Document-Level Machine Translation with Hierarchical Attention

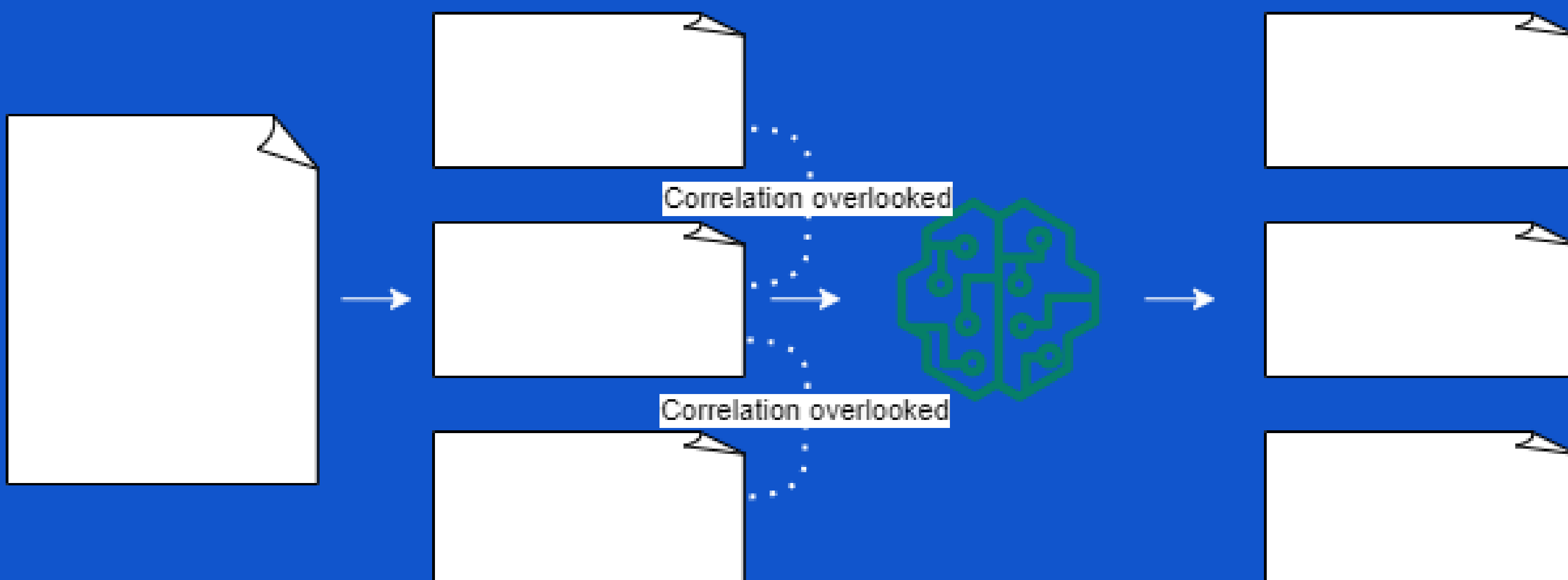
Master's Defense by Yu-Tang Shen
Advisor: Dr. Chris Pollett
Committee: Dr. Thomas Austin
Committee: Dr. William Andreopoulos

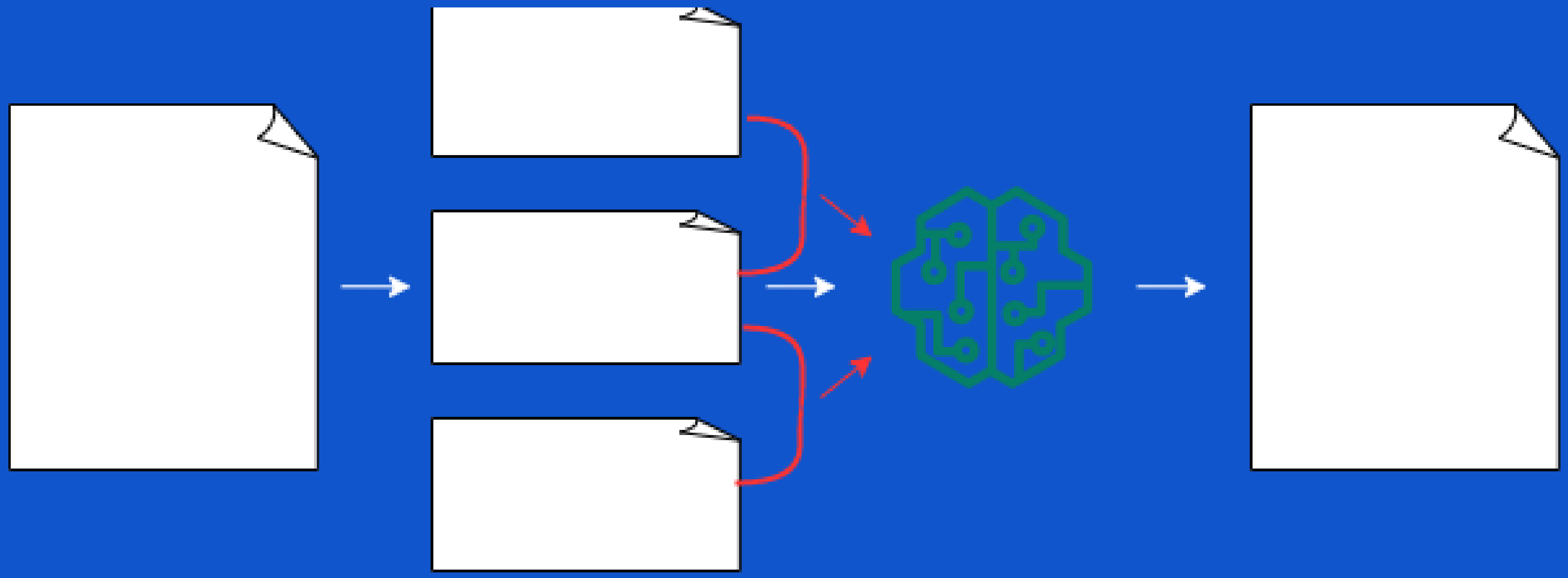
Agenda

- Project Goals
- Background
- Implementation
- Results
- Demo
- Future Work

Project Goals
Problem statement

Document translation requires too much computing power.





Provide correlations during the translation process.

Background

Terminologies

Machine translation	**MT**
Source language	SL
Target language	TL
Hierarchical attention	HAN

Background
Machine translation history

- Rule-based MT
- Statistical MT
- Neural MT
    - before the Transformer
    - the Transformer

Background

Rule-based machine translation

- Rules

Mary is	瑪莉是
the	那位
lady	女士
in a white dress	穿白裙子的
<Noun> <Postnominal Adj.>	<Postnominal Adj.> <Noun>

- Results

瑪莉是那位穿白裙子的女士

Background

Rule-based machine translation

Pros

- Simple algorithm to implement
- Deterministic results
  - Easy to debug

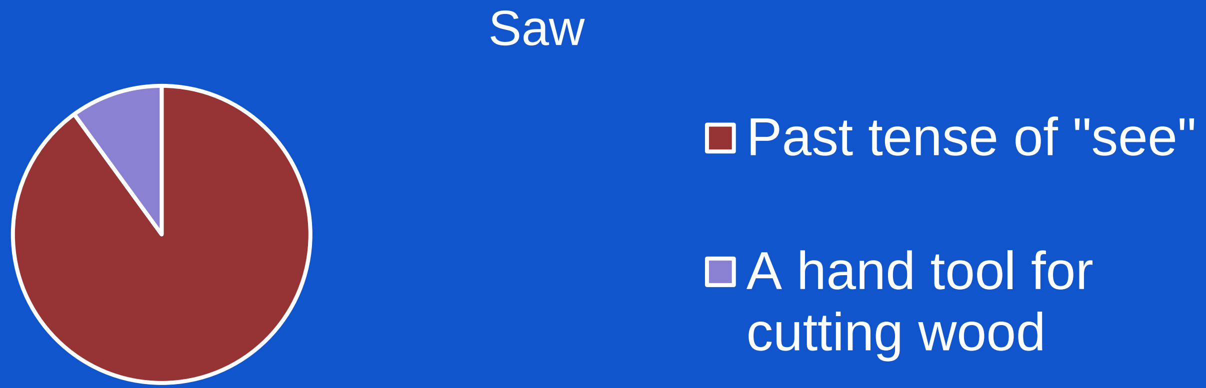
Cons

- Labor-intensive for listing out the rules
- Deterministic results
  - Can’t resolve lexical ambiguity
    - I saw bats

Background

Statistical machine translation

- Translates based on language statistics
- Flexible translations



Background
Statistical machine translation

- Language statistics
  - Frequency of SL *s* being translated TL *t*

saw	看到 (see)	90%
	鋸子 (tool)	10%

  - Frequency of TL t_1 following TL t_0

P(saw = 看到(see) | 我 (I))	90%
P(saw = 鋸子 (tool) | 我 (I))	10%

Background

Statistical machine translation

- Statistics

I	我	100%
saw	看到 (see)	90%
	鋸子 (tool)	10%
bats	蝙蝠 (animal)	50%
	球棒 (stick)	50%
P(saw = 看到(see) | 我)		90%
P(saw = 鋸子 (tool) | 我)		10%
P(bats =蝙蝠 (animal) | 看到 (see))		55%
P(bats =球棒 (stick) | 看到 (see))		45%

P(saw = 看到(see) | band)	10%
P(saw = 鋸子 (tool) | band)	90%
P(saw = 鋸子 (tool) | a)	99%
…	…

Background

Statistical machine translation

- Statistics

I	我	100%
saw	看到 (see)	90%
	鋸子 (tool)	10%
bats	蝙蝠 (animal)	50%
	球棒 (stick)	50%
P(saw = 看到(see) | 我)		90%
P(saw = 鋸子 (tool) | 我)		10%
P(bats =蝙蝠 (animal) | 看到 (see))		55%
P(bats =球棒 (stick) | 看到 (see))		45%

- argmax_x P (I = x)
  - 我

Background

Statistical machine translation

- Statistics

I	我	100%
saw	看到 (see)	90%
	鋸子 (tool)	10%
bats	蝙蝠 (animal)	50%
	球棒 (stick)	50%
P(saw = 看到(see) | 我)		90%
P(saw = 鋸子 (tool) | 我)		10%
P(bats =蝙蝠 (animal) | 看到 (see))		55%
P(bats =球棒 (stick) | 看到 (see))		45%

