove Punctuation
- Remove Stop-words
- Spelling Correction
- Stemming & Lemmatization
- Remove Numbers
- …

EDA

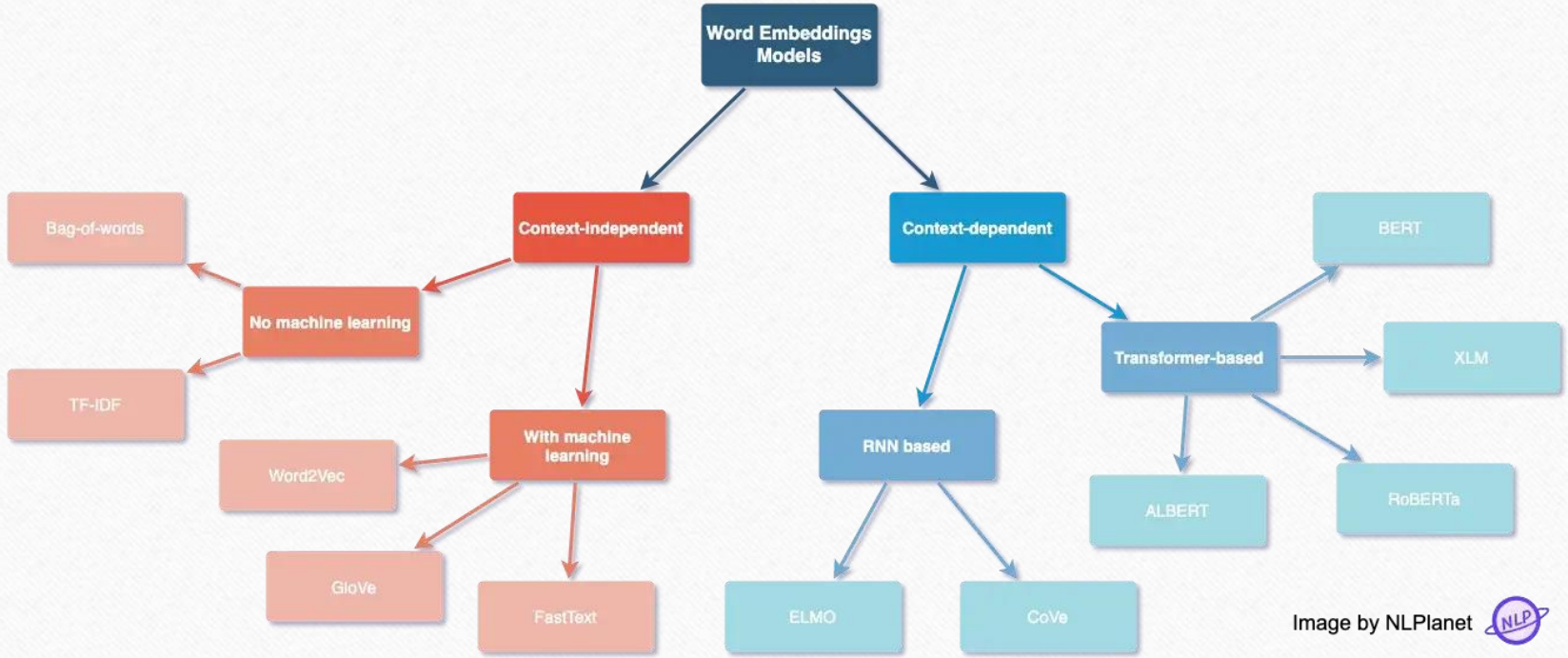
- Unbalanced Data
- Text Length Distribution
- Visualizing the Most Frequent Word
- WordCloud
- Visualizing of the Sentimental Words

Feature Engineering
or
Text Representation
or
Word Embedding

Good Embedding



Word Embedding Techniques



Word Embedding

- Non-Context Based
  - One-Hot
  - TF-IDF
- Context Based (Language Model)
  - RNN
  - Transformers

Non-Context Based

One-Hot

Dictionary

words

	آبی	ایران	اتمام	است	بیمار	بندر	...
ذخایر	0	0	0	0	0	0	
آبی	1	0	0	0	0	0	
ایران	0	1	0	0	0	0	
در	0	0	0	0	0	0	
حال	0	0	0	0	0	0	
اتمام	0	0	1	0	0	0	
است	0	0	0	1	0	0	

One-Hot

Advantages

Disadvantages

Term Frequency (TF)

Documents

words

	BBC	Iran-int	اقتصاد	ورزشی	کیهان	....
رئیسی	100	50	30	5	500	
دلار	50	100	300	3	60	
علی دایی	60	50	0	130	5	
مهسا امینی	400	500	5	10	1	
اعدام	100	50	45	2	70	
یارانه	10	15	80	0	27	
…						

TF-IDF

TF-IDF is a measure of originality of a word by comparing the number of times a word appears in a doc with the number of docs the word appears in.

$$\text{TF-IDF} = \text{TF}(t, d) \times \text{IDF}(t)$$

Term frequency

Number of times term t appears in a doc, d

Inverse document frequency

$$\log \frac{1 + n}{1 + \text{df}(d, t)} + 1$$

n ← # of documents

df(d, t) ← Document frequency of the term t

TF-IDF

Advantages

Disadvantages

Context Based

Language Model

- Limited words $\sum$
- Unlimited Sentence $\sum^*$
- Are all possible sentences produced with these alphabets and words?

$$p(start, w_1, w_2, \ldots, w_n, stop)$$

Language Model

- Classic
  - **n-gram**
- Neural Network
  - **BERT**
  - CBOW
  - Skip-gram
  - ELMO

n-gram

$$p(\text{start}, w_1, w_2, \ldots, w_n, \text{stop}) = \prod_{i=1}^{n+1} \gamma(w_i \mid w_1, w_2, \ldots, w_{i-1})$$

P("تاب بنفشه می دهد طره مشک سای تو") =

P(تاب) × P(تاب|بنفشه) × P(تاب بنفشه|می دهد)

× P(تاب بنفشه می دهد|طره) × P(تاب بنفشه می دهد طره |مشک سای)

× P(تاب بنفشه می دهد طره مشک سای |تو)

Unigram (Bag of Word)

$$p(\text{start}, w_1, w_2, \ldots, w_n, \text{stop}) = \prod_{i=1}^{n+1} \gamma(w_i)$$

P("تاب بنفشه می دهد طره مشک سای تو") =

P(تاب) × P(بنفشه|تاب) × P(می دهد|تاب بنفشه)

× P(طره|تاب بنفشه می دهد) × P(مشک سای|تاب بنفشه می دهد طره)

× P(تو|تاب بنفشه می دهد طره مشک سای)

= P(تاب) × P(بنفشه) × P(می دهد) × P(طره) × P(مشک سای) × P(تو)

