MACQUARIE University

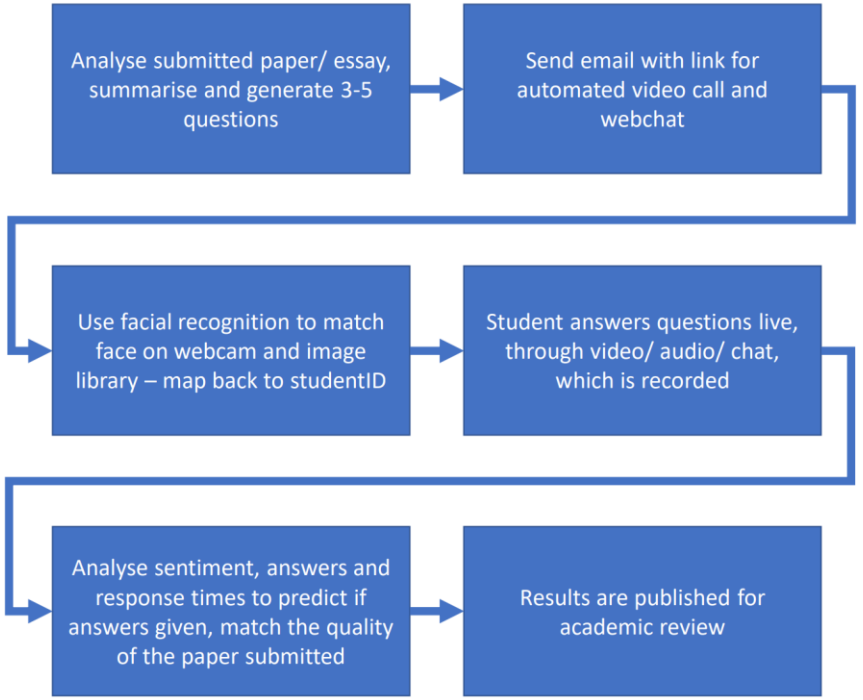
ITIC systems

Automatic Exam Marking Project

Phase 1: Exam Management Product

Product ML Engine

Required Core Modules:
Bert & Transformers



MACQUARIE University

ITIC systems

Automatic Exam Marking Project

Phase 1: Exam Management Product

PART 1 : Language Understanding Models & Bert Overview

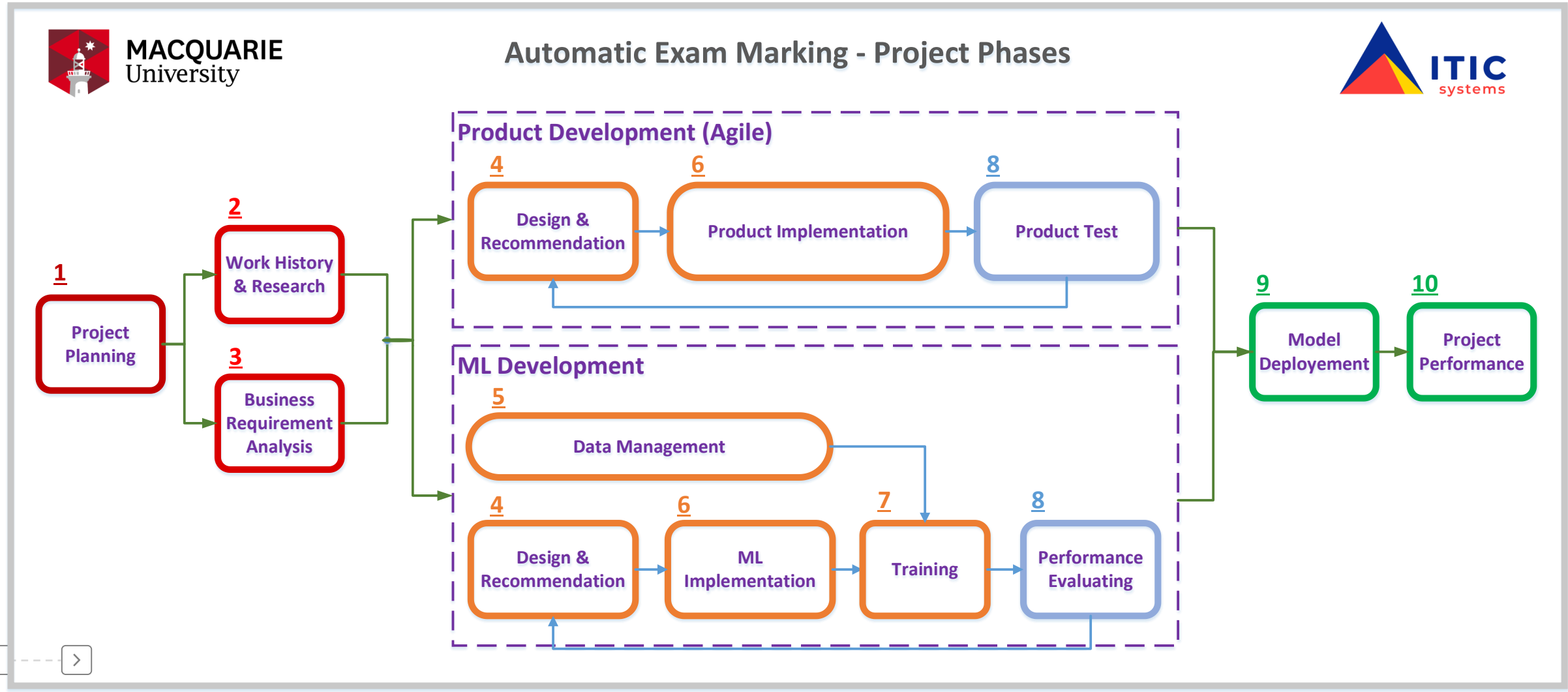
PART 2 : Transformer Architecture Overview &
Encoder Part Technical Detail