- argmax_x P (I = x)
  - 我
- argmax_x P (saw = x | 我)

Background

Statistical machine translation

- Statistics

I	我	100%
saw	看到 (see)	90%
	鋸子 (tool)	10%
bats	蝙蝠 (animal)	50%
	球棒 (stick)	50%
P(saw = 看到(see) | 我)		90%
P(saw = 鋸子 (tool) | 我)		10%
P(bats = 蝙蝠 (animal) | 看到 (see))		55%
P(bats = 球棒 (stick) | 看到 (see))		45%

- argmax_x P (I = x)
  - 我
- argmax_x P (saw = x | 我)
  - 看到 (see)

Background

Statistical machine translation

- Statistics

I	我	100%
saw	看到 (see)	90%
	鋸子 (tool)	10%
bats	蝙蝠 (animal)	50%
	球棒 (stick)	50%
P(saw = 看到(see) | 我)		90%
P(saw = 鋸子 (tool) | 我)		10%
P(bats =蝙蝠 (animal) | 看到 (see))		55%
P(bats =球棒 (stick) | 看到 (see))		45%

- argmax$_x$ P (I = x)
  - 我
- argmax$_x$ P (saw = x | 我)
  - 看到 (see)
- argmax$_x$ P (bats = x | 看到 (see))

Background

Statistical machine translation

- Statistics

I	我	100%
saw	看到 (see)	90%
	鋸子 (tool)	10%
bats	蝙蝠 (animal)	50%
	球棒 (stick)	50%
P(saw = 看到(see) | 我)		90%
P(saw = 鋸子 (tool) | 我)		10%
P(bats = 蝙蝠 (animal) | 看到 (see))		55%
P(bats = 球棒 (stick) | 看到 (see))		45%

- argmax_x P (I = x)
  - 我
- argmax_x P (saw = x | 我)
  - 看到 (see)
- argmax_x P (bats = x | 看到 (see))
  - 蝙蝠 (animal)
- Final result:

我看到 (see)蝙蝠 (animal)

Background

Statistical machine translation

Pros

- Flexible translations
  - Better readability[1]

Cons

- Doesn't analyze / interpret the context thoroughly
  - The band saw some audience leaving the concert early.

P(saw = 看到(see) | band)	10%
P(saw = 鋸子 (tool) | band)	90%

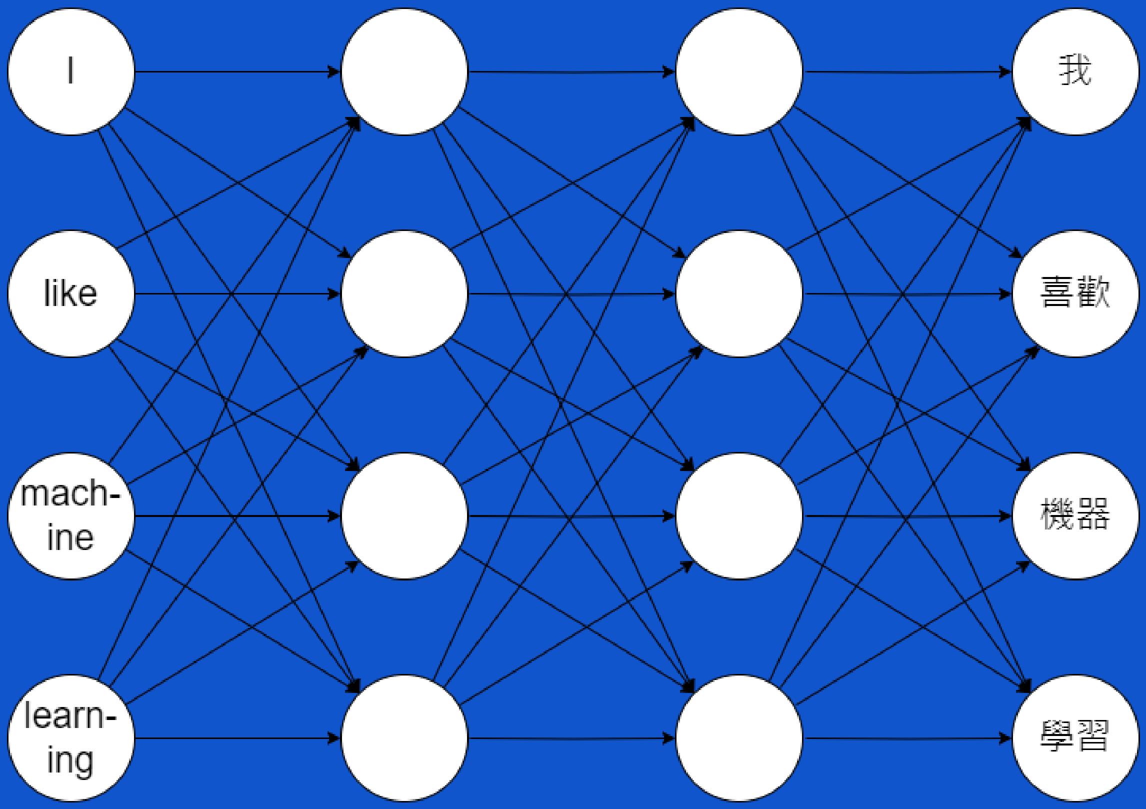
[1] As seen in CS 297 experiments at http://www.cs.sjsu.edu/faculty/pollett/masters/Semesters/Fall22/thomas/index.php?297Deliverable2.php

Background
Neural machine translation

- Neural networks are good at resolving complex correlations
  - Translations are complicated matching