Bigram

$$p(\text{start}, w_1, w_2, \ldots, w_n, \text{stop}) = \prod_{i=1}^{n+1} \gamma(w_i \mid w_{i-1})$$

P("تاب بنفشه می دهد طره مشک سای تو") =

= P(تاب) × P(بنفشه|تاب) × P(می دهد|بنفشه تاب)

× P(طره|می دهد بنفشه تاب) × P(مشک سای|طره می دهد بنفشه تاب)

× P(تو|مشک سای طره می دهد بنفشه تاب)

= P(تاب|شروع) × P(بنفشه|تاب) × P(می دهد|بنفشه تاب)

× P(طره|می دهد بنفشه تاب) × P(مشک سای|طره می دهد بنفشه تاب)

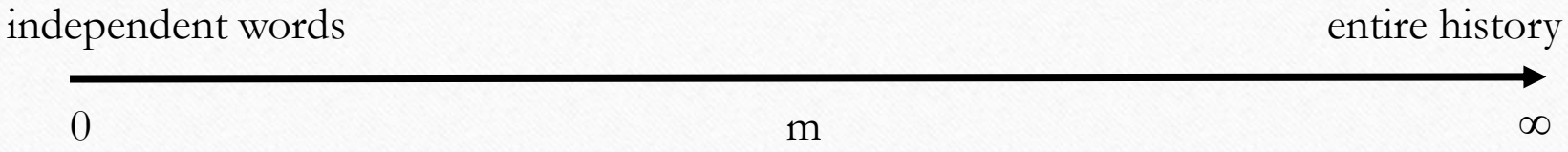
× P(تو|مشک سای طره می دهد بنفشه تاب)

m-gram

- Estimation of probabilities
  - MLE :

$$P(w_n | w_{n-m+1:n-1}) = \frac{Count(w_{n-m+1:n-1} w_n)}{Count(w_{n-m+1:n-1})}$$

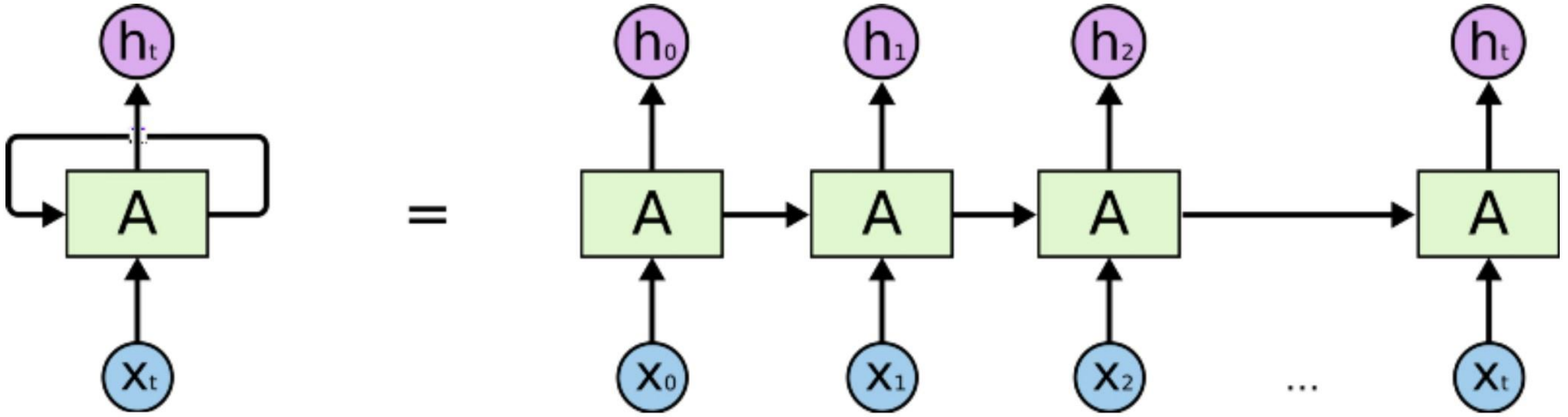
- Effect of m:



Neural Network

- MLP
- RNN
  - Vanilla
  - LSTM
- Attention Mechanism
  - Transformers
  - Bert

Vanilla

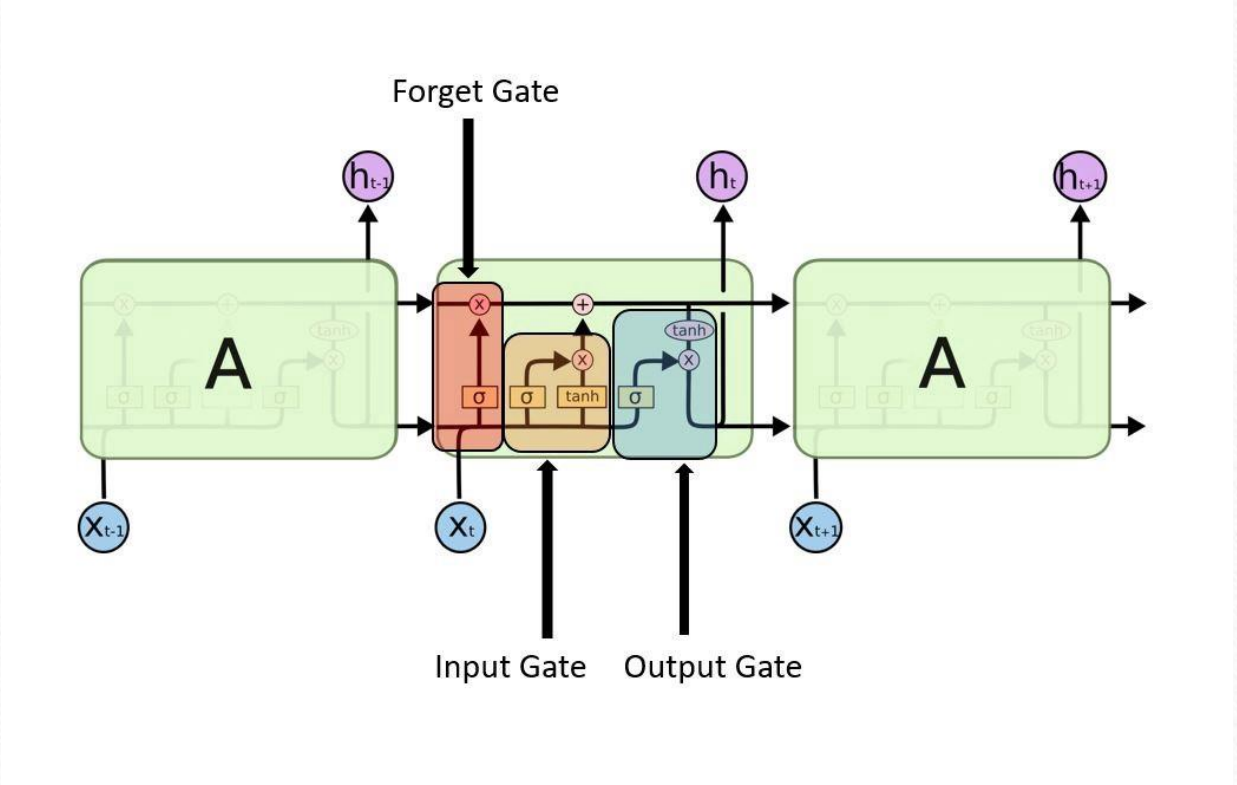


$$h_t = f_W(h_{t-1}, x_t)$$

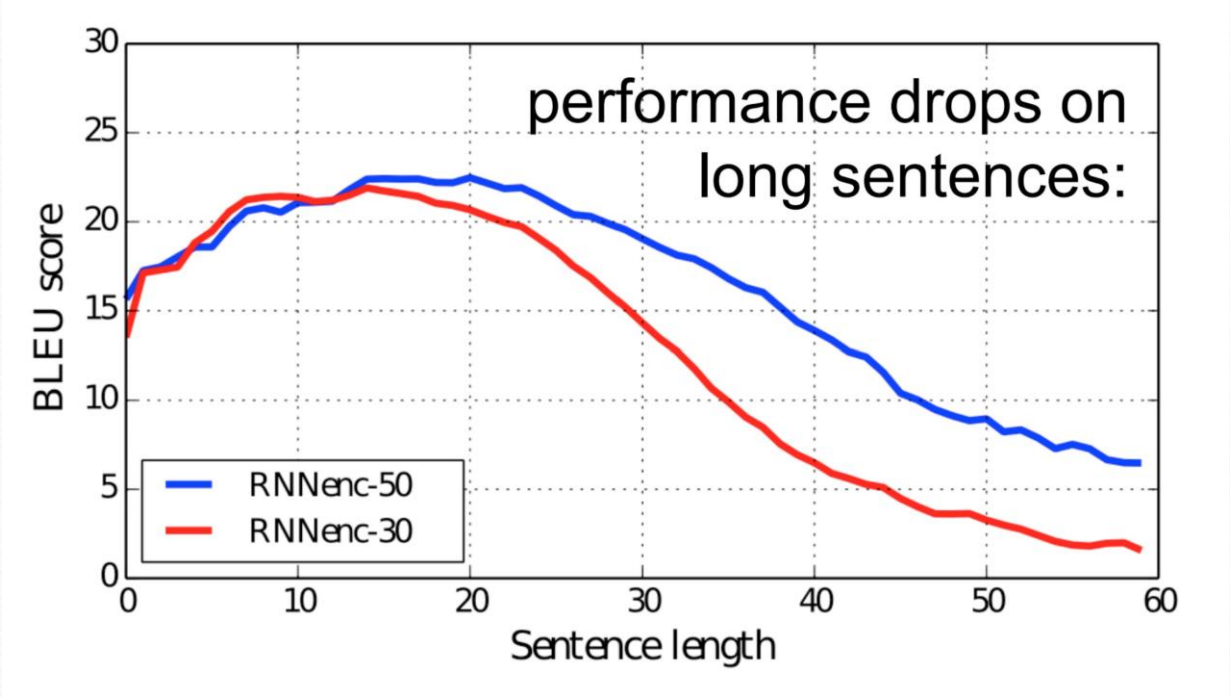
$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$

$$y_t = W_{hy}h_t$$

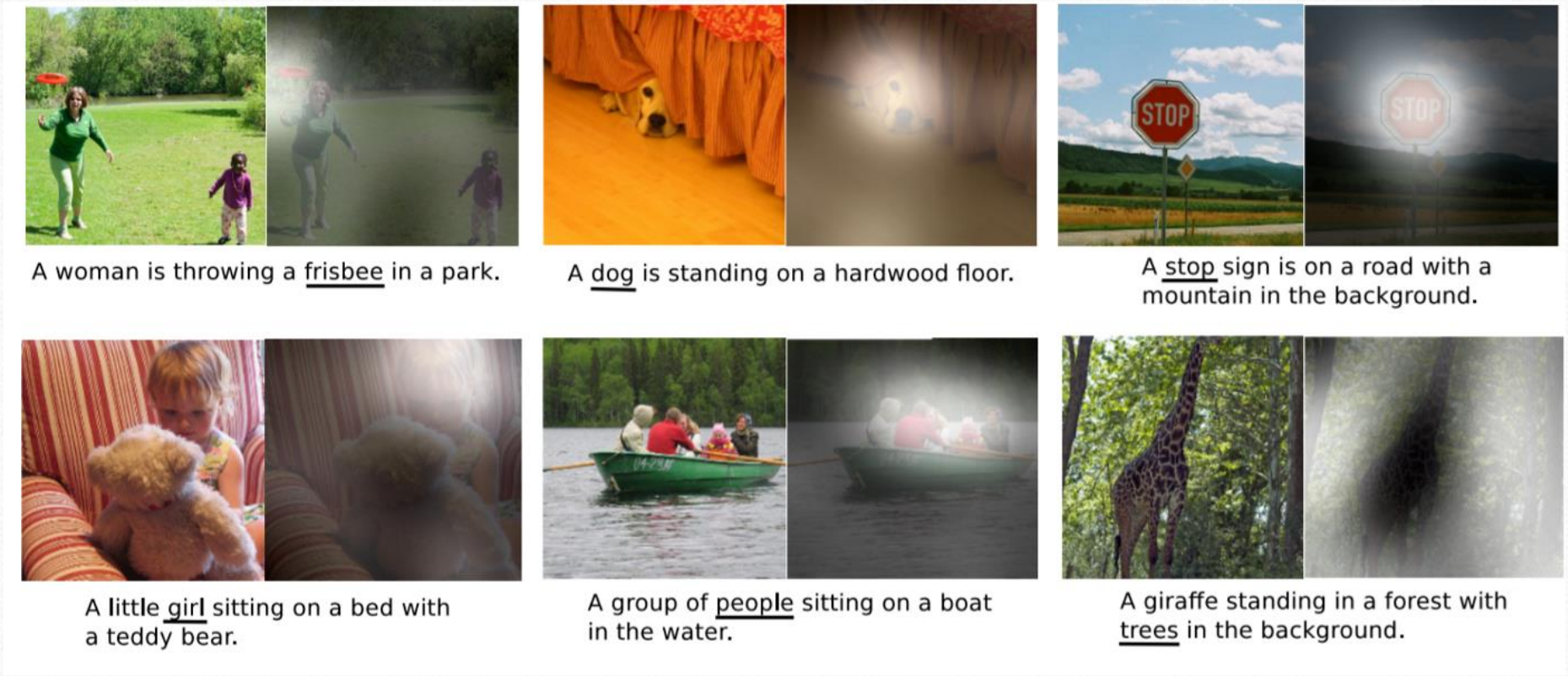
LSTM


Forget Gate
h_{t-1}
h_t
h_{t+1}
A
A
σ
σ
tanh
σ
tanh
X_{t-1}
X_t
X_{t+1}
Input Gate
Output Gate