PART 3 : Bert Technical Detail & Fine Tuning Examples

PART 4 : New Task & Code Analysis Definition

PART 5 : References

Project Current Status



MACQUARIE University

ITIC systems

Automatic Exam Marking Project

Phase 1: Exam Management Product

Product ML Engine

PART 1 : Language Understanding Models

&

Bert Overview

NLP Tasks

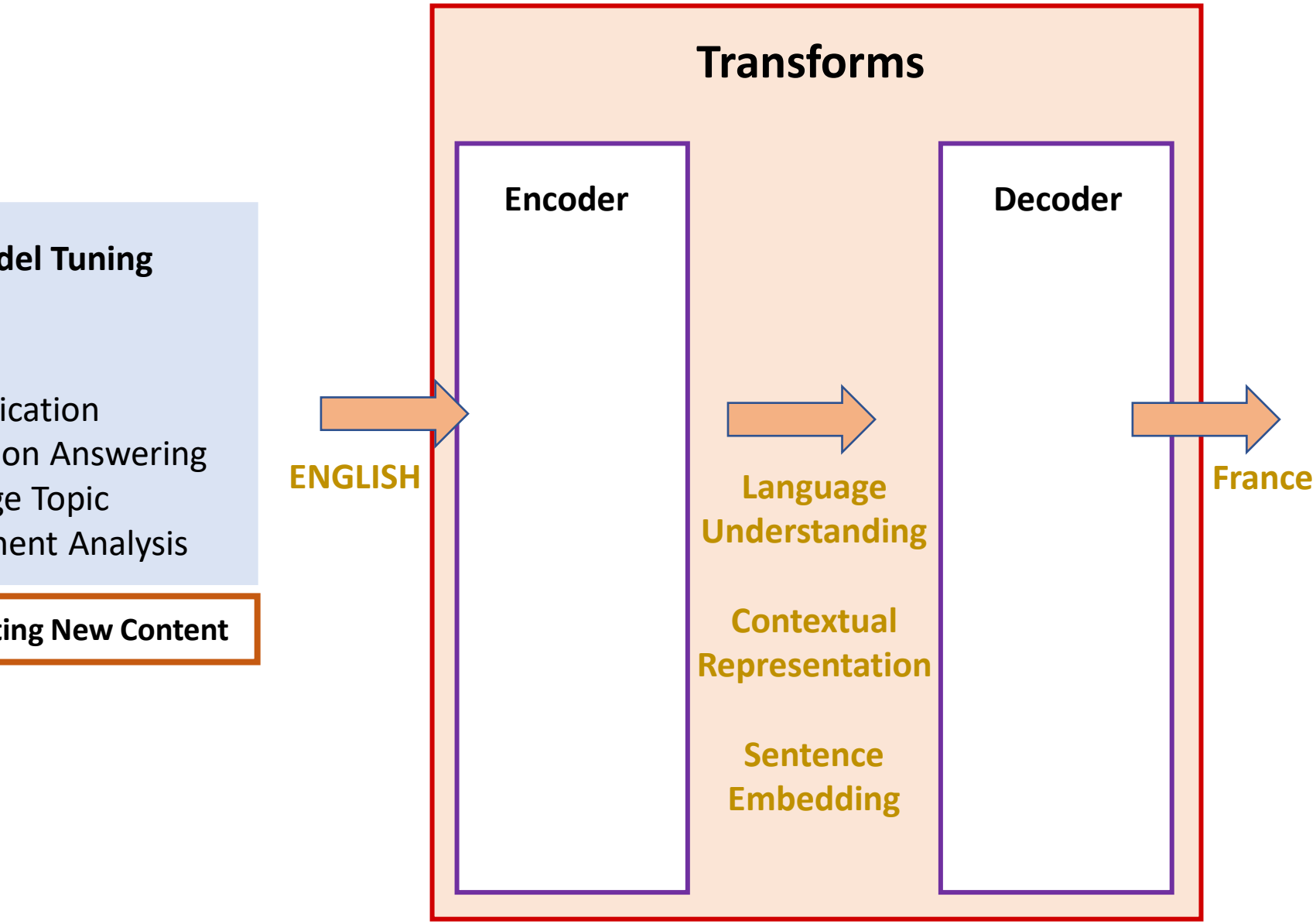
Pre-train Phase

Language Understanding

Model Tuning

Classification
Question Answering
Passage Topic
Sentiment Analysis

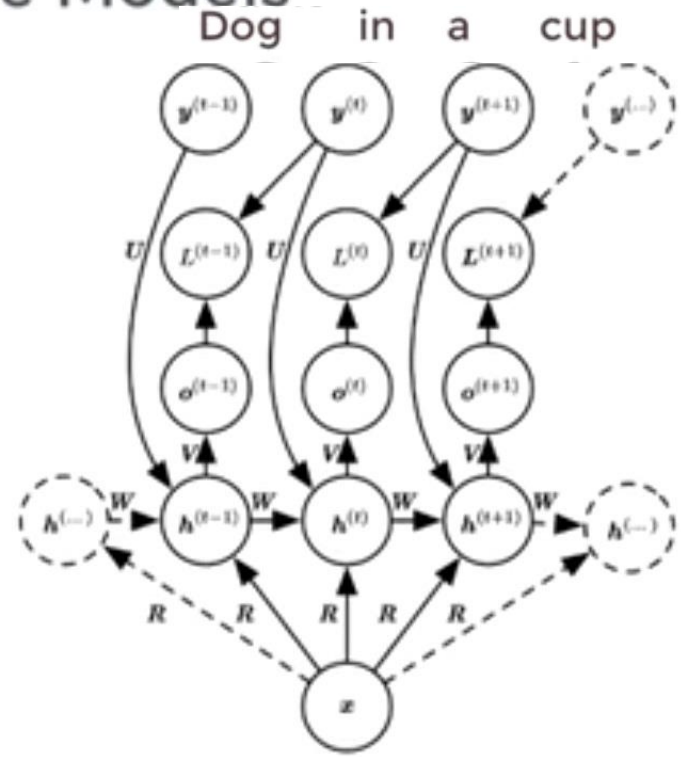
Generating New Content



ITIC systems

Recurrent Neural Networks

1. Vector-Sequence Models

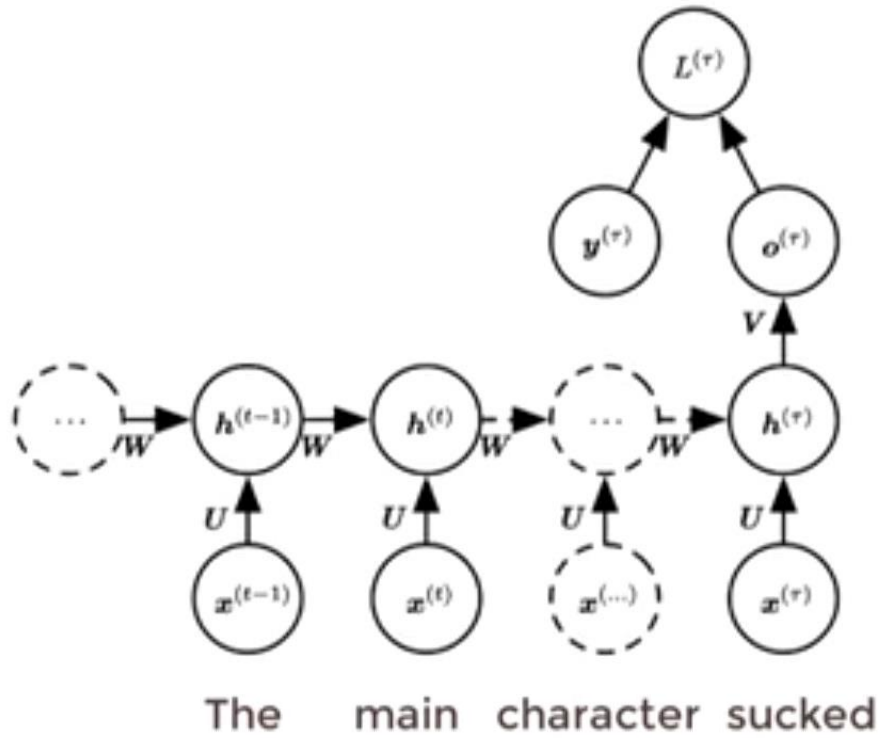


ITIC systems

Recurrent Neural Networks

2. Sequence-Vector Models

$$\begin{bmatrix} 0.11 \\ 0.89 \end{bmatrix}$$ good / bad

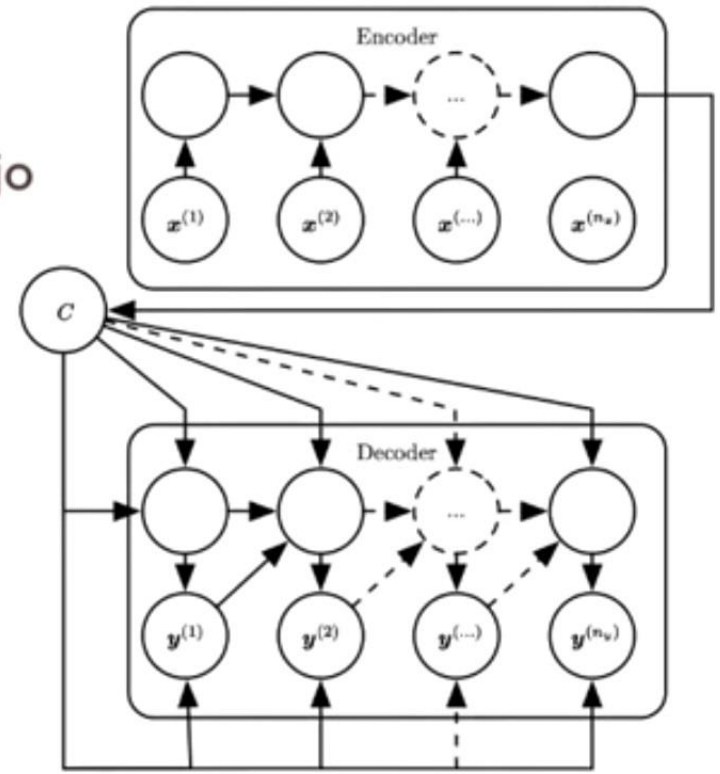


ITIC systems

Recurrent Neural Networks

3. Sequence-Sequence Models

El gran perro rojo



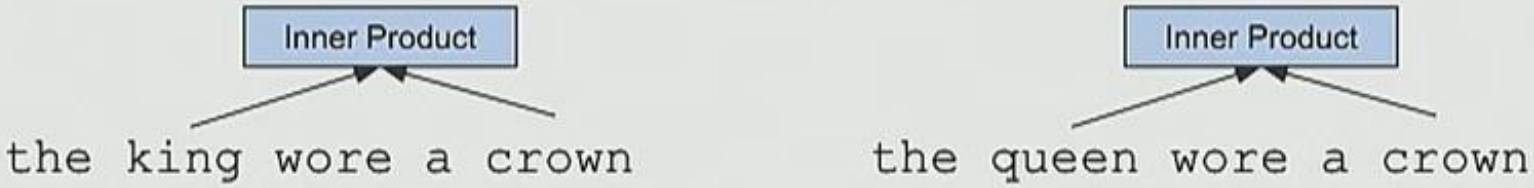
The big red dog

ITIC systems

Pre-training in NLP

- Word embeddings are the basis of deep learning for NLP

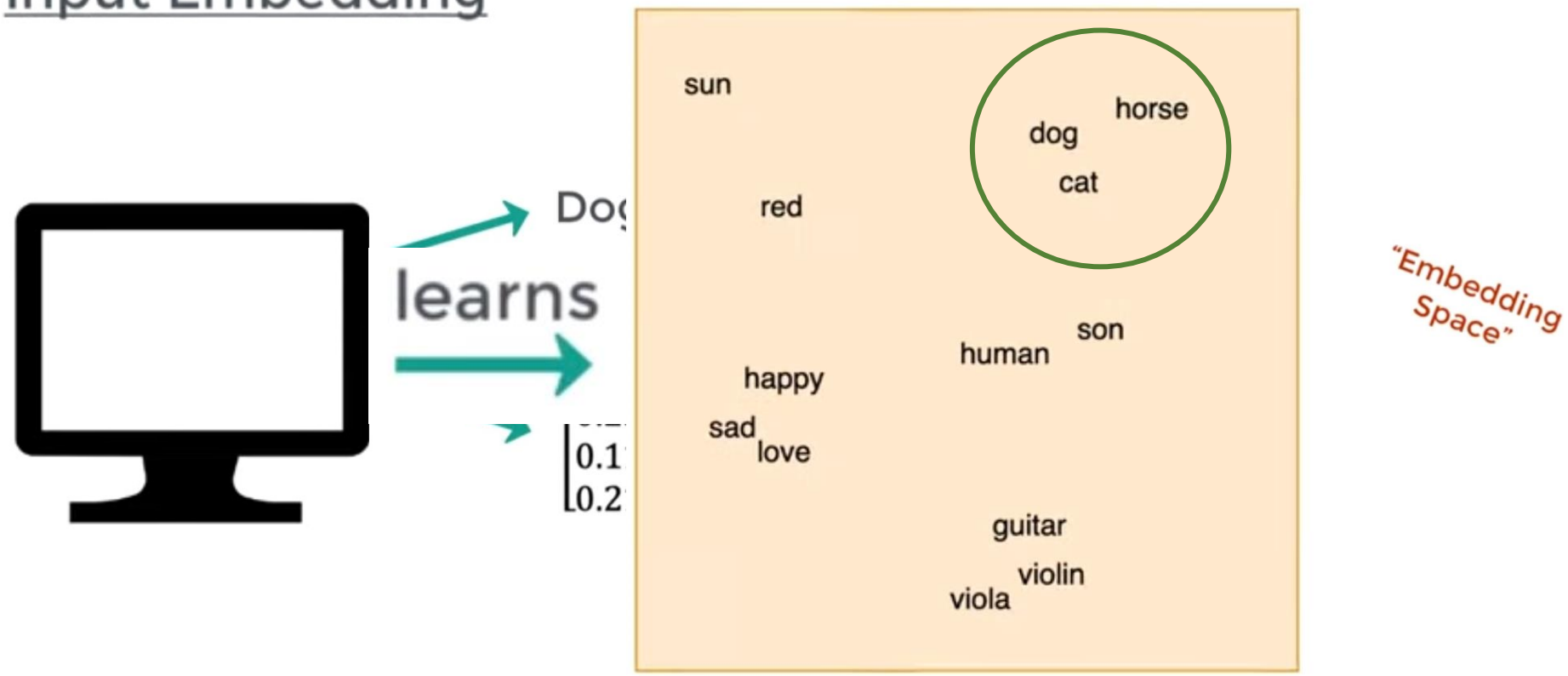
- Word embeddings (`word2vec`, `GloVe`) are often *pre-trained* on text corpus from co-occurrence statistics



ITIC systems

Transformer Components

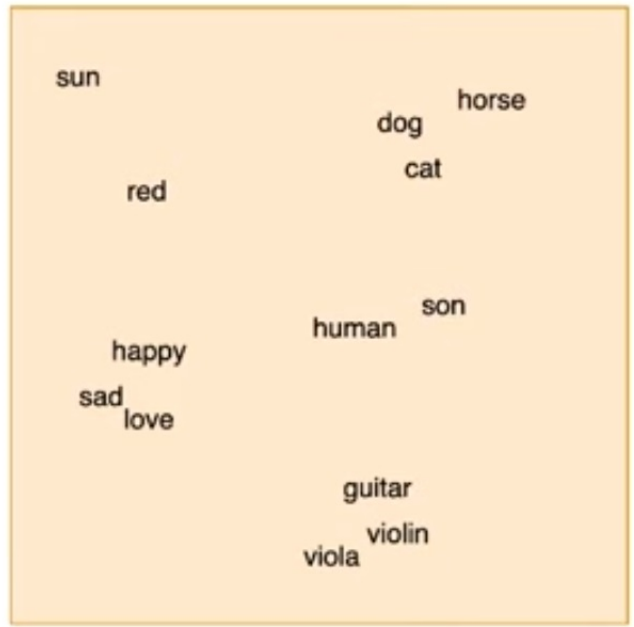
Input Embedding



Transformer Components

Input Embedding

dog →



→ $\begin{bmatrix} 0.37 \\ 0.99 \\ 0.01 \\ 0.08 \end{bmatrix}$

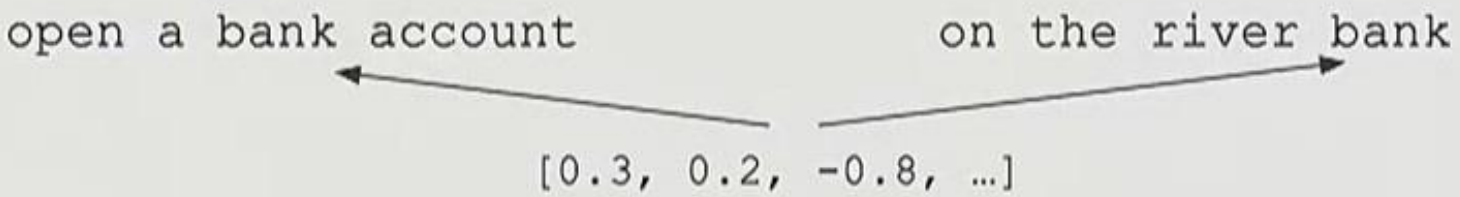
AJ's dog is a cutie

AJ looks like a dog

ITIC systems

Contextual Representations

- **Problem**: Word embeddings are applied in a context free manner



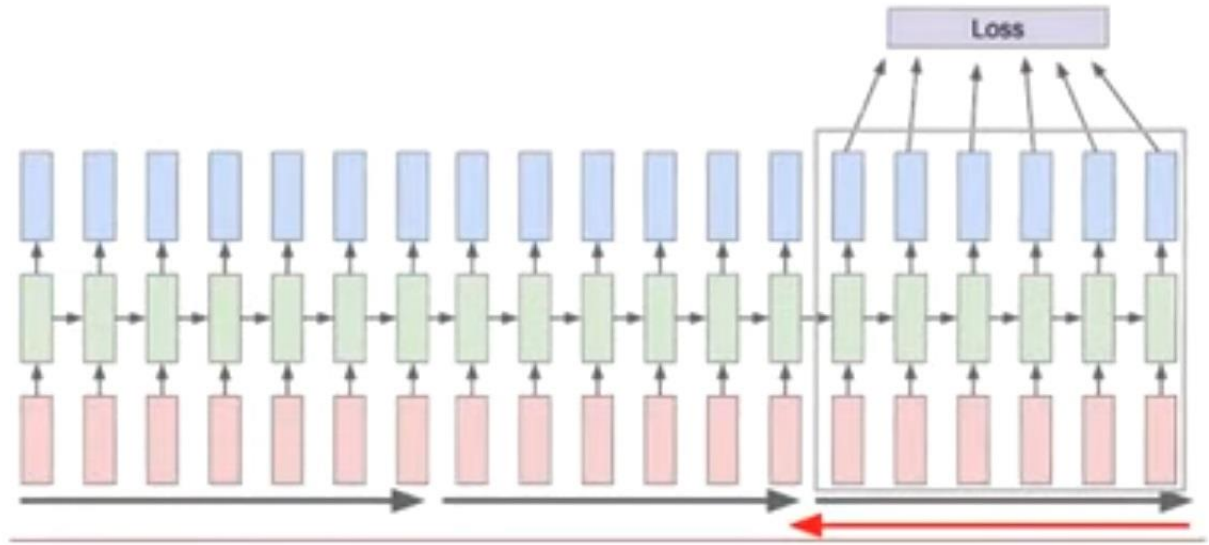
- **Solution**: Train *contextual* representations on text corpus