Background

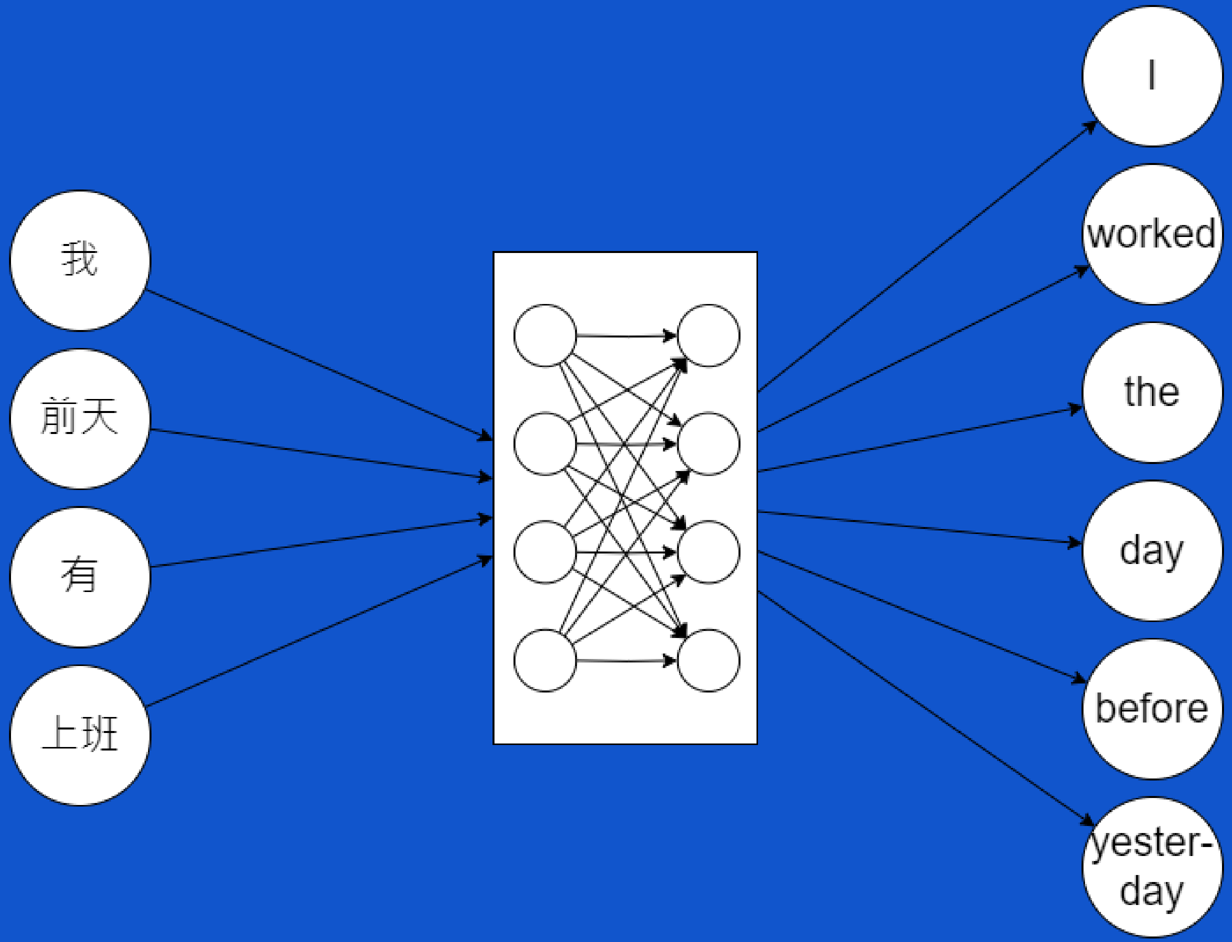
Neural machine translation



Background

Neural machine translation

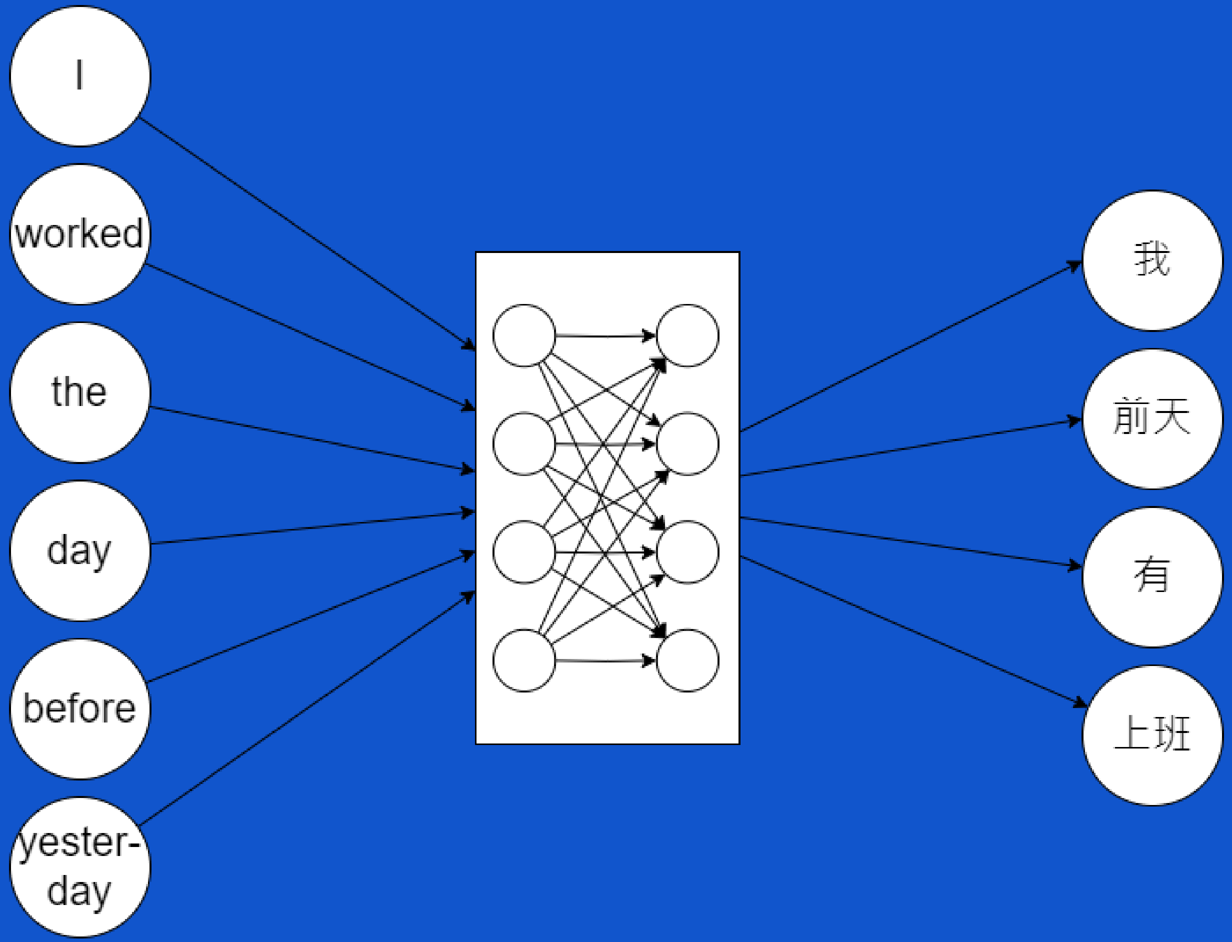
Encoder-decoder



Background

Neural machine translation

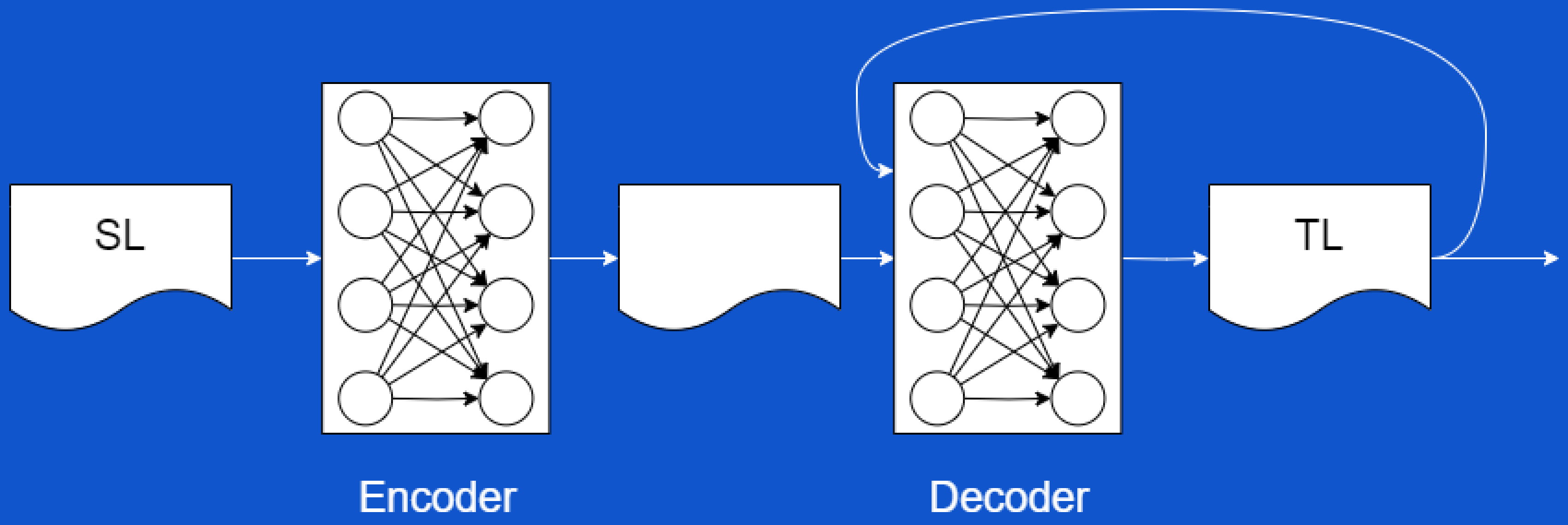
Encoder-decoder



Background

Neural machine translation

Encoder-decoder



Background

Neural machine translation

Interlingua

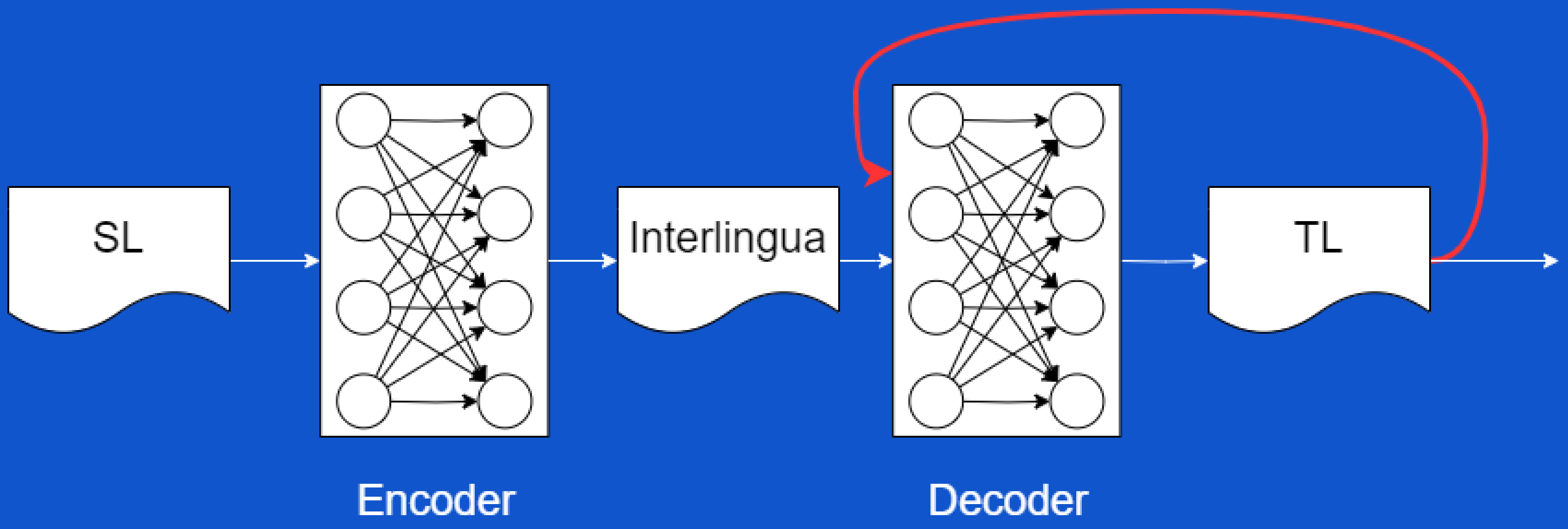
- Direct translation
  - SL => TL
- Indirect translation / transfer translation
  - SL => Interlingua => TL

Background

Neural machine translation

Encoder-decoder

- Indirect translation / transfer translation
  - SL => Interlingua => TL



Background
Neural machine translation
Encoder-decoder

- Encoder / decoder options
  - RNN
  - LSTM

Background

Neural machine translation

The Transformer [1]

- Each token attends with others
  - α(“surgeon”, “surgery”) > α(“engineer”, “surgery”)
- What is important to translate this token

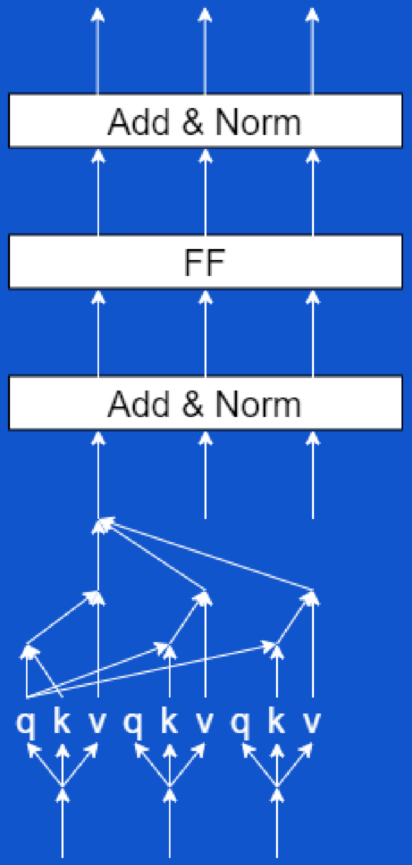
[1] A. Vaswani, et. al., "Attention is all you need," Advances in neural information processing systems, 2017.