Weakness RNN

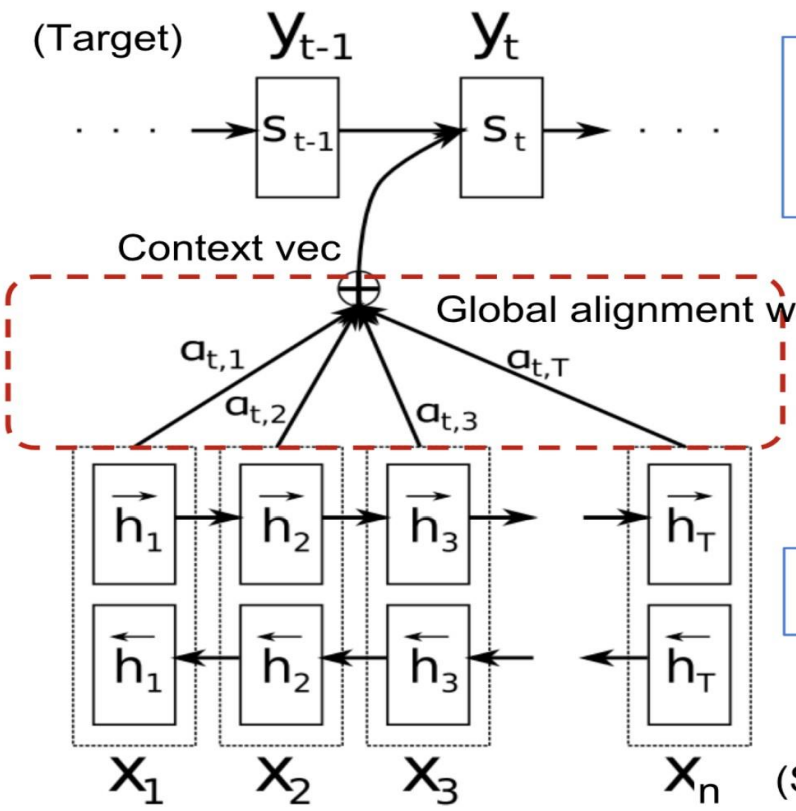


Attention Mechanism



In psychology, attention is the cognitive process of selectively concentrating on one or a few things while ignoring others

Attention + RNN



Decoder: RNN with input from previous state + dynamic context vector.

Attention layer: parameterized by a simple feed-forward network

Encoder: bidirectional RNN

$$\mathbf{c}_t = \sum_{i=1}^{n} \alpha_{t,i} \boldsymbol{h}_i$$

$$\alpha_{t,i} = \text{align}(y_t, x_i) = \frac{exp\left(\text{score}(\boldsymbol{s}_{t-1}, \boldsymbol{h}_i)\right)}{\sum_{j=1}^{n} \exp\left(\text{score}(\boldsymbol{s}_{t-1}, \boldsymbol{h}_j)\right)}$$

Instead of building a single context vector out of the encoder's last hidden state, the goal of attention is to create shortcuts between the context vector and the entire source input (Bahdanau, Cho, and Bengio 2015).

Self Attention

1. Each word is key, query and value.
2. Each word w is represented by a vector $\mathbf{x} \in \mathbb{R}^d$ by using an embedding method.
3. Calculate query ($\mathbf{q} \in \mathbb{R}^p$) for $\mathbf{x} \in \mathbb{R}^d$, which is projection of $\mathbf{x}$ to a new space.

$$\mathbf{q} = \mathbf{w}_q^\top \mathbf{x}.$$

4. Calculate key ($\mathbf{k} \in \mathbb{R}^p$) for $\mathbf{x} \in \mathbb{R}^d$, which is projection of $\mathbf{x}$ to a new space.

$$\mathbf{k} = \mathbf{w}_k^\top \mathbf{x}.$$

5. Calculate value ($\mathbf{w} \in \mathbb{R}^p$) for $\mathbf{x} \in \mathbb{R}^d$, which is projection of $\mathbf{x}$ to a new space.

$$\mathbf{v} = \mathbf{w}_v^\top \mathbf{x}.$$

6. A single word $\mathbf{x}$ has three different representations. Sometimes, we look at this word as query, sometimes as key, and sometimes as value.
7. The self-attention means that looking a word as query and compute the similarity of the query with all of the words seen as key.
8. Then use the softmax for computing the weights and compute the weighted average all of the words seen as value.
9. This computes the attention vector.

Self Attention

1. Consider the following sentence.

2. Calculating the attention for word apple.
3. Taking the inner product of the query vector of apple to the key vector of the previous words.