ITIC systems

Recurrent Neural Networks

Disadvantages

1. Slow to train.



Truncated Backprop Through Time

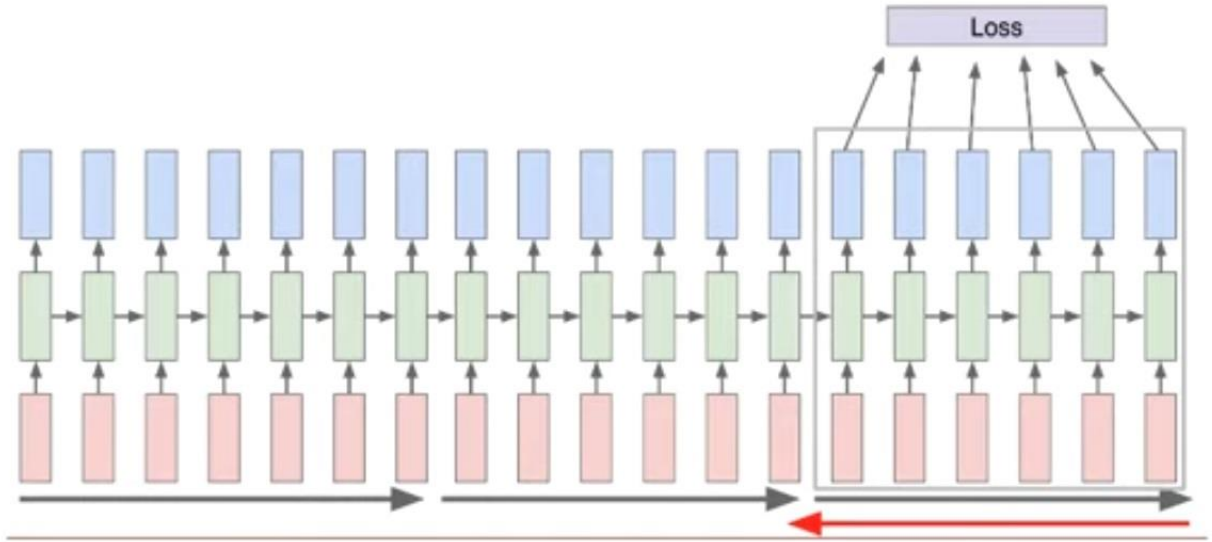
Recurrent Neural Networks

Disadvantages

1. Slow to train.

2. Long sequences lead to vanishing/ exploding gradients

3. It not bidirectional

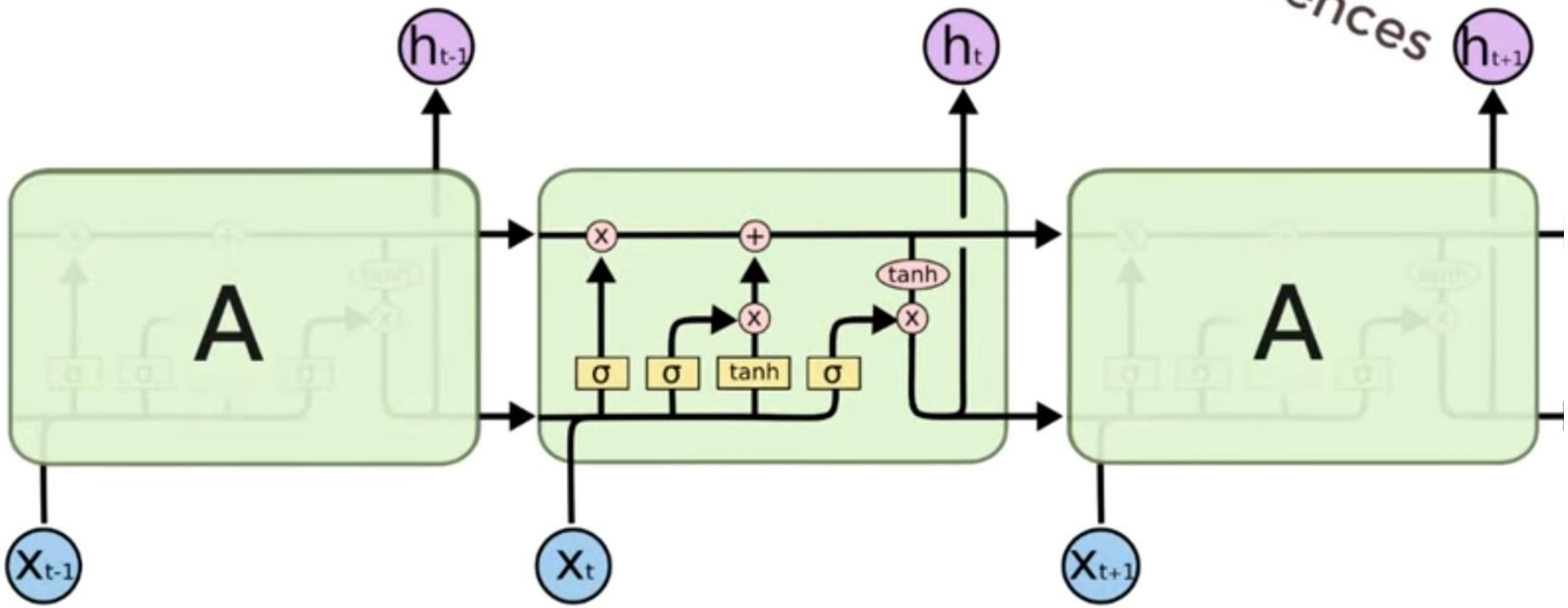


Truncated Backprop Through Time

ITIC systems

LSTM Networks

Longer sequences

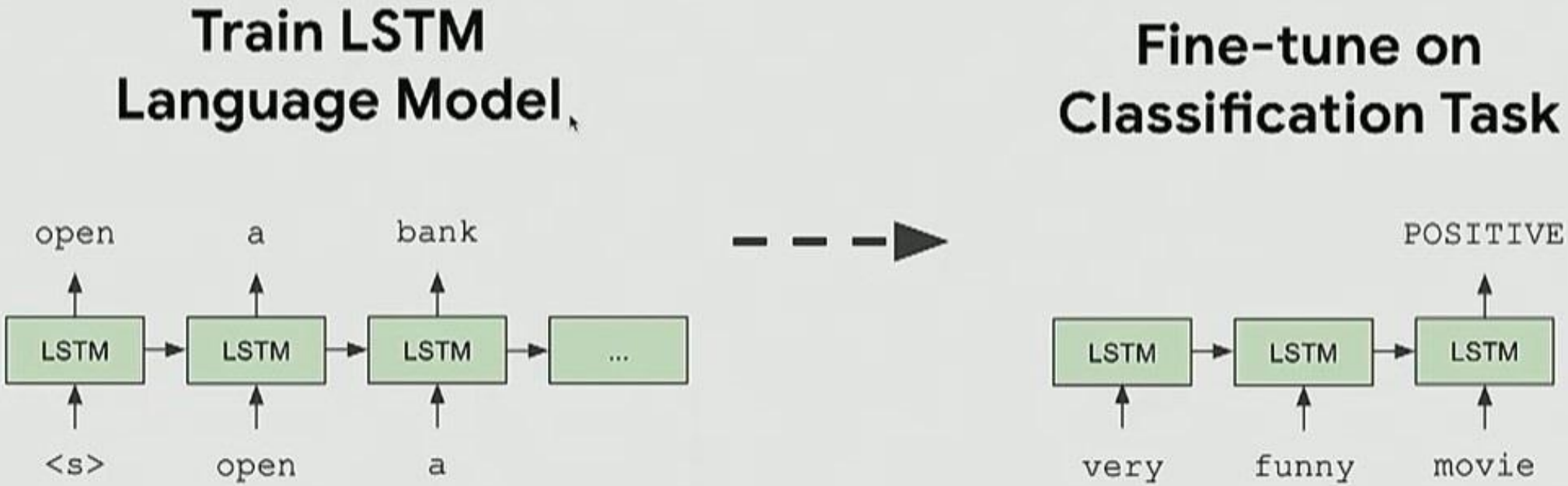


The repeating module in an LSTM contains four interacting layers.

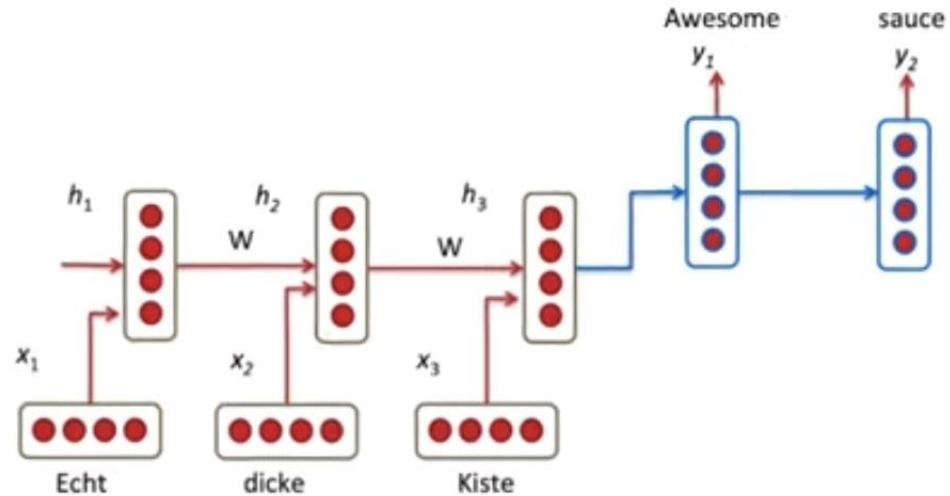
ITIC systems

History of Contextual Representations

- *Semi-Supervised Sequence Learning*, Google, 2015



LSTM



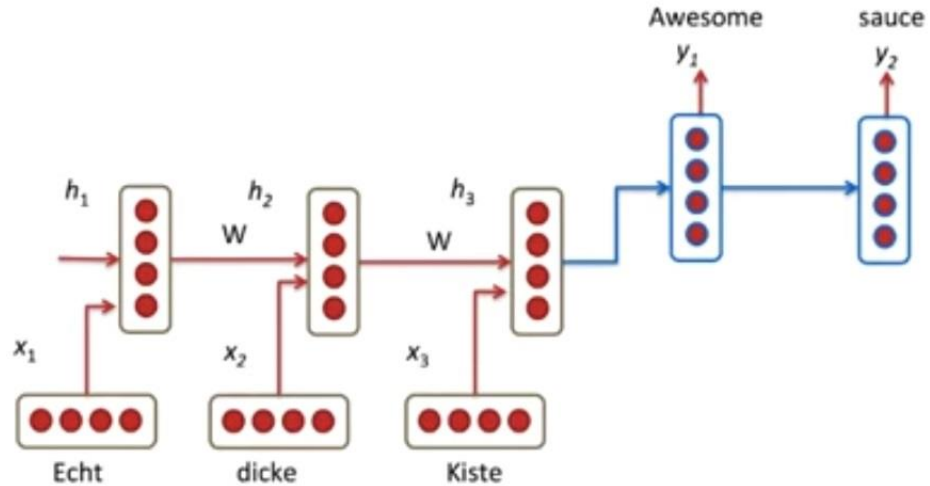
Source [Blog]: Attention & Memory in Deep Learning (Britz, 2017)