Background

Neural machine translation

The Transformer

- Pros
  - Avoid gradient vanishing / exploding
  - More parallel => efficient

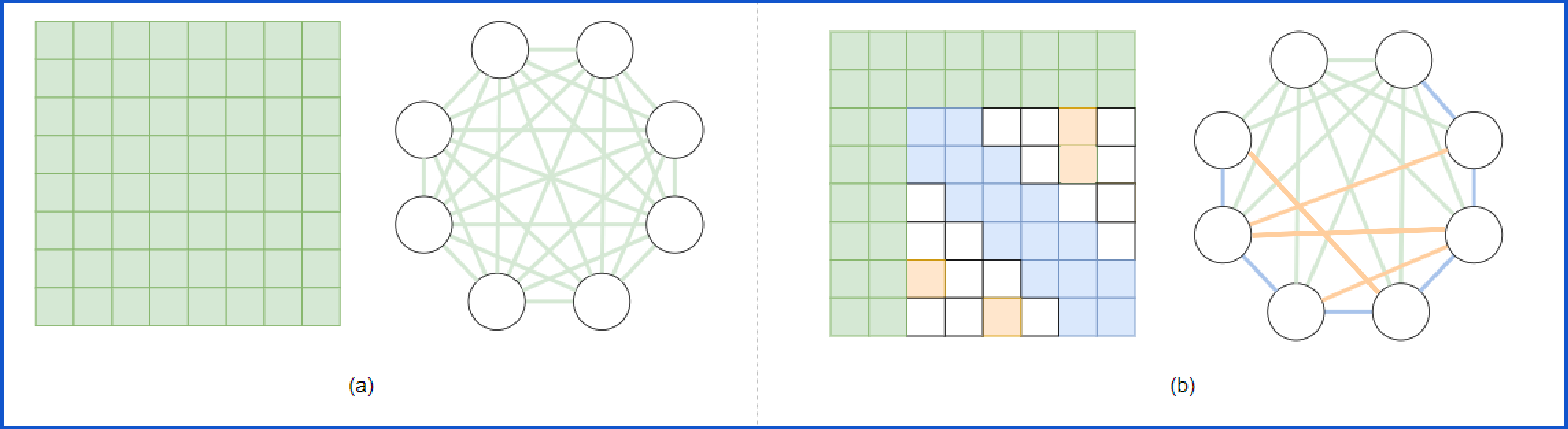


Background

Neural machine translation

Big Bird [2] Attention mechanism

- Attend to the tokens that are more likely to be relevant



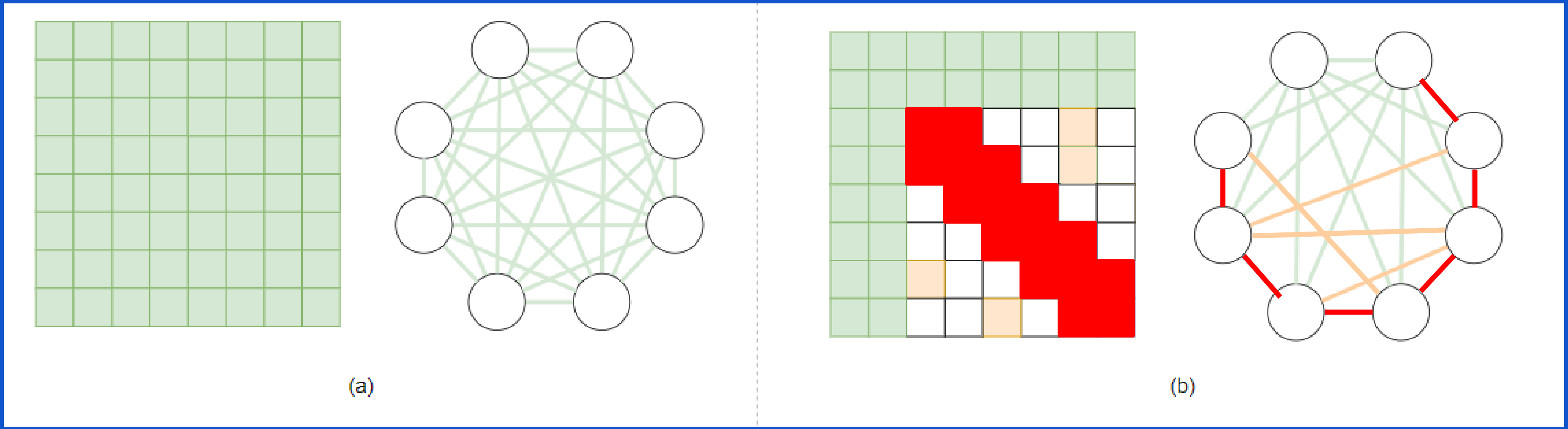
[2] Z. Manzil, et.al., "Big bird: Transformers for longer sequences," Adv. Neural Inf. Process. Syst., 2020.