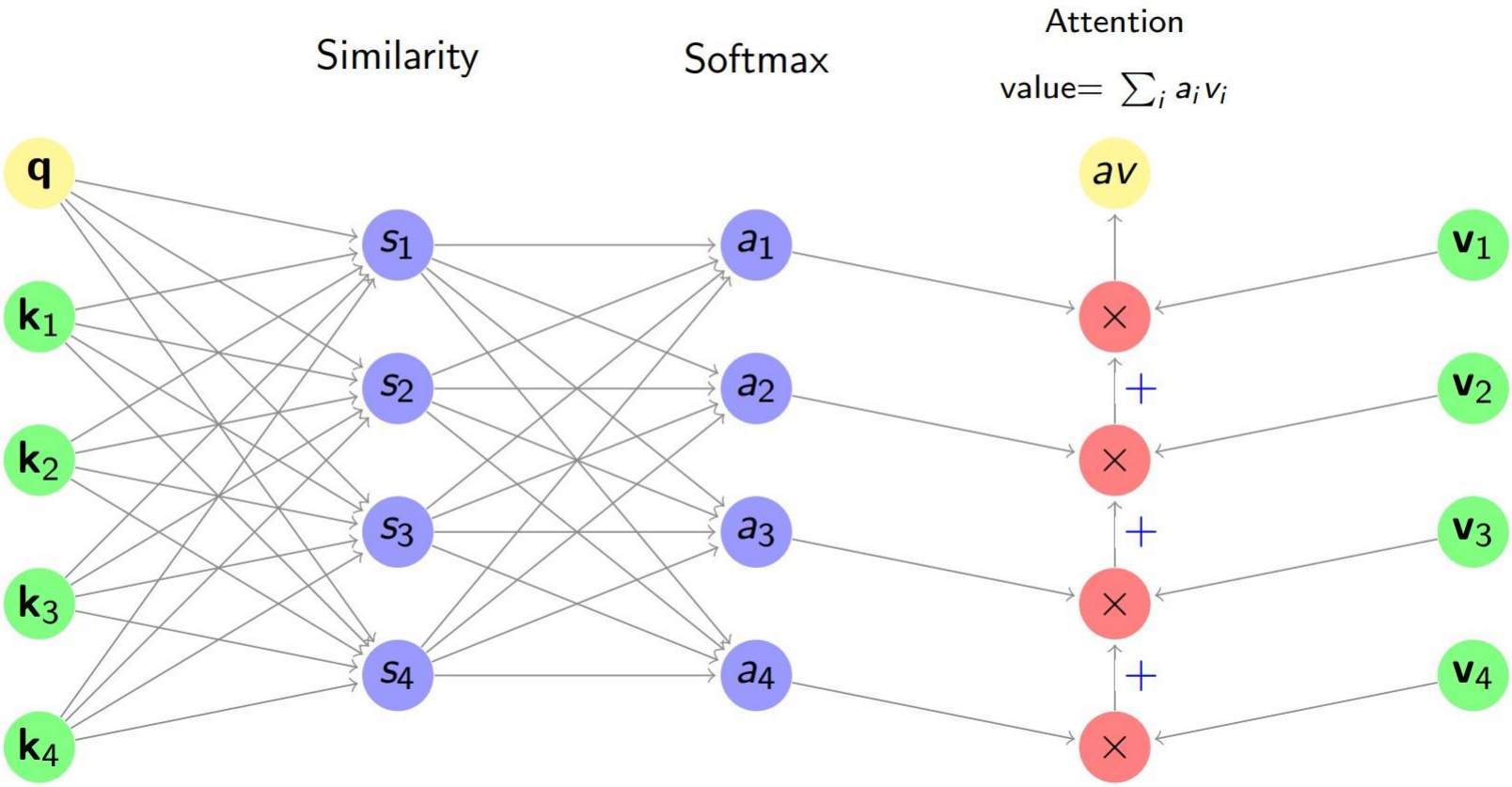
$$\mathbf{a} = softmax\left(\mathbf{q}_{apple}^{\top}\mathbf{k}_{she}, \mathbf{q}_{apple}^{\top}\mathbf{k}_{is}, \mathbf{q}_{apple}^{\top}\mathbf{k}_{eating}, \mathbf{q}_{apple}^{\top}\mathbf{k}_{a}, \mathbf{q}_{apple}^{\top}\mathbf{k}_{green}\right)$$

4. Suppose that we obtain $\mathbf{a} = (0.1, 0.1, 0.5, 0.1, 0.2)$. Then we obtain

$$\mathbf{v}_{apple} = 0.1\mathbf{v}_{she} + 0.1\mathbf{v}_{is} + 0.5\mathbf{v}_{eating} + 0.1\mathbf{v}_{a} + 0.2\mathbf{v}_{green}$$

Attention is all you need (Vaswani et al., 2017)

1. Self-attention uses the following neural network architecture.



Attention Example

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

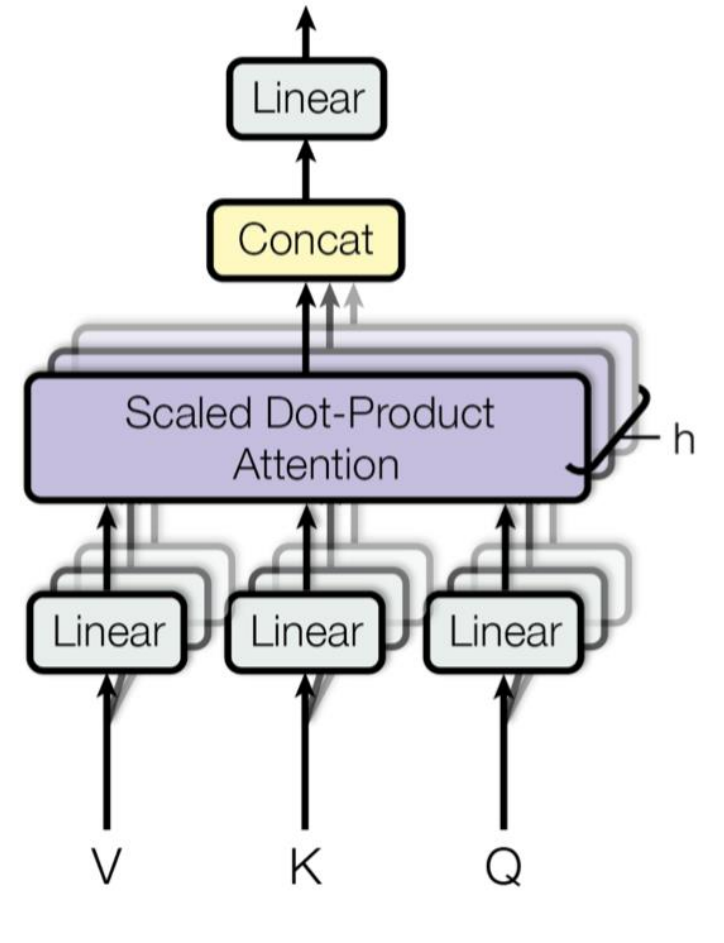
The FBI is chasing a criminal on the run .