ITIC systems

LSTM Vs Transformer

LSTM Networks

1. Slow



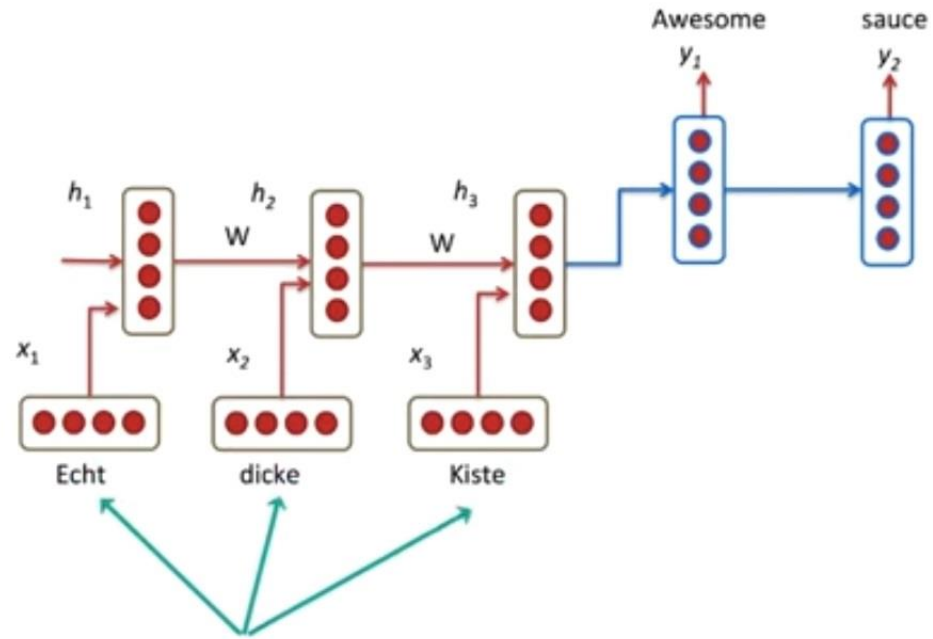
Source [Blog]: Attention & Memory in Deep Learning (Britz, 2017)

ITIC systems

LSTM

1. Slow

2. Not truly Bidirectional



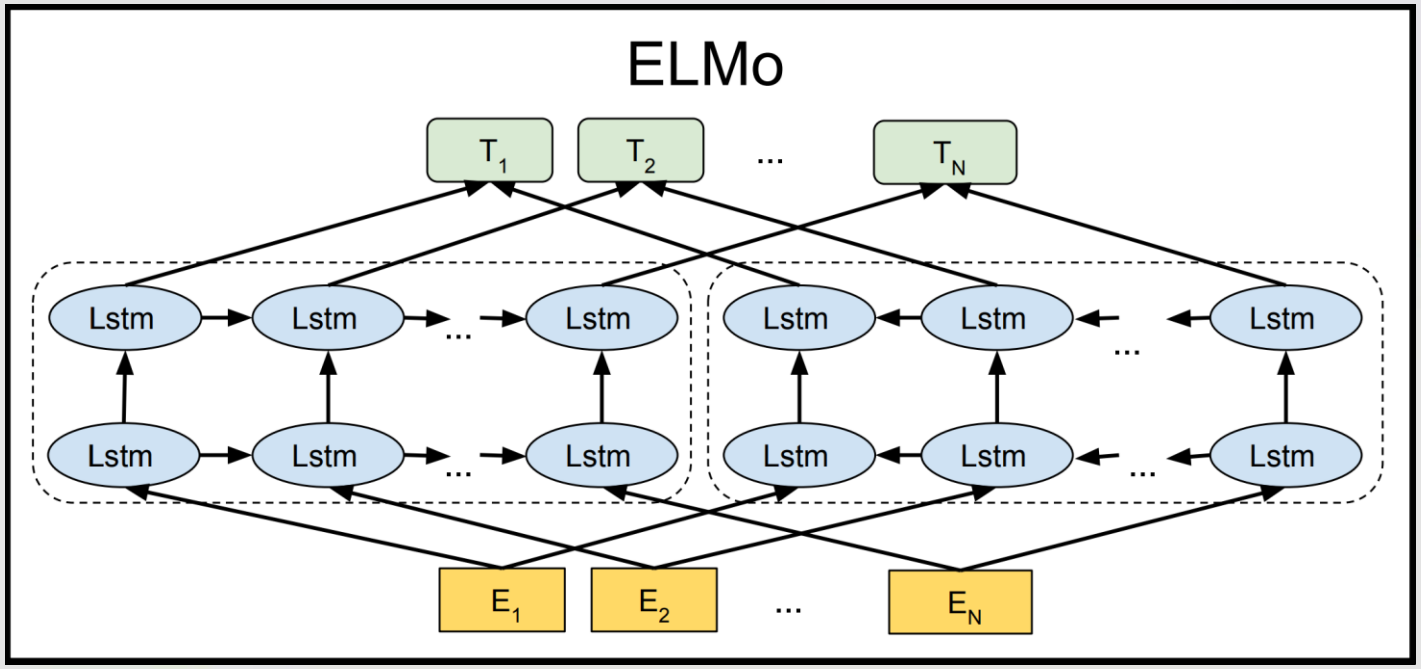
Source [Blog]: Attention & Memory in Deep Learning (Britz, 2017)

ITIC systems

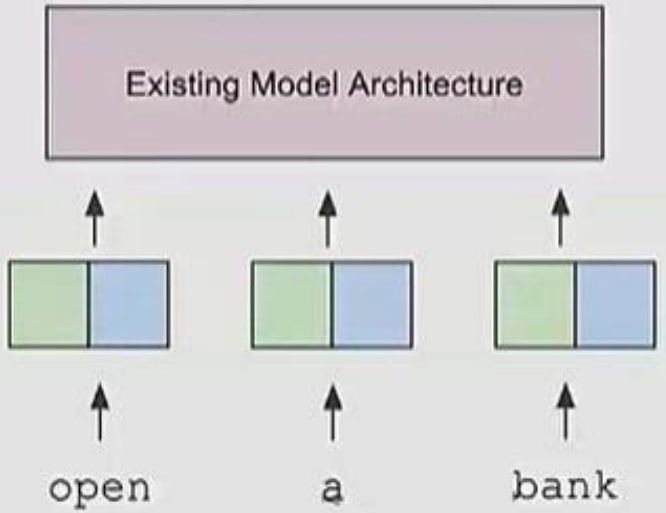
History of Contextual Representations

- *ELMo: Deep Contextual Word Embeddings*, AI2 & University of Washington, 2017

Train Separate Left-to-Right and Right-to-Left LMs



Apply as "Pre-trained Embeddings"



ITIC systems

Model Architecture

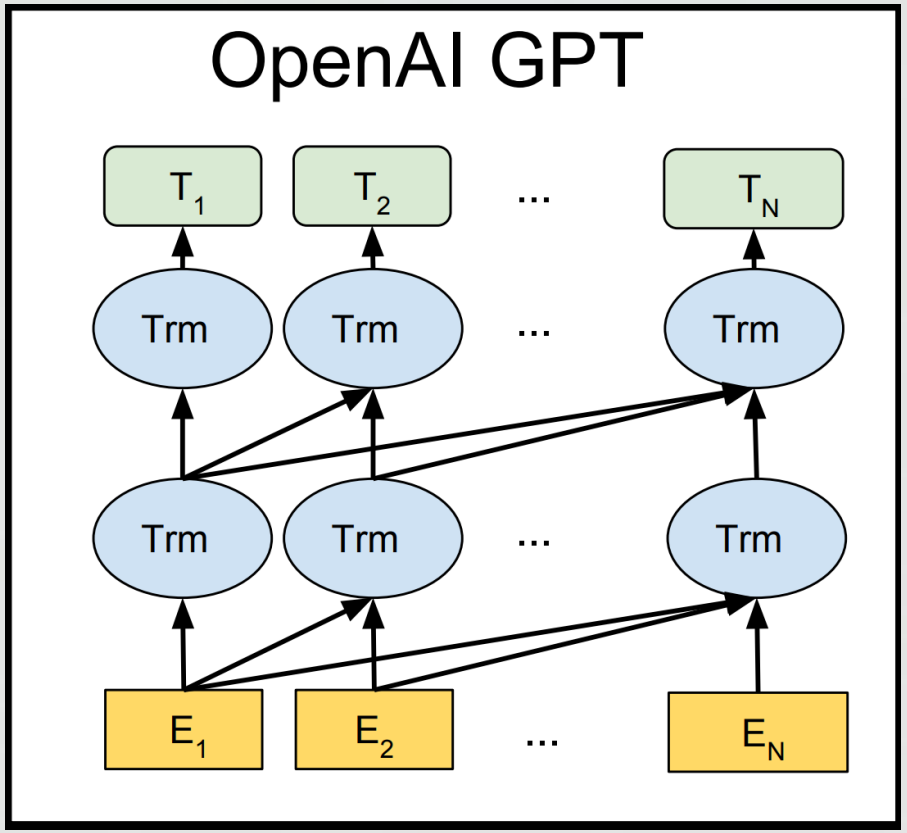
- Empirical advantages of Transformer vs. LSTM:

1. Self-attention == no locality bias
   - Long-distance context has “equal opportunity”
2. Single multiplication per layer == efficiency on TPU
   - Effective batch size is number of *words*, not *sequences*

nidirectional vs. Bidirectional Models

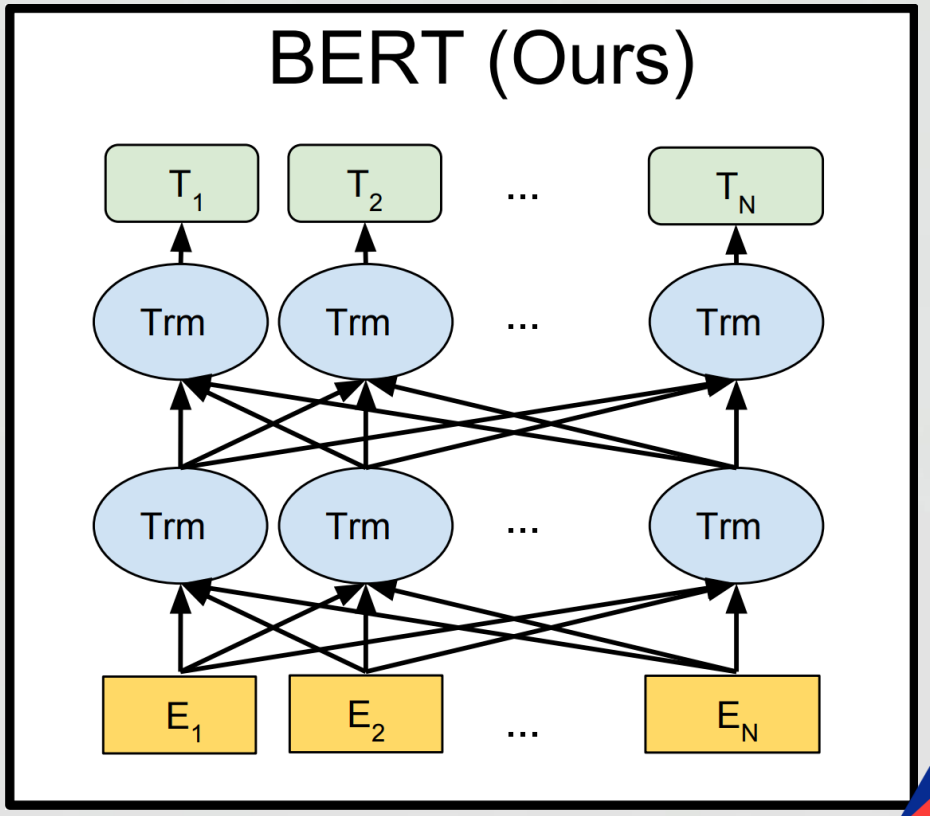
Unidirectional context

Build representation incrementally



Bidirectional context

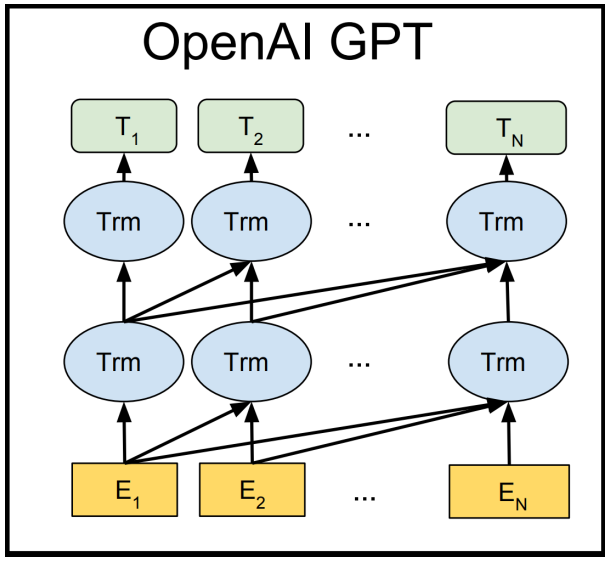
Words can “see themselves”