Background

Neural machine translation

Big Bird Attention mechanism

- Attend to the tokens that are more likely to be relevant

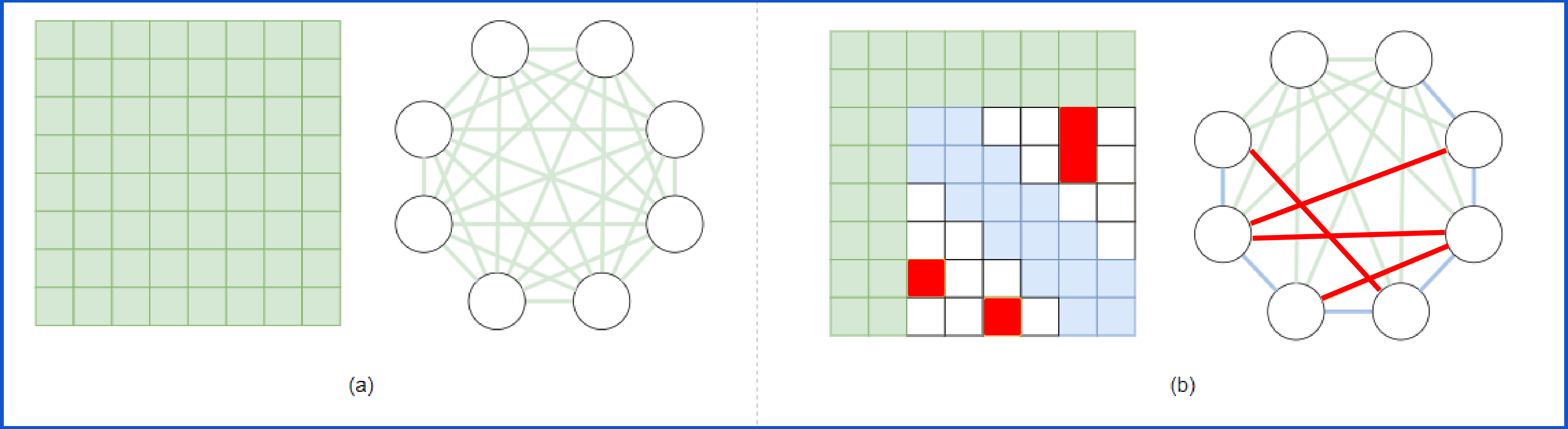


Background

Neural machine translation

Big Bird Attention mechanism

- Attend to the tokens that are more likely to be relevant

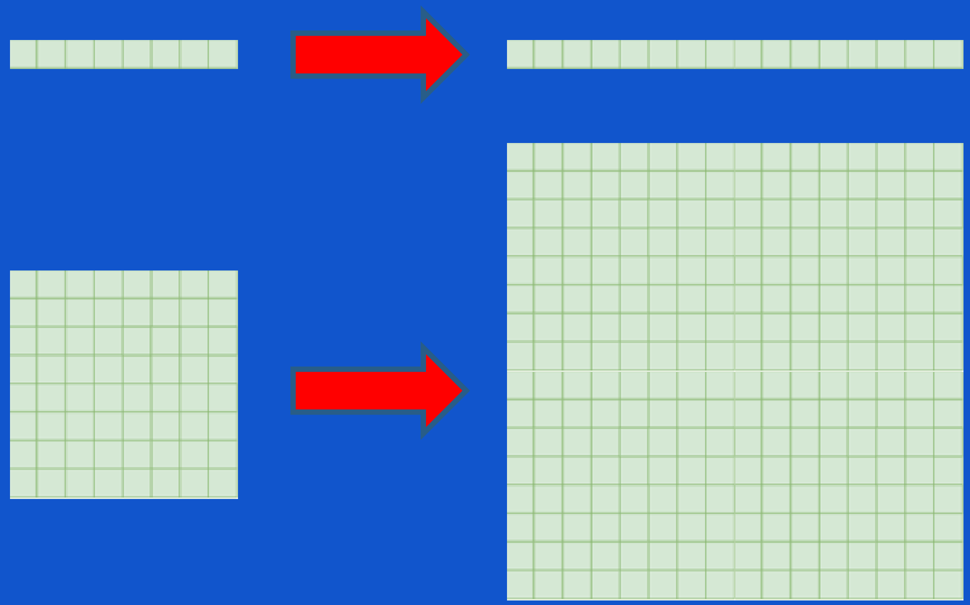


Background

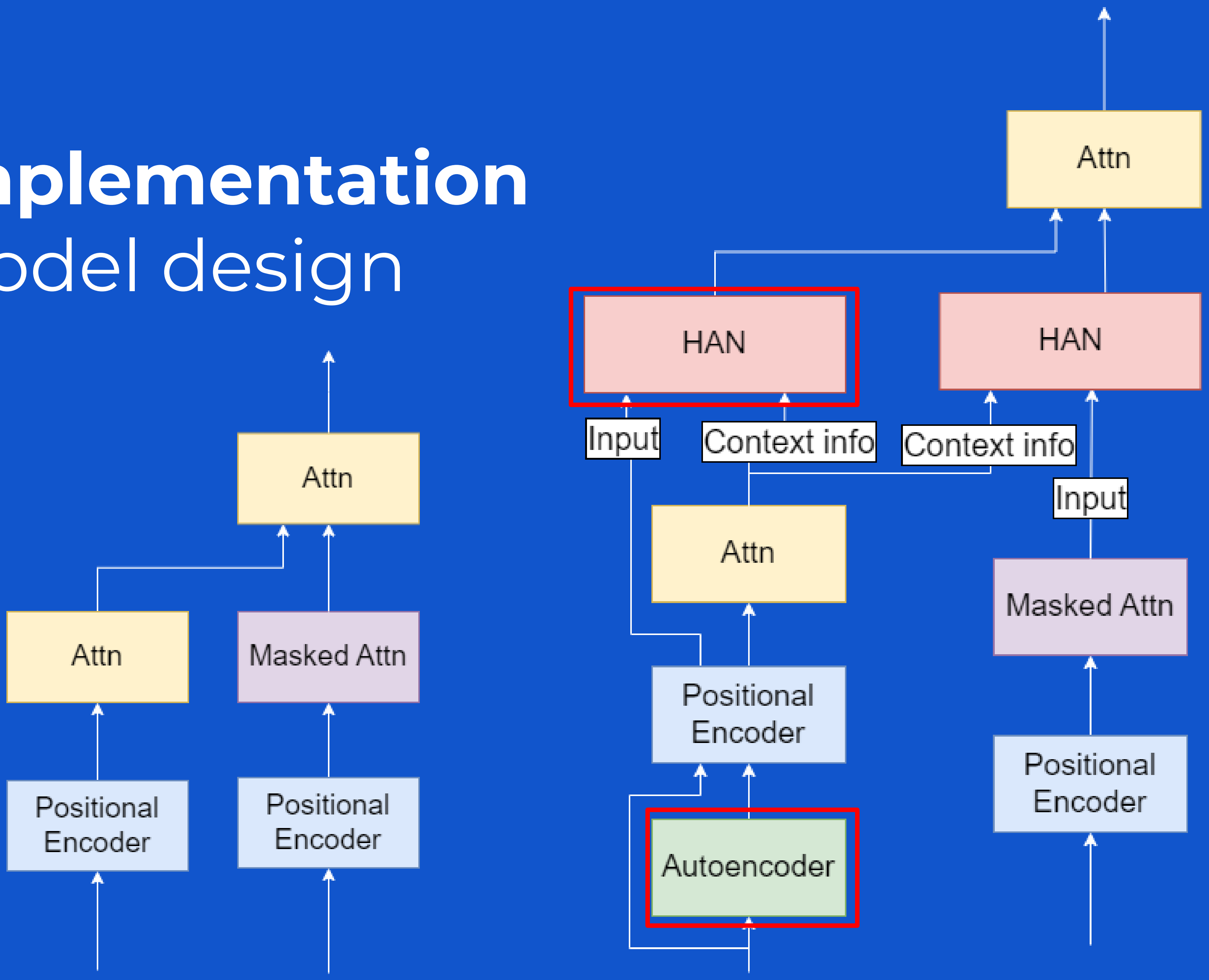
Neural machine translation

Flaw

- **Encoder input length limit**



Implementation
Model design

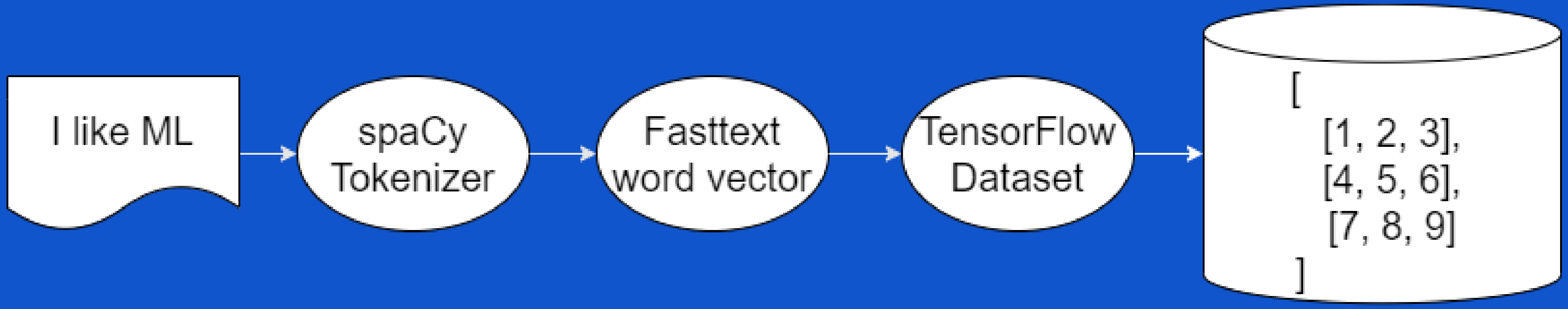


Implementation
Overview

- Preprocessing
  - K dimensional tree (k-d tree)
  - The digit issue
- New layers
  - Autoencoder
  - Hierarchical attention (HAN)
    - Big Bird attention
- New model metric

Implementation
Preprocessing

- Store data as vectors instead of words
  - spaCy tokenization
  - Fasttext word vectors
  - Store into TensorFlow Dataset



Implementation

Preprocessing

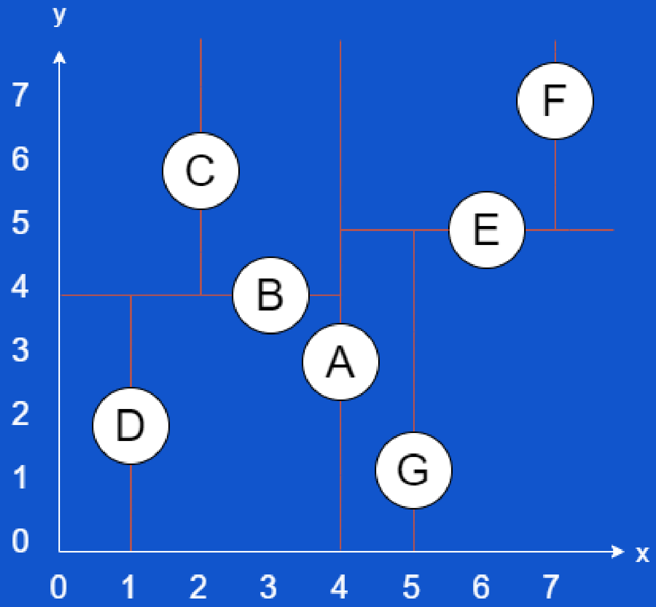
K-d tree

- Decipher word vector outputs
- K-d tree is a fast algorithm ($O(\log n)$) for spatial search

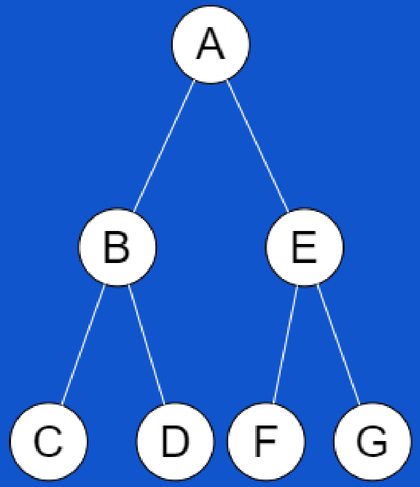
Implementation

Preprocessing

K-d tree



(a)



(b)

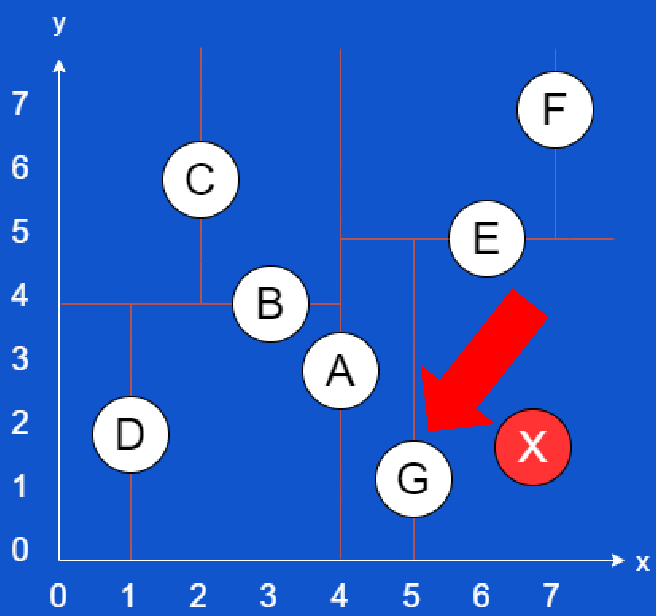
Dictionary	
Key	**Value**
[1, 2]	D
[2, 6]	C
[3, 4]	B
[4, 3]	A
[5, 1]	G
[6, 5]	E
[7, 7]	F

(c)

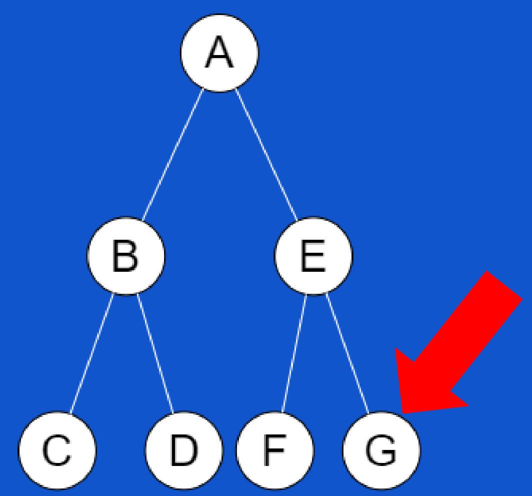
Implementation

Preprocessing

K-d tree



(a)



(b)

Dictionary	
Key	**Value**
[1, 2]	D
[2, 6]	C
[3, 4]	B
[4, 3]	A
[5, 1]	G
[6, 5]	E
[7, 7]	F

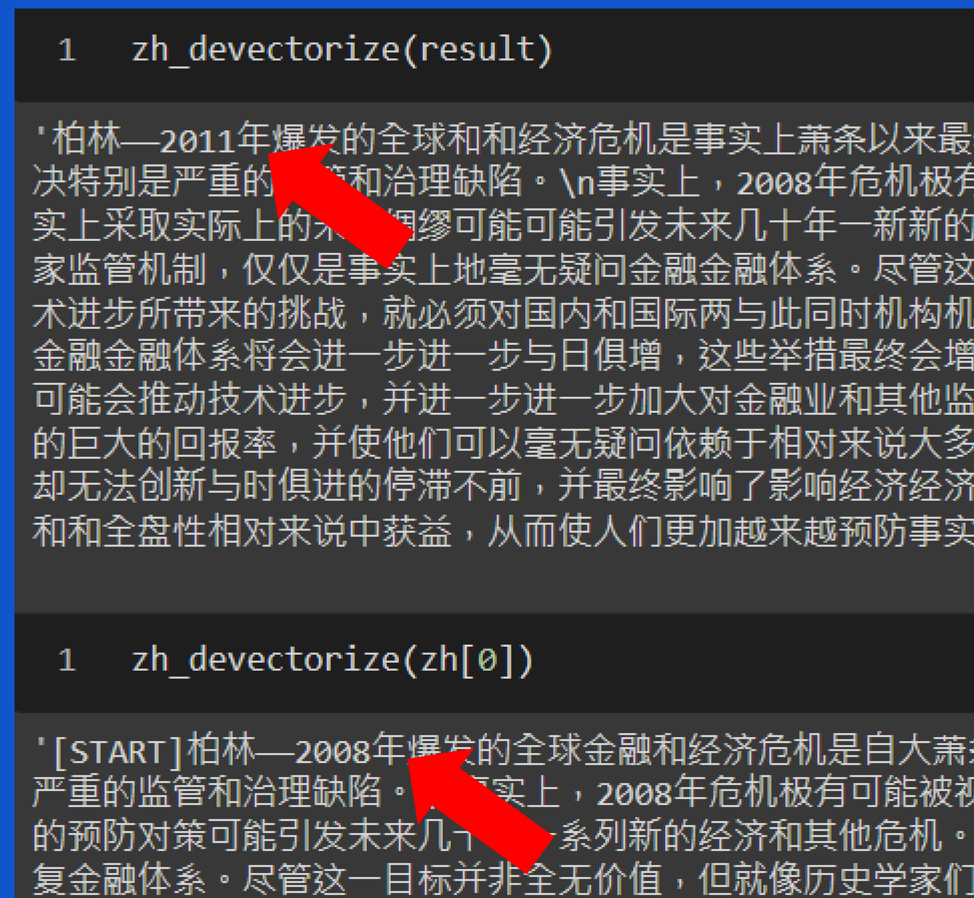
(c)