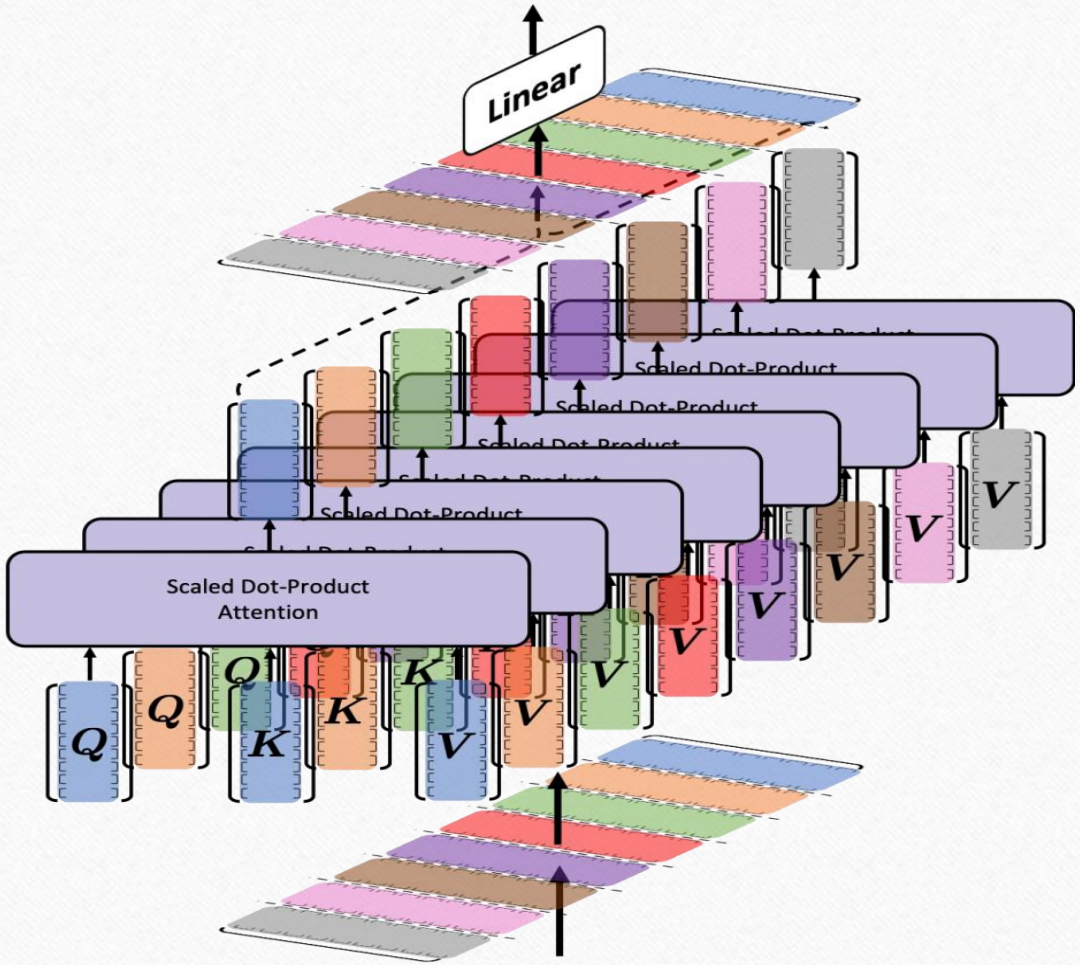
The FBI is chasing a criminal on the run .

Multi-head Attention

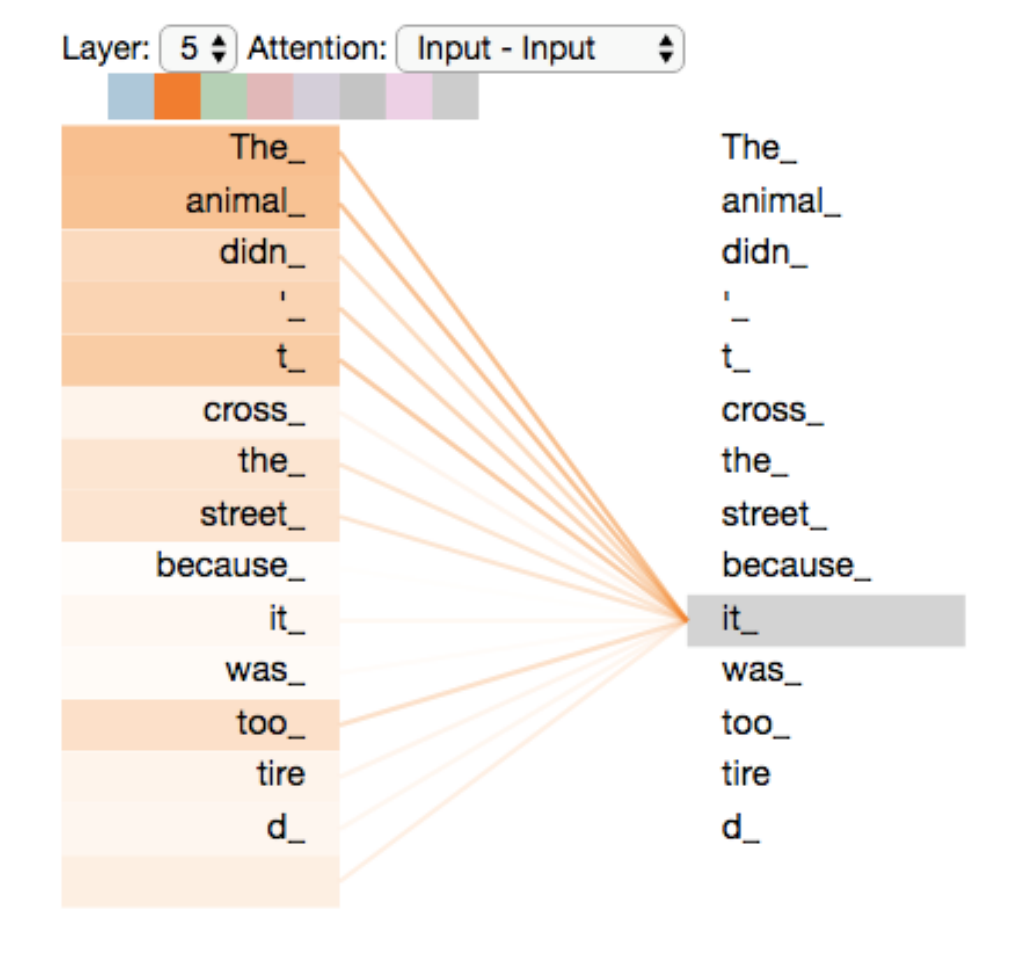
1. By defining three different vectors corresponding to each word.
   - Key $\mathbf{k} \in \mathbb{R}^p$ and $\mathbf{k} = \mathbf{W}_k^\top \mathbf{x}$, where $\mathbf{W}_k \in \mathbb{R}^{d \times p}$ and $\mathbf{x} \in \mathbb{R}^d$.
   - Query $\mathbf{q} \in \mathbb{R}^p$ and $\mathbf{q} = \mathbf{W}_q^\top \mathbf{x}$, where $\mathbf{W}_q \in \mathbb{R}^{d \times p}$ and $\mathbf{x} \in \mathbb{R}^d$.
   - Value $\mathbf{v} \in \mathbb{R}^p$ and $\mathbf{v} = \mathbf{W}_v^\top \mathbf{x}$, where $\mathbf{W}_v \in \mathbb{R}^{d \times p}$ and $\mathbf{x} \in \mathbb{R}^d$.
2. By defining the following matrices
   - $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$, where $\mathbf{X} \in \mathbb{R}^{d \times n}$.
   - $\mathbf{K} = [\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_n]$, where $\mathbf{K} \in \mathbb{R}^{p \times n}$.
   - $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n]$, where $\mathbf{Q} \in \mathbb{R}^{p \times n}$.
   - $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]$, where $\mathbf{V} \in \mathbb{R}^{p \times n}$.
3. Then, the new value $\mathbf{Z} \in \mathbb{R}^{p \times n}$ equals to $\mathbf{Z} = \mathbf{V}\ \mathit{Softmax}\left(\frac{\mathbf{Q}^\top \mathbf{K}}{\sqrt{p}}\right)$.