ITIC systems

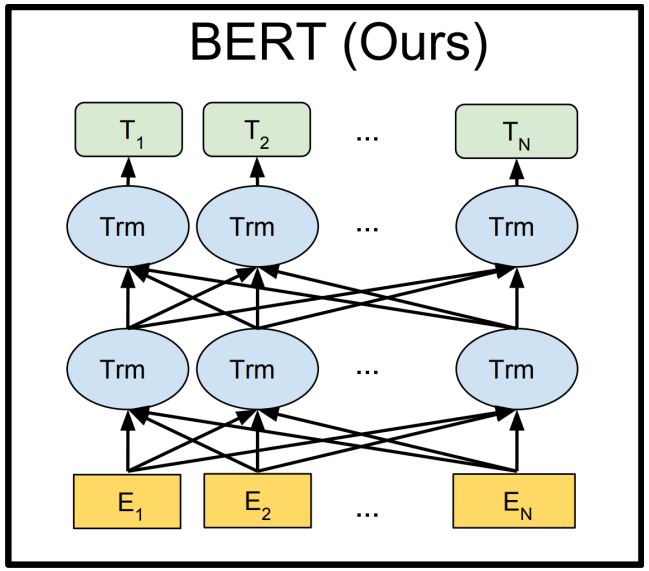
Unidirectional vs. Bidirectional Models

Unidirectional



It is good to understand a Language
But it required to input one by on
It uses for specific task
Training Task is: next word prediction

Bidirectional

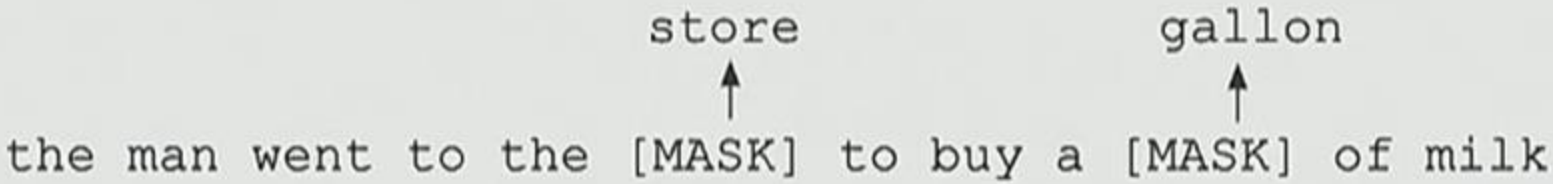


It gives us better Understanding because of Bidirectional
It input all words simultaneously
It is not good for all part generation tasks
So ho we cant trian the model to language undertaning
Different Training Tasks

ITIC systems

Masked LM

- **Solution**: Mask out *k*% of the input words, and then predict the masked words
  - We always use *k* = 15%



- Too little masking: Too expensive to train
- Too much masking: Not enough context

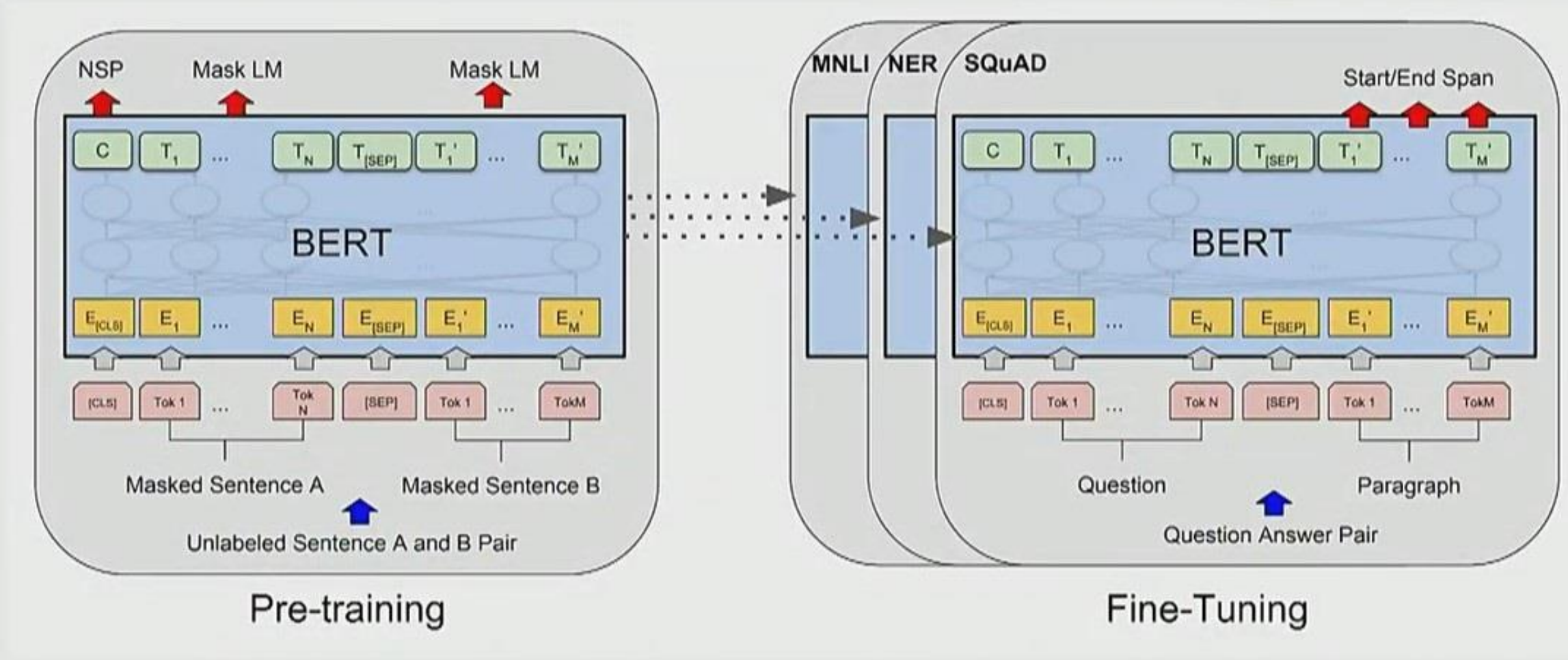
Next Sentence Prediction

- To learn *relationships* between sentences, predict whether Sentence B is actual sentence that proceeds Sentence A, or a random sentence

Sentence A = `The man went to the store.`
Sentence B = `He bought a gallon of milk.`
Label = `IsNextSentence`

Sentence A = `The man went to the store.`
Sentence B = `Penguins are flightless.`
Label = `NotNextSentence`