Implementation

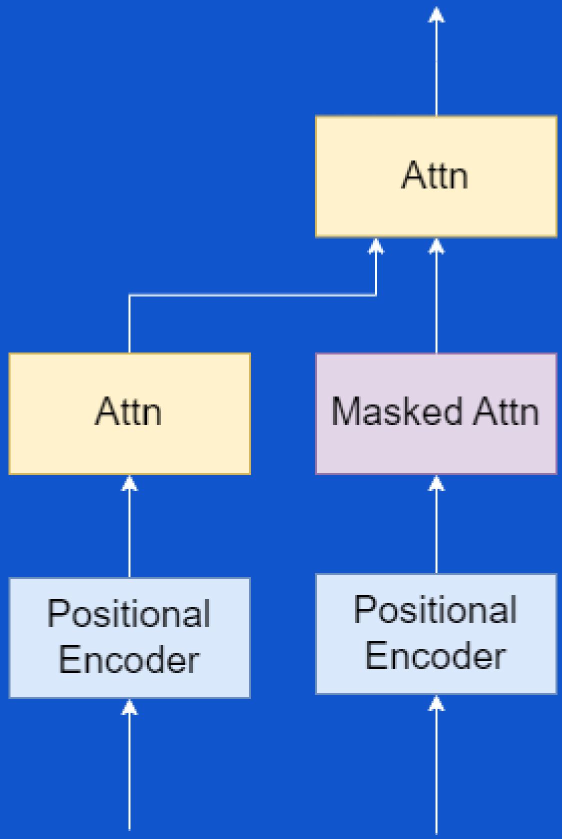
Preprocessing

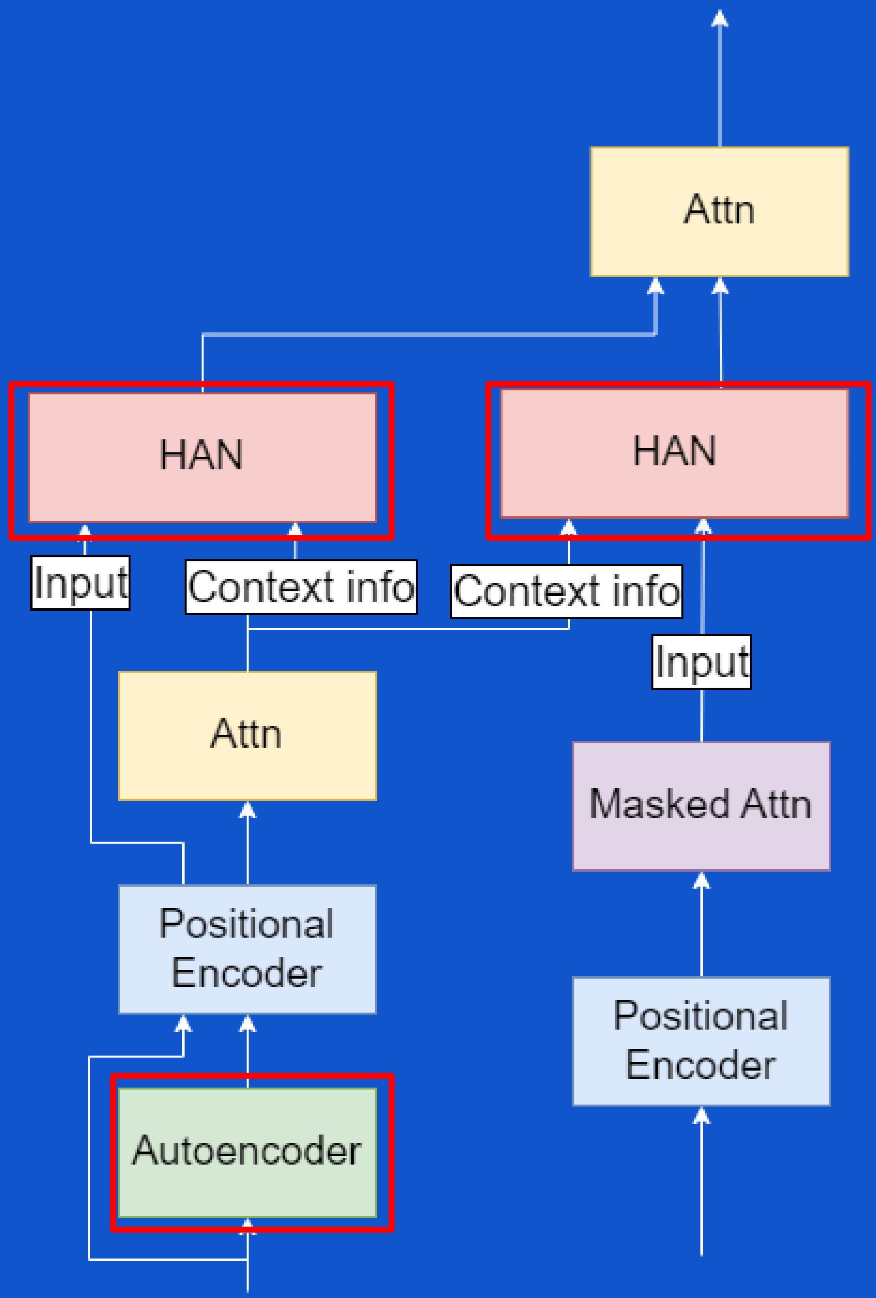
The digit issue

- Fail to translate digits
  - Guess
    - Occurrence of 2008 was too little
  - Solution
    - Split 2008 into 2, 0, 0, and 8



Implementation
Model design



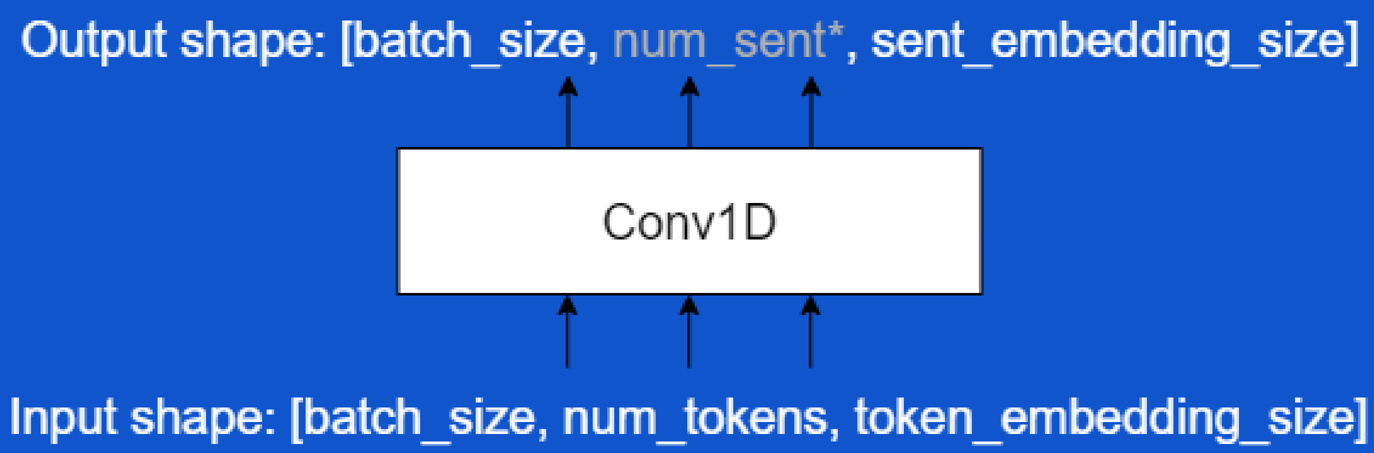


Implementation

New layers

Autoencoder

- Autoencoder summarizes the input on a sentence level* to get context information