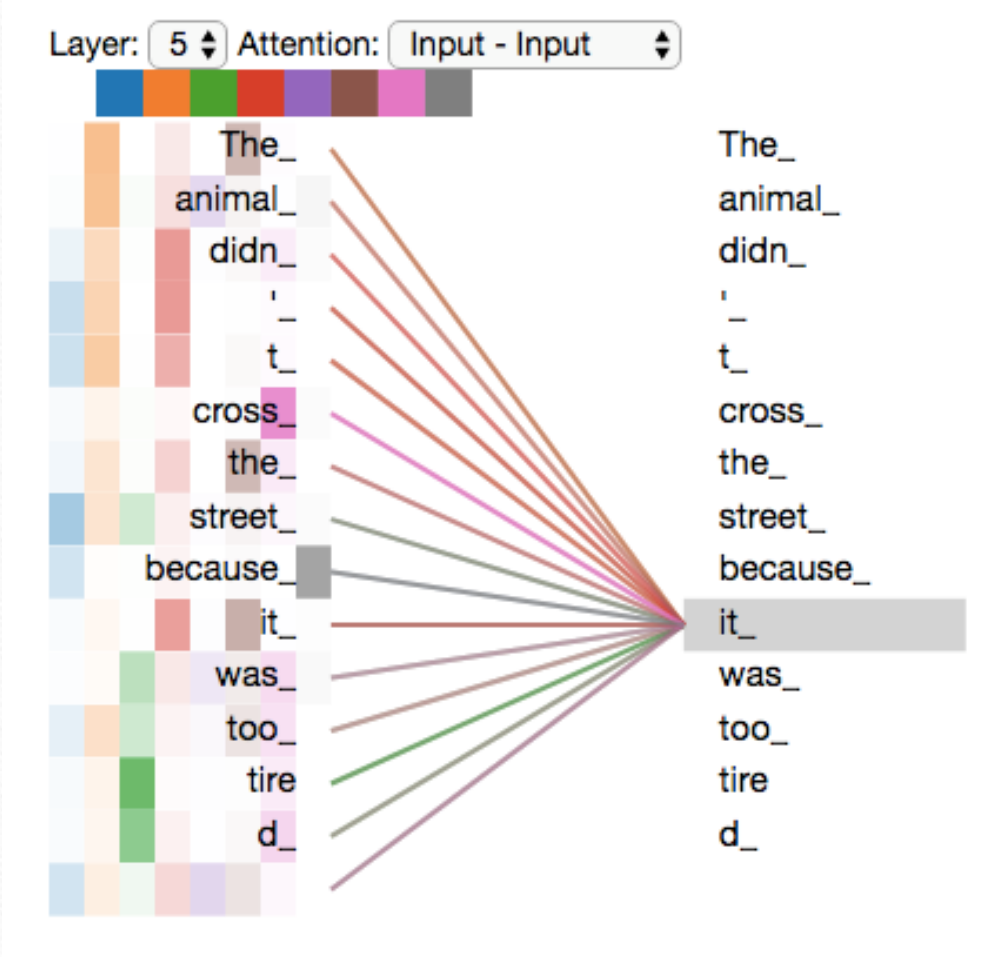
Multi-head Attention



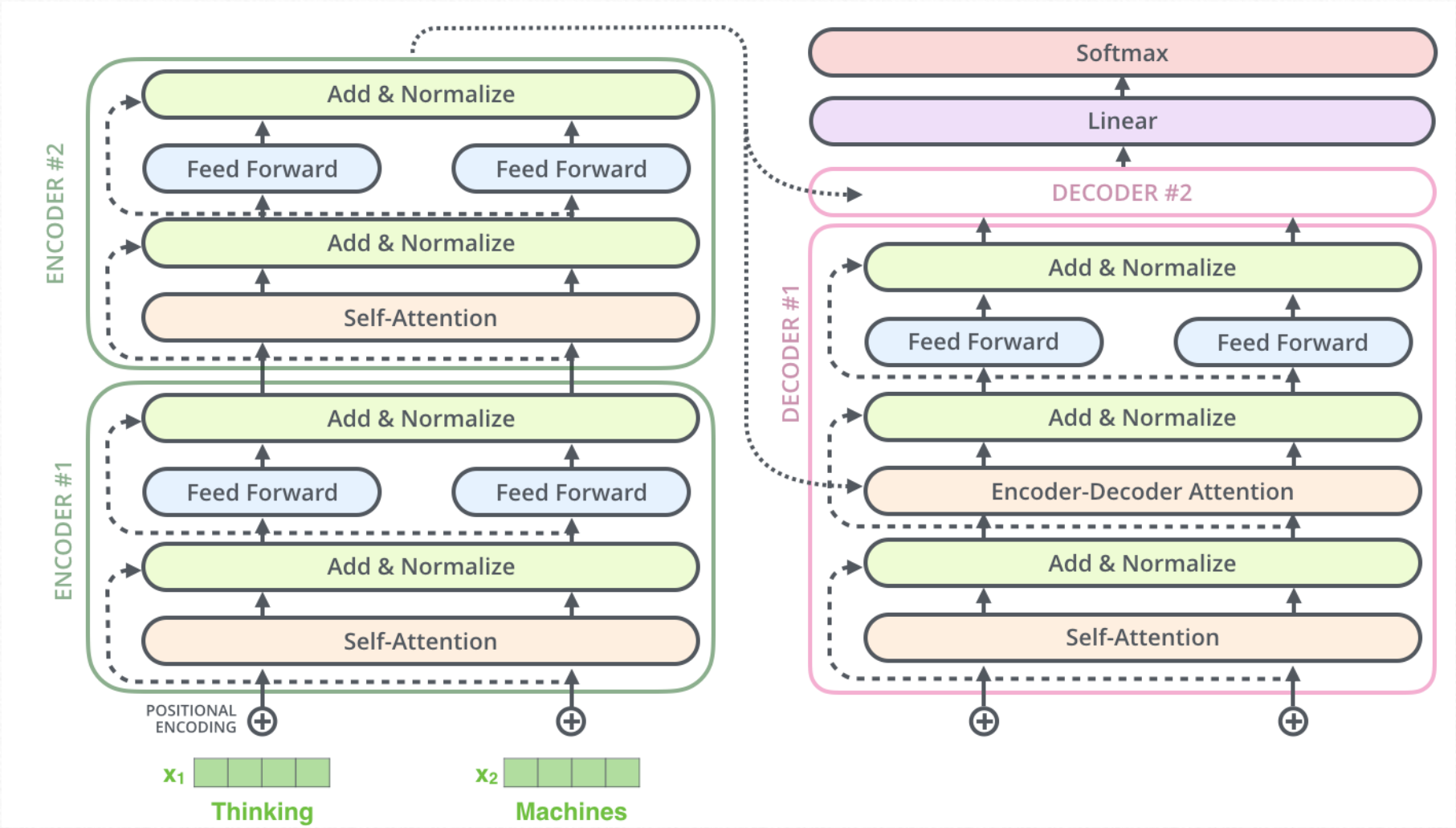


Single Attention V.S Multi-head Attention



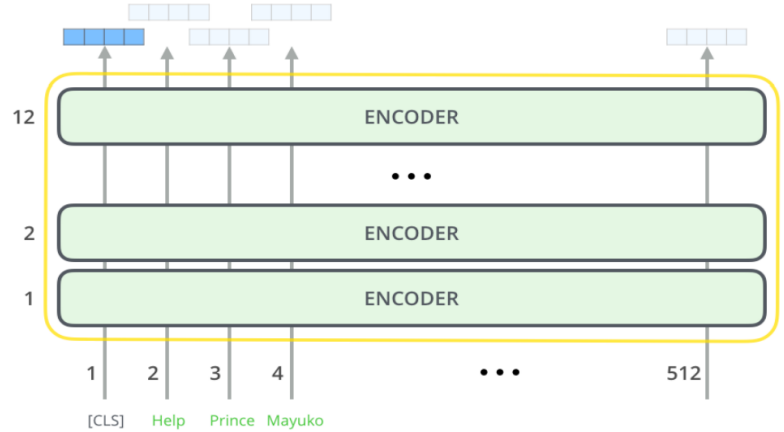


Transformer Block



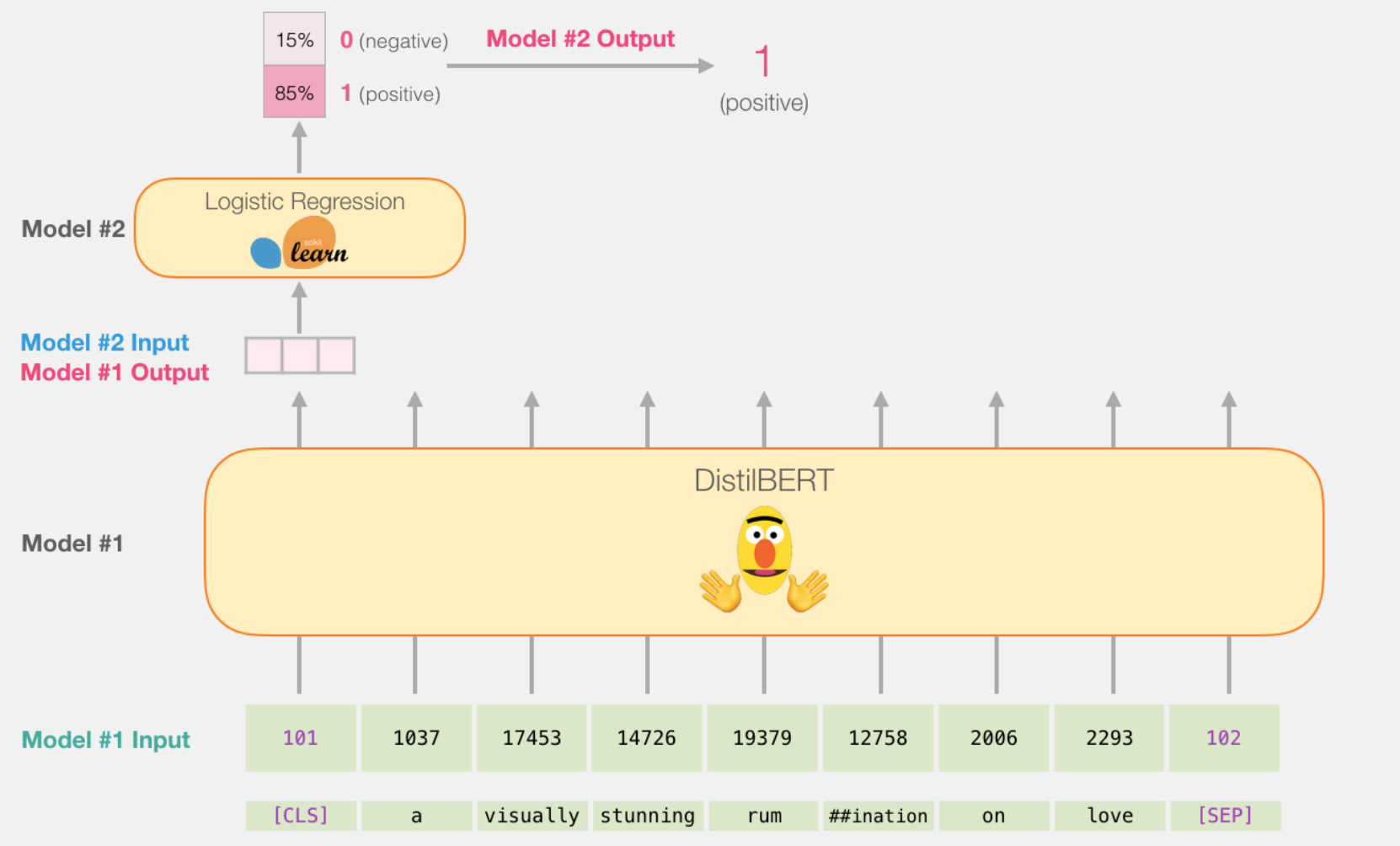
BERT for Language Model

1. BERT (Pre-training of Deep Bidirectional Transformers for Language Understanding) is basically a trained Transformers Encoder stack (Devlin et al. 2019).
2. Each position outputs a vector. For the sentence classification, we focus on the output of only the first position ([CLS]).
3. That vector can now be used as the input for a classifier. The paper achieves great results by just using a single-layer neural network as the classifier.



4. BERT makes use of a novel technique called Masked LM (MLM): it randomly masks words in the sentence and then it tries to predict them.

Classification with BERT Model



References

- Alain, Guillaume and Yoshua Bengio (2014). "What Regularized Auto-Encoders Learn from the Data-Generating Distribution". In: Journal of Machine Learning Research 15.110, pp. 3743–3773. url: http://jmlr.org/papers/v15/alain14a.html.
- Vaswani, Ashish et al. (2017). "Attention is All you Need". In: Advances in Neural Information Processing Systems, pp. 5998–6008.
- Devlin, Jacob et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: Proc. of Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186.