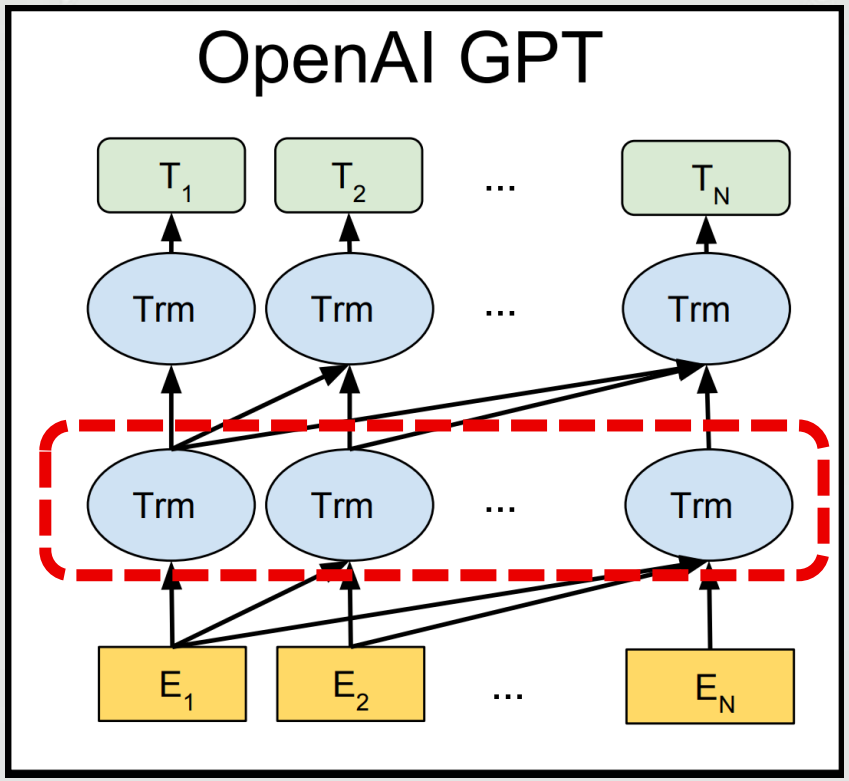
Fine-Tuning Procedure



Unidirectional vs. Bidirectional Models

Unidirectional context

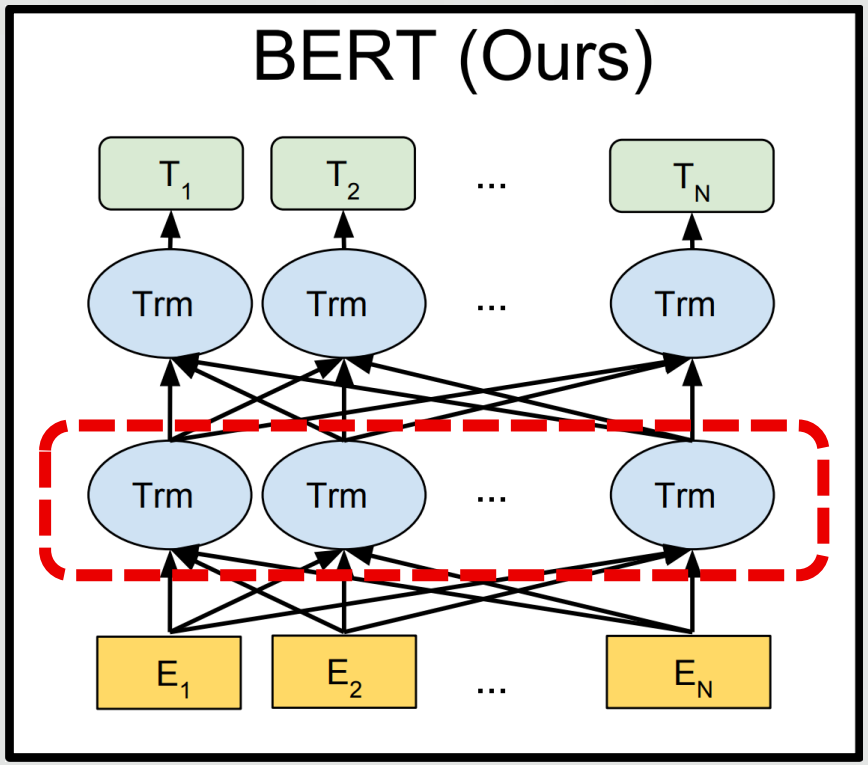
Build representation incrementally



Transformer Encoder

Bidirectional context

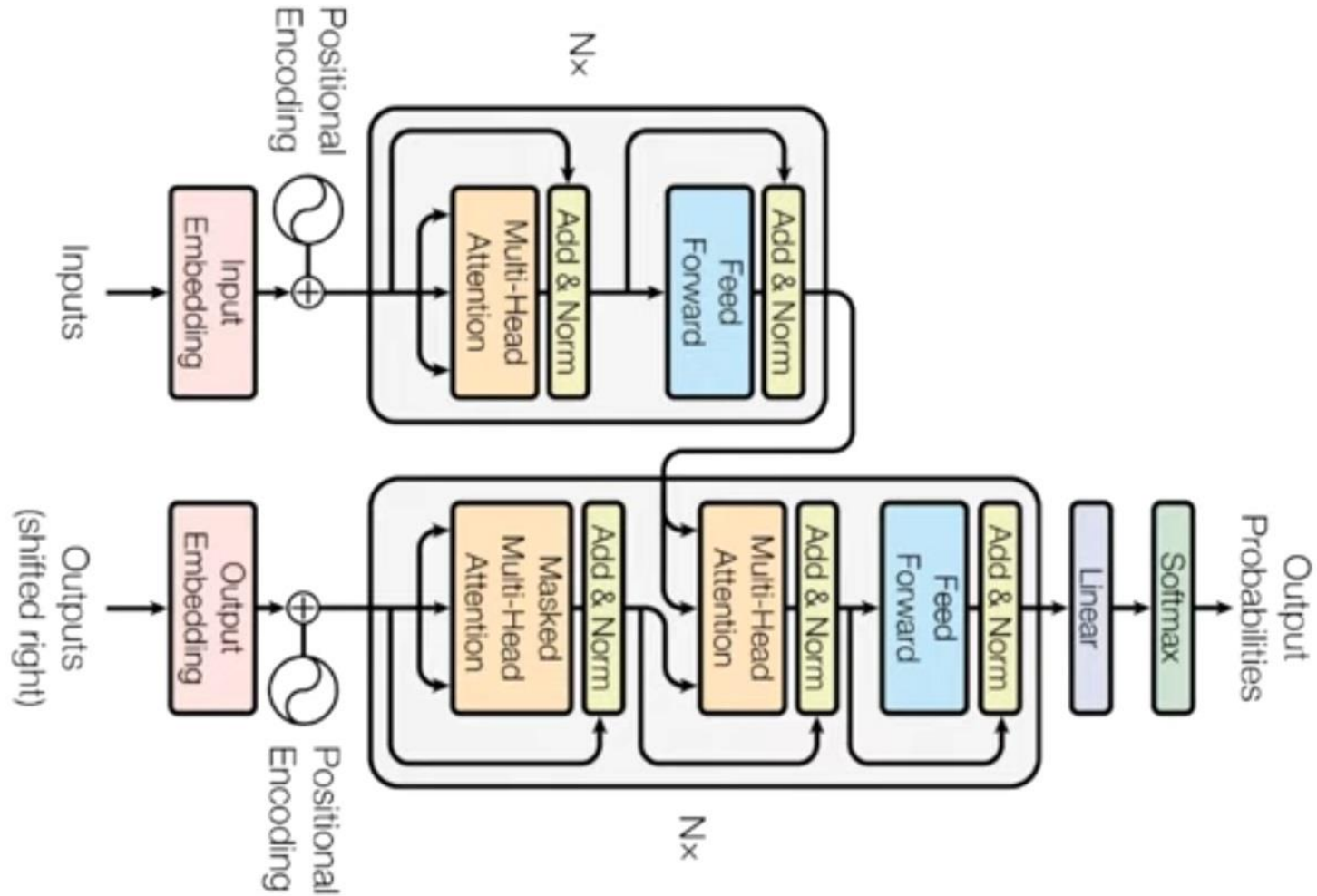
Words can “see themselves”



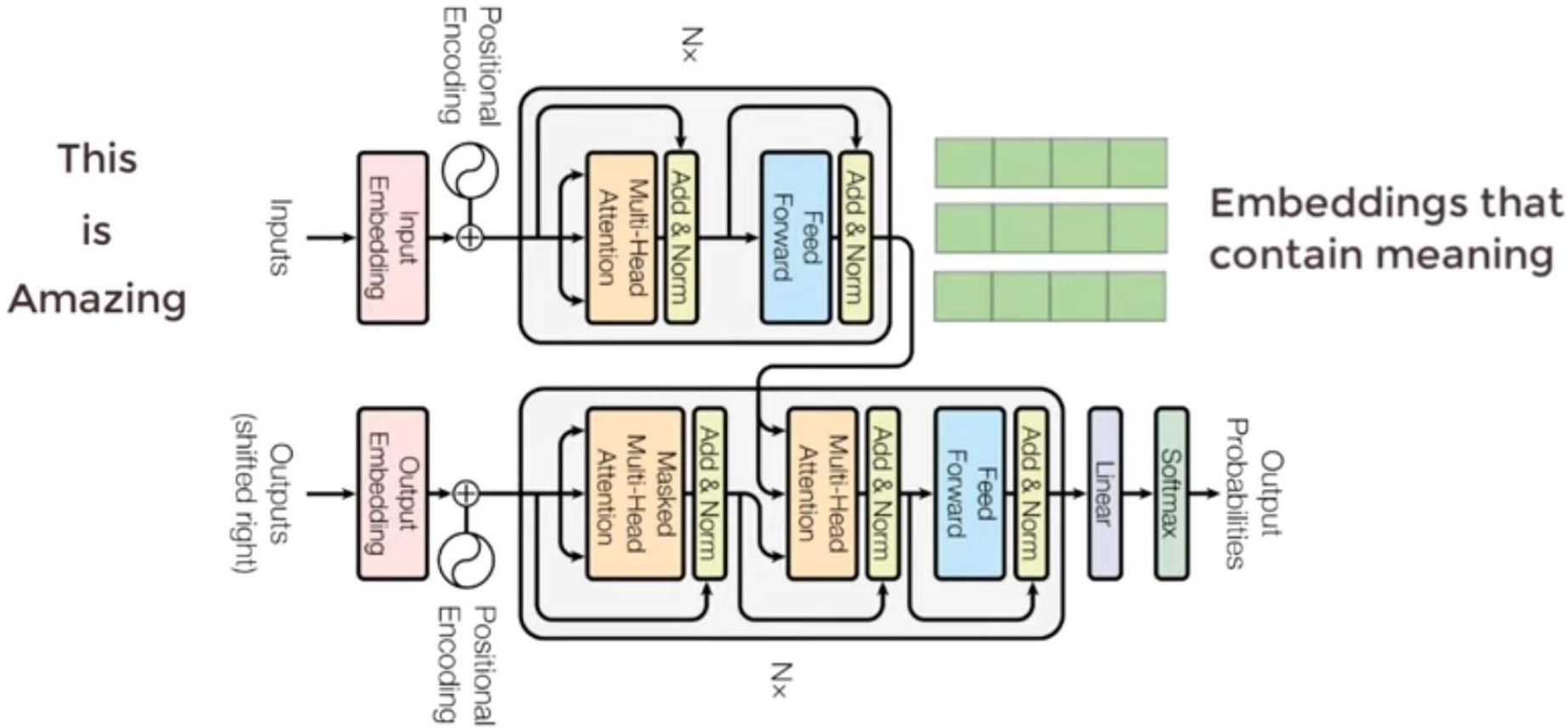
Transformer Decoder

ITIC systems

Transformer Flow

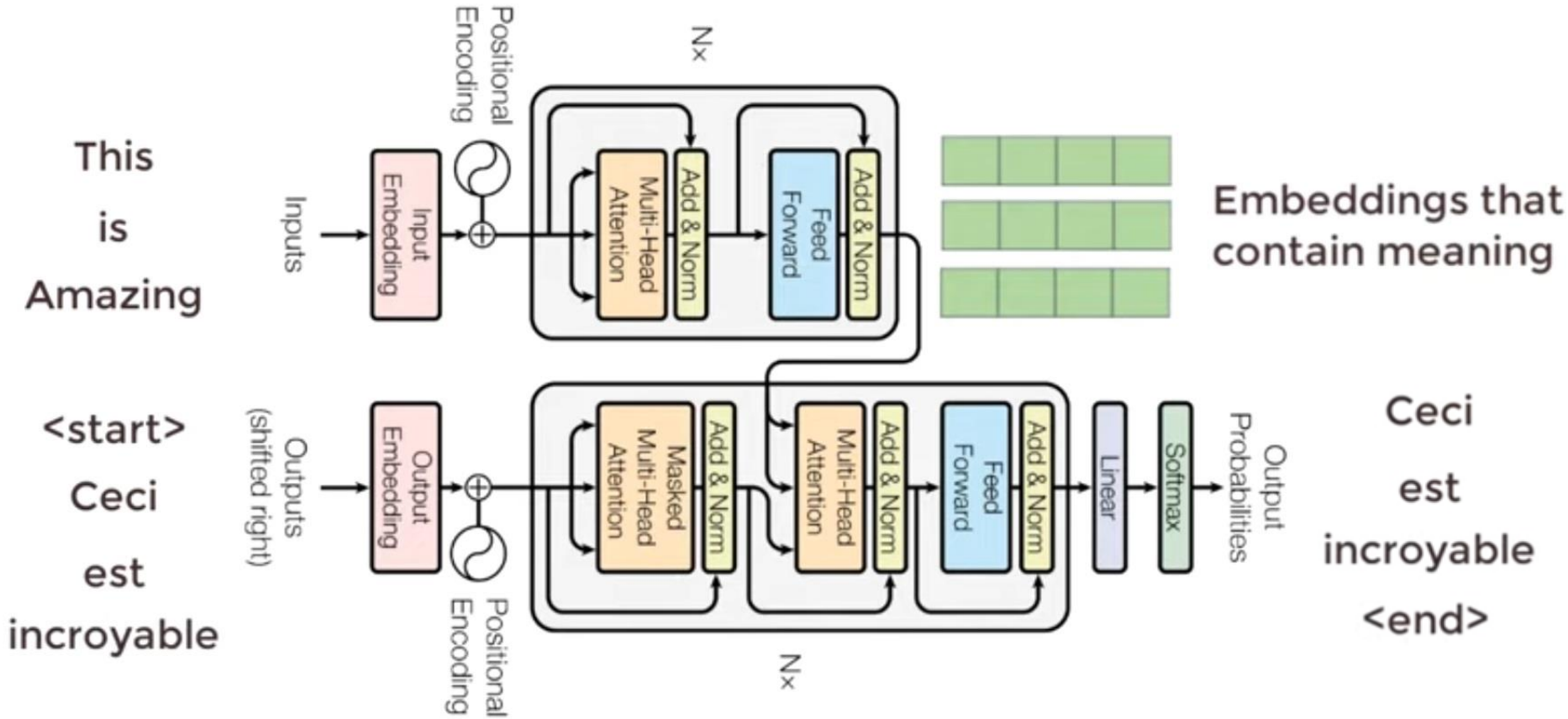


Transformer Flow



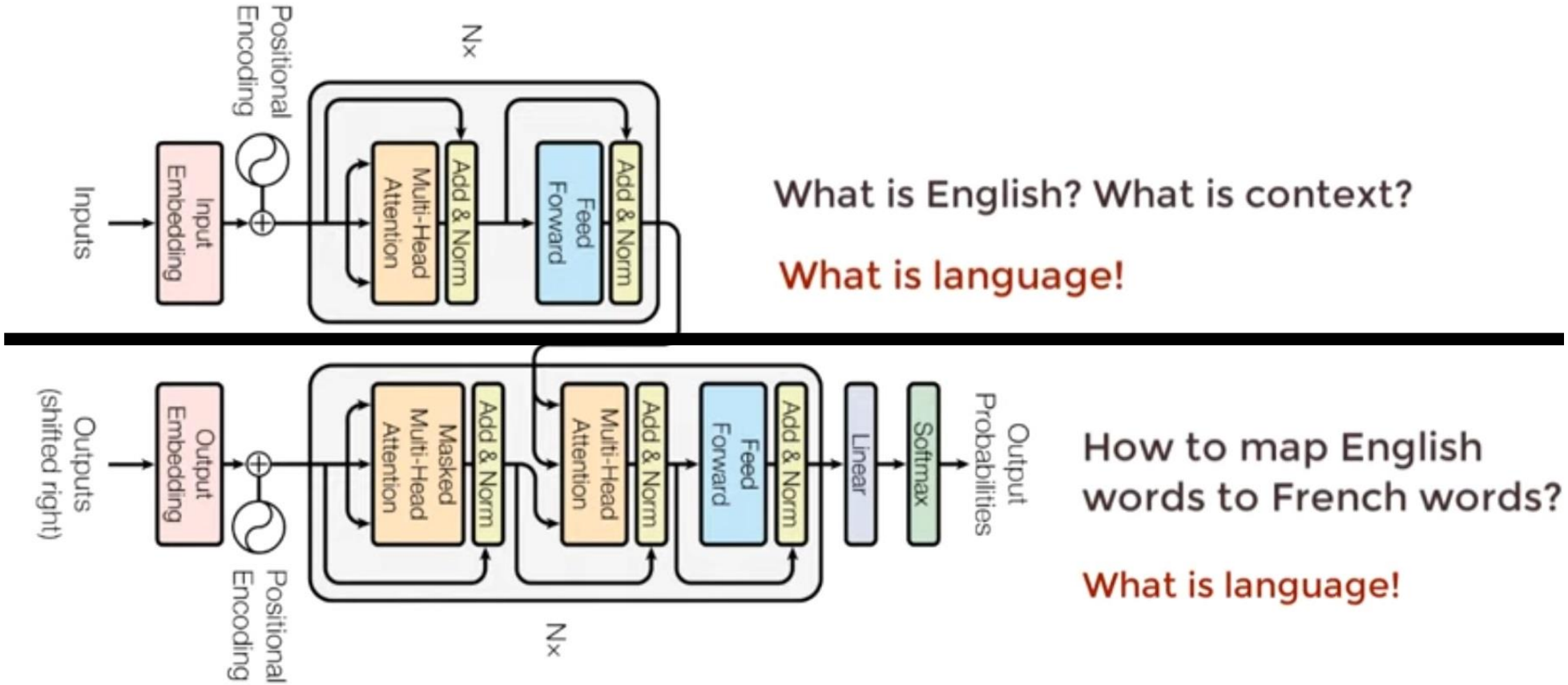
ITIC systems

Transformer Flow



ITIC systems

Transformer Flow



What is English? What is context?