Implementation

New layers

Autoencoder

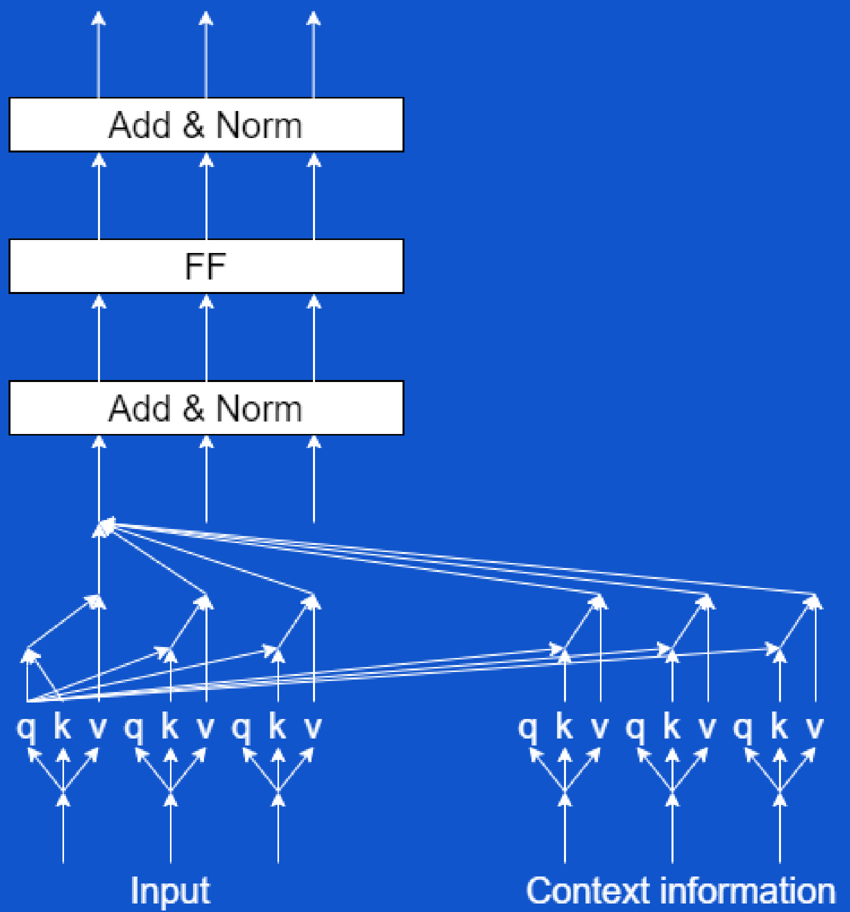
- Summarizing actual sentences is slow
- Set a fixed size window and stride to split input into sentences
    - Window: 16 (average sentence length in dataset = 14)
    - Stride: 8

Implementation

New layers

HAN

- Each token attends with
  - Other tokens, and
  - Context information nodes
- Information
  - At that position, and
  - From distant nodes



Implementation
New model metric

- Penalize by distance
  - Mean squared error as loss function
- Evaluate with hit or miss
  - Hit if a dimension of output vector is within a defined threshold
  - Miss if otherwise
  - Accuracy = #hits / total dimensions

Truth	Prediction	Absolute difference	Threshold
[[10, 10], [20, 20]]	[[11, 23], [19, 24]]	[[1, 13], [1, 4]]	5
Accuracy = 3 / 4			

Results
Overview

- English to Chinese translation
  - Full attention
  - Big Bird attention
- Chinese to English translation
  - Big Bird attention

Results

English to Chinese translation

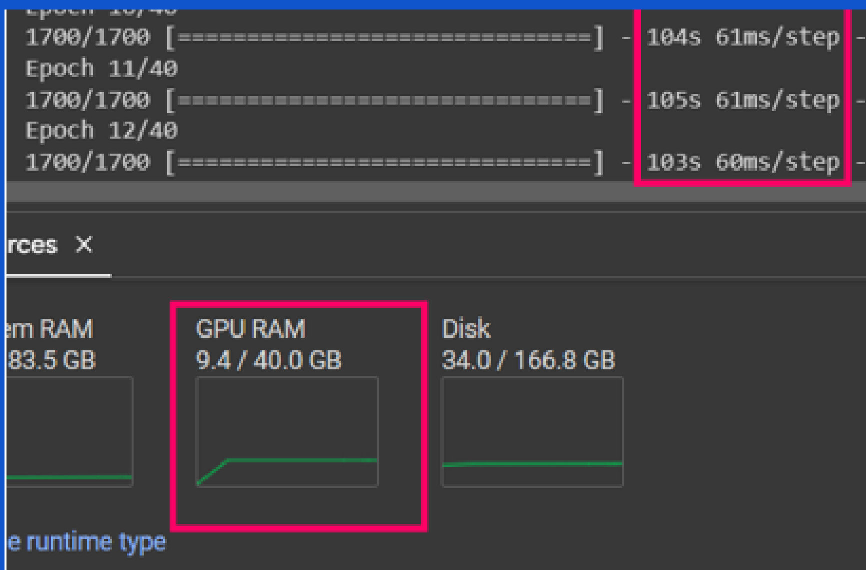
Full attention

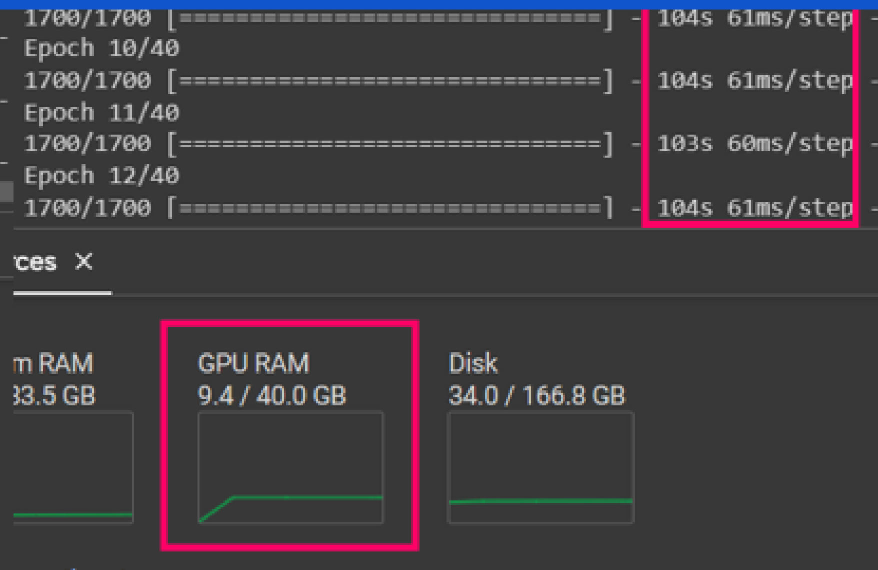
Input	Prediction	Ground truth
BERLIN – The global financial and economical crisis that began in 2008 was the greatest economic stress-test since the Great Depression, and the greatest challenge to social and political systems since World War II.	柏林——2008年的全球全球金融经济经济危机自大萧条以来最严峻的一次经济经济测试测试，是自自以来社会社会政治政治所面临的最严重挑战。	柏林——2008年爆发的全球金融和经济危机是自大萧条以来最严峻的一次经济压力测试，也是自二战以来社会和政治制度所面临的最严重挑战。

Results

English to Chinese translation

Big Bird attention





Full attention HAN

Big Bird HAN

Results

English to Chinese translation

Big Bird attention

Input	Prediction	Ground truth
BERLIN – The global financial and economical crisis that began in 2008 was the greatest economic stress-test since the Great Depression, and the greatest challenge to social and political systems since World War II.	柏林——2008年爆发的全球金融和经济危机是自大萧条以来最严峻的一次经济压力测试测试也是自二战以来社会和政治制度所面临的最严重挑战。	柏林——2008年爆发的全球金融和经济危机是自大萧条以来最严峻的一次经济压力测试，也是自二战以来社会和政治制度所面临的最严重挑战

Results

Chinese to English translation

Big Bird attention

- Model complexity reduced due to larger input space
  - Chinese documents are longer
  - Number of attention layers: 4 => 2
  - Number of attention heads: 8 => 4

Results

Chinese to English translation

Big Bird attention

Input	Prediction	Ground truth
柏林——2008年爆发的全球金融和经济危机是自大萧条以来最严峻的一次经济压力测试，也是自二战以来社会和政治制度所面临的最严重挑战。	Amsterdam – the global financial and economic crisis that began in 2 0 0 8 was the greatest economic stress - test since the great depression , and the greatest challenge to social and political systems since world war II .	BERLIN – The global financial and economical crisis that began in 2008 was the greatest economic stress-test since the Great Depression, and the greatest challenge to social and political systems since World War II.

Results
Comparison

Model	Training Configuration		Results		
	Attention layers	Attention heads	Validation Accuracy	BLEU score	Training cost (ms/step)
HAN	4	8	0.8	0.44	172
HAN-^{SD}	4	8	0.92	**0.9**	169
BB-HAN-$^{EN\_ZH}$	4	8	**0.96**	0.86	171
BB-HAN-$^{ZH\_EN}$	2	4	0.95	0.79	**81**

Results
Observation

Desirable feature

- Comprehendible texts

Bugs

- Stuttering effect
- Smaller models may cause incorrect translations

“

Demo

4 min

[ZH-EN][Split-digit][demo] BB HAN.ipynb

File Edit View Insert Runtime Tools Help All changes saved

```
VALIDATION_SIZE = 0.3
MAX_TOKENS = 4096
CHUNK_SIZE = 16 #  English sentence average sentence legth: 15~20 / Chinese sentence: 8~14
LATENT_SIZE = 300
BB_RANDOM_RATIO = 0.3
BATCH_SIZE = 4
THRESHOLD = 0.05
```

Show code

```
from google.colab import drive
drive.mount('/content/drive')
%cd drive/MyDrive/HAN
```

```
Mounted at /content/drive
```

56s completed at 3:45 AM

Future work

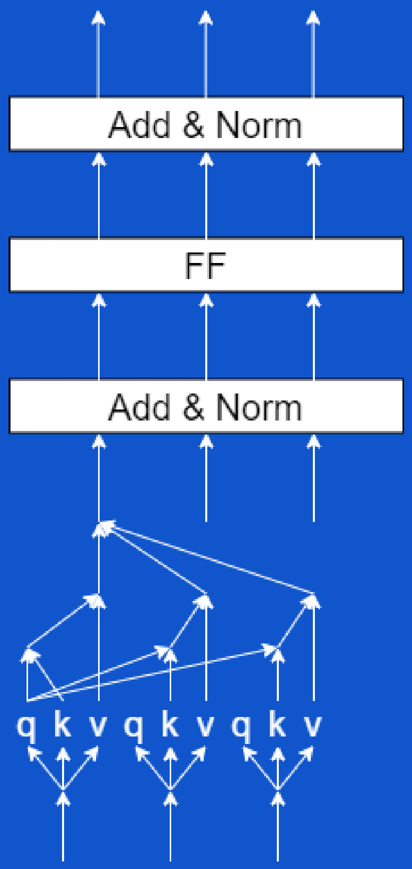
- Optimize word vectors during training
- Post editing
  - Removing stuttering effect in English
- Optimize Big Bird attention mechanism to boost efficiency

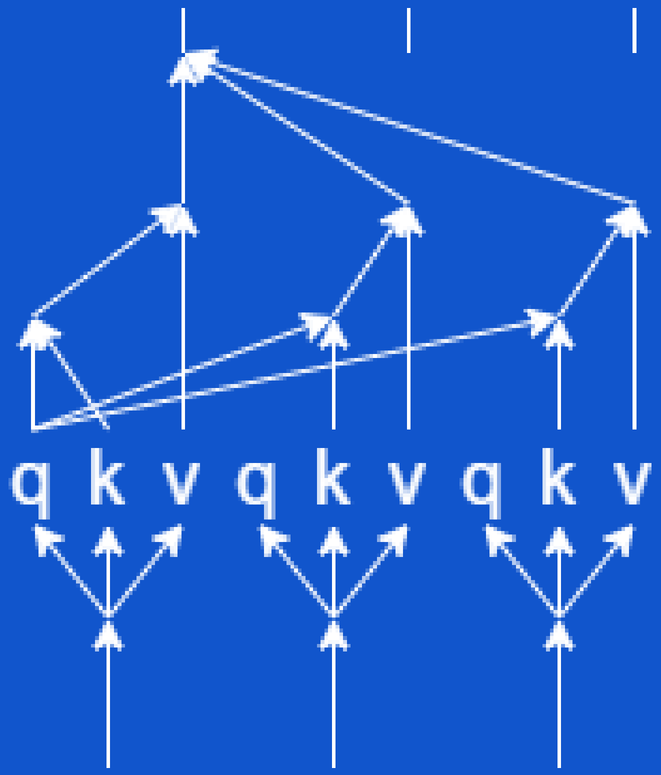
Thank you!

Attention Q, K, V?

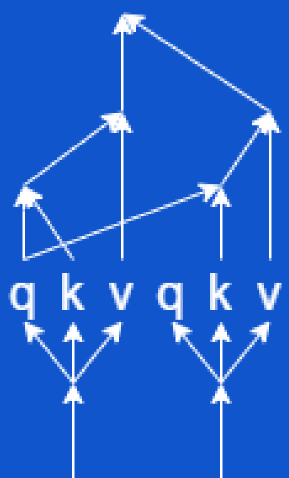
- Input sequence [1, 2, 3]
  - 3x3 matrix
- Q => **input * 2** = [2, 4, 6]
  - Q => input * $\boldsymbol{W_q}$

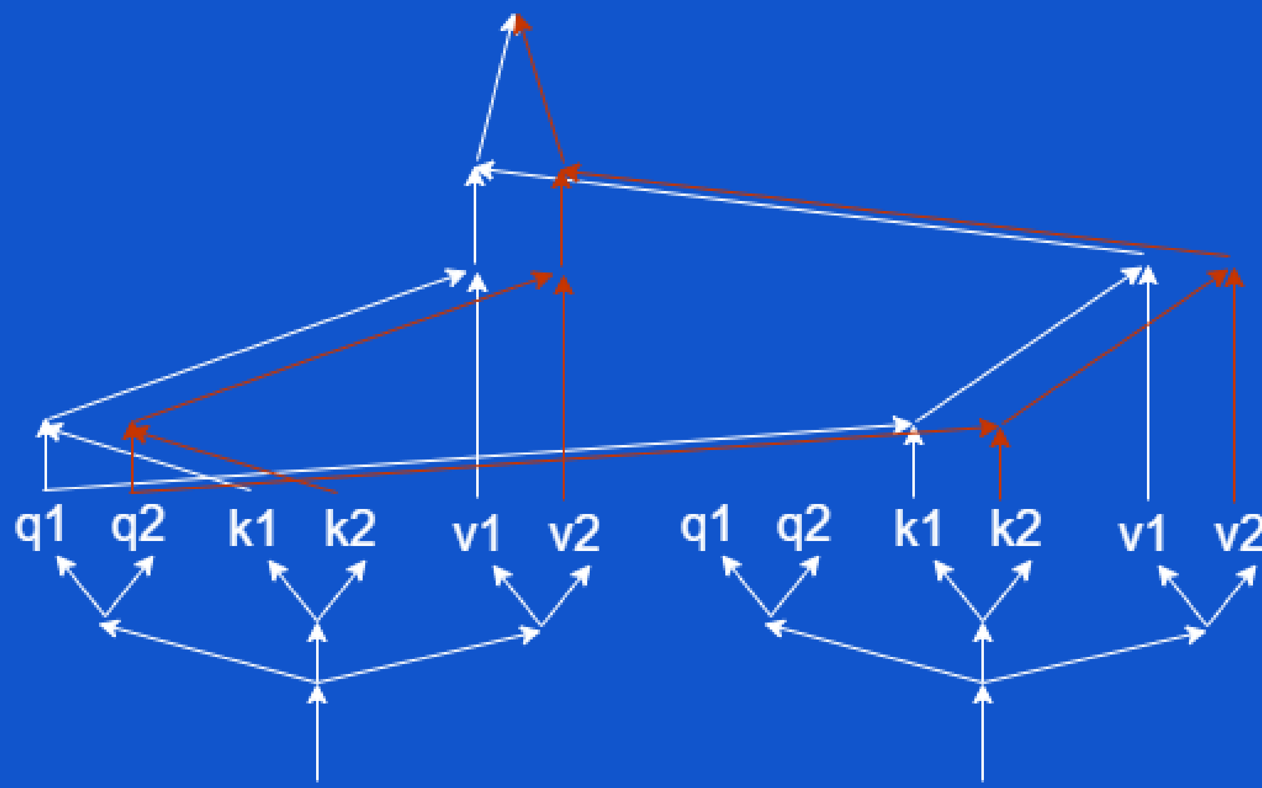
Multihead attention

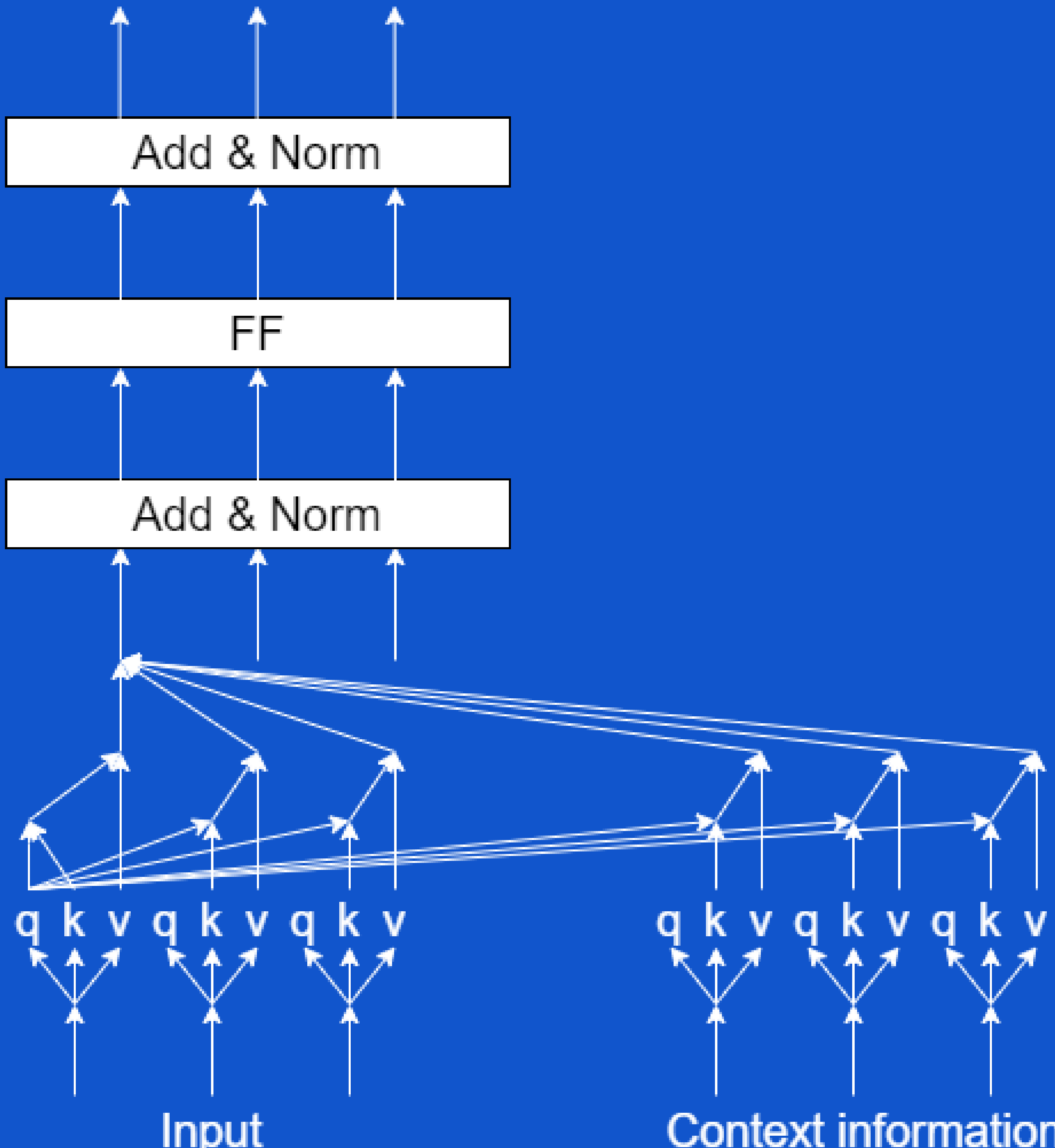




Multihead attention






Add & Norm
FF
Add & Norm
q k v q k v q k v
q k v q k v q k v
Input
Context information

Implementation

New model metric

Prediction	Truth
$[y_0, y_1, y_2, \ldots, y_{\#wordsInOutput}]$	$[v_0, v_1, v_2, \ldots, v_{\#wordsInTruth}]$
$y_0 = [y_0^0, y_0^1, y_0^2, \ldots, y_0^{299}]$	$v_0 = [v_0^0, v_0^1, \ldots, v_0^{299}]$
Distance between y_0 and v_0	**MSE between y_0 and v_0**
$\sqrt{(y-v)^2}$ $= \sqrt{(y_0^0 - v_0^0)^2 + \cdots + (y_0^{299} - v_0^{299})^2}$ $= \sqrt{\sum_{i=0}^{\dim(y)} (y_0^i - v_0^i)^2}$	$\frac{1}{\dim(y)} \sum_{i=0}^{\dim(y)} (y_0^i - v_0^i)^2$
$\sqrt{\sum_{i=0}^{\dim(y)} (y_0^i - v_0^i)^2} \propto \frac{1}{\dim(y)} \sum_{i=0}^{\dim(y)} (y_0^i - v_0^i)^2$	