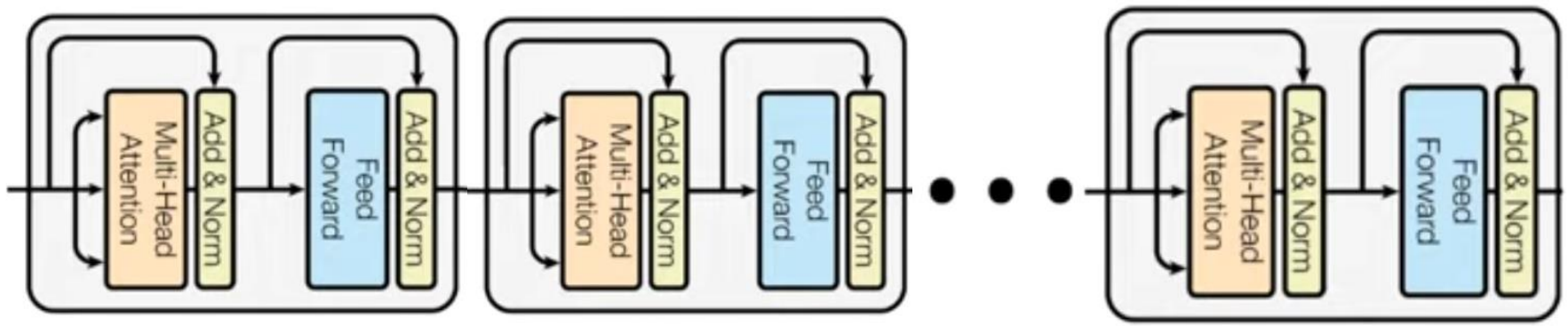
What is language!

How to map English words to French words?

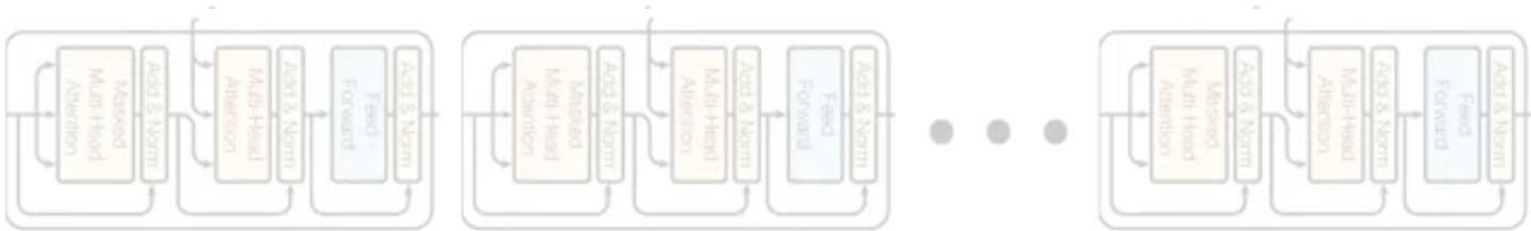
What is language!

ITIC systems

Transformer Flow



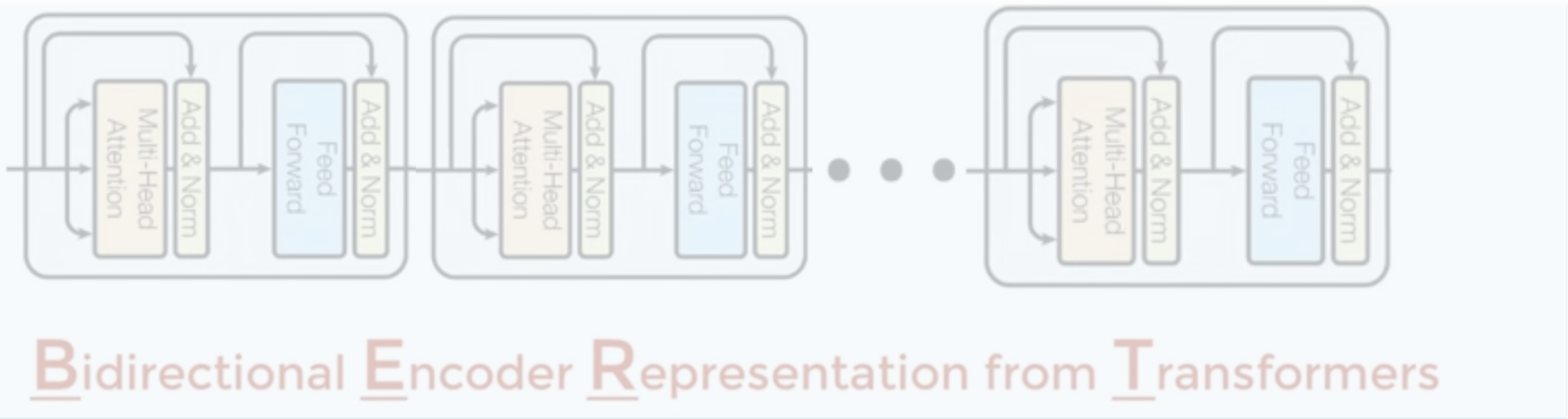
Bidirectional **E**ncoder **R**epresentation from **T**ransformers

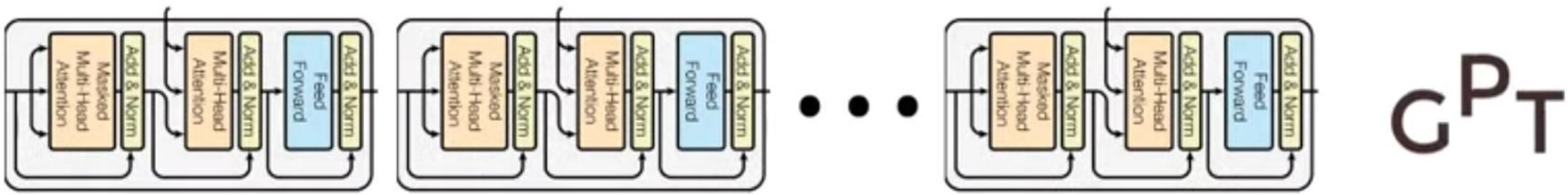


GPT

ITIC systems

Transformer Flow





ITIC systems

MACQUARIE University

ITIC systems

Automatic Exam Marking Project

Phase 1: Exam Management Product

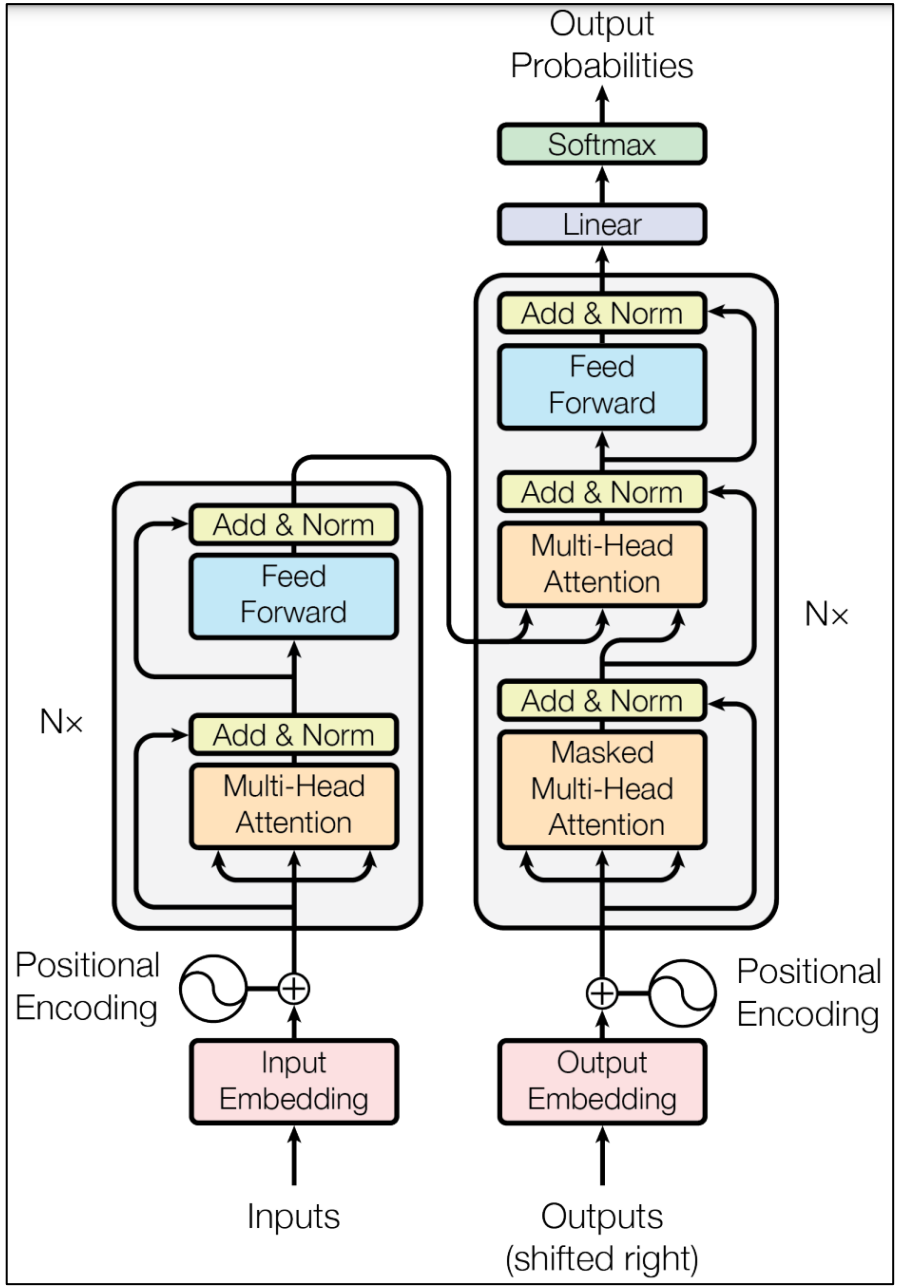
Product ML Engine

PART 2 : Transformer Architecture Overview

&

Encoder Part Technical Detail

Transformer Flow



ITIC systems

Transformer Components

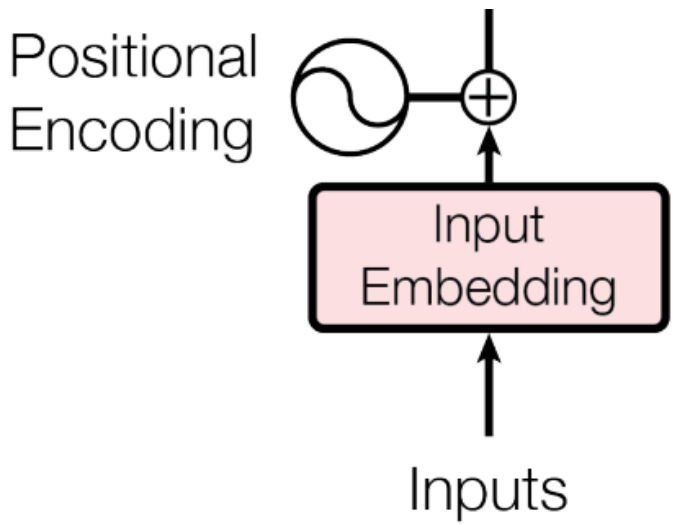
Positional Encoder :vector that gives context based on position of word in sentence

AJ's dog is a cutie → Position 2

AJ looks like a dog → Position 5

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$



ITIC systems

Transformer Components

Positional Encoder :vector that gives context based on position of word in sentence

Embedding of "Dog"

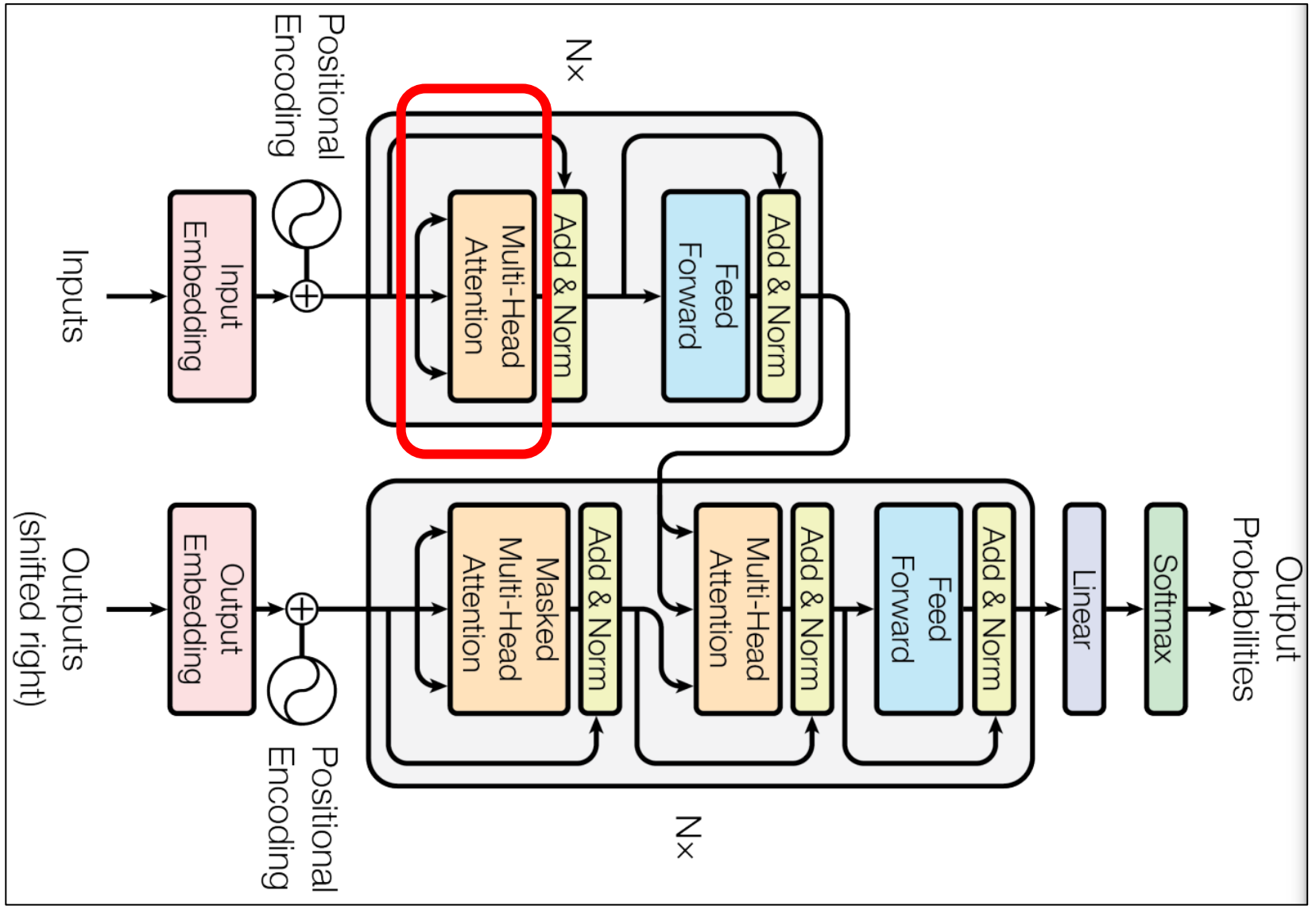
Vector Encoding of position in sentence

Embedding of Dog (with context info)

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

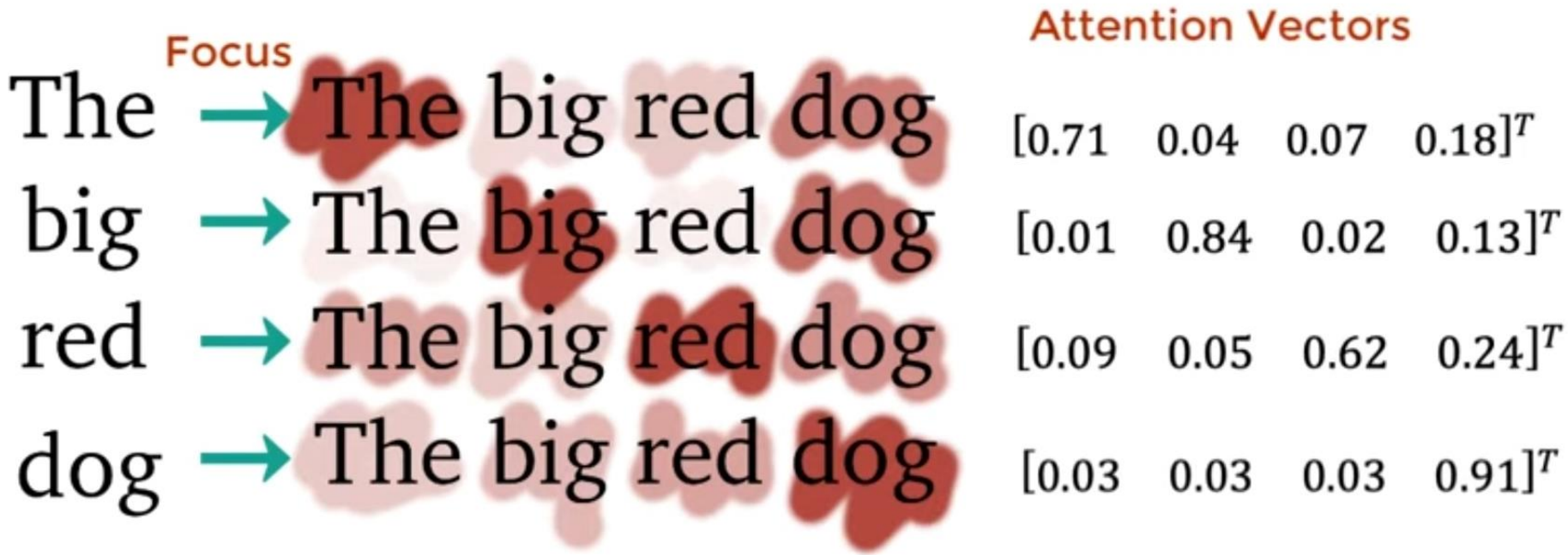
ITIC systems

Transformer Flow – Multi Head Attention



Transformer Components

Attention : What part of the input should we focus?

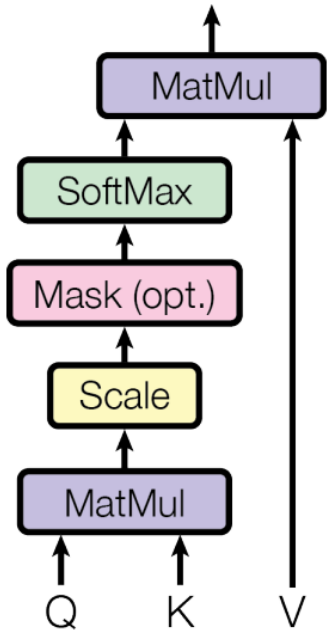


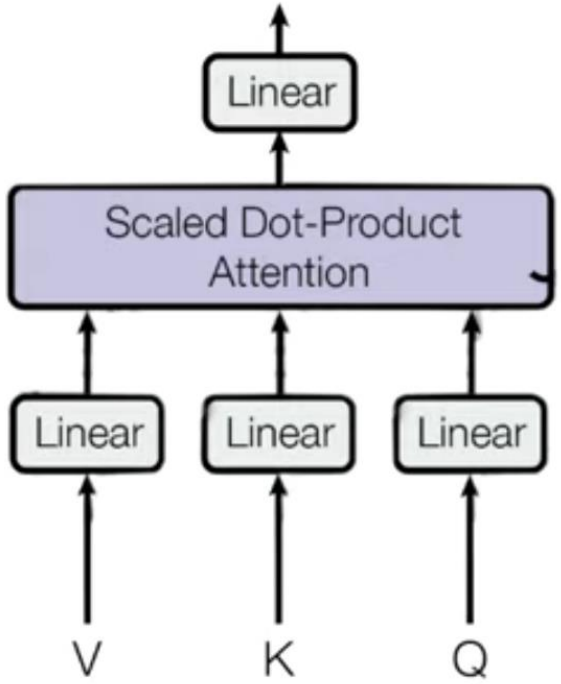
ITIC systems

Attention Matrix Generation

Q = What we are looking for
K = What can we offer
V= What we actually offer

Scaled Dot-Product Attention



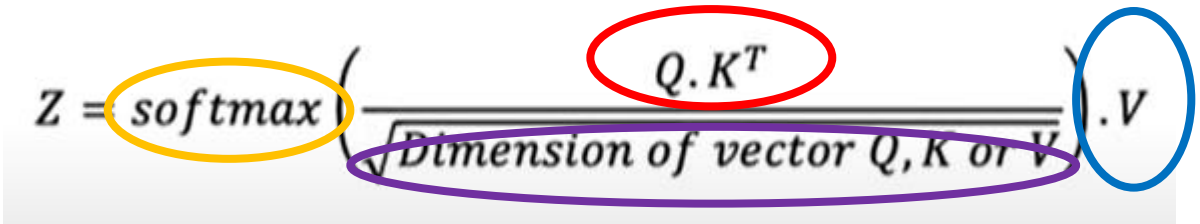


$$Z = softmax\left(\frac{Q.K^T}{\sqrt{Dimension\ of\ vector\ Q, K\ or\ V}}\right).V$$

ITIC systems

Attention Matrix Generation

Q = What we are looking for
K = What can we offer
V= What we actually offer



1- To Comparison between Query and Key vectors

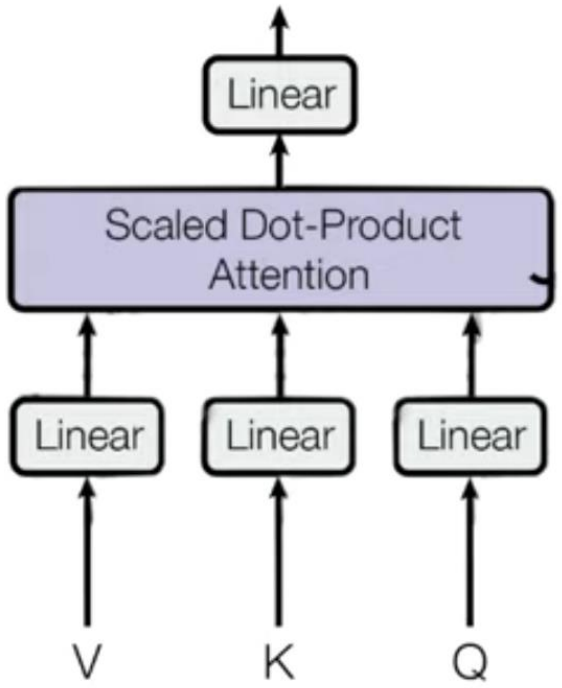
2- Divide to constant in order to better stabilize the value and better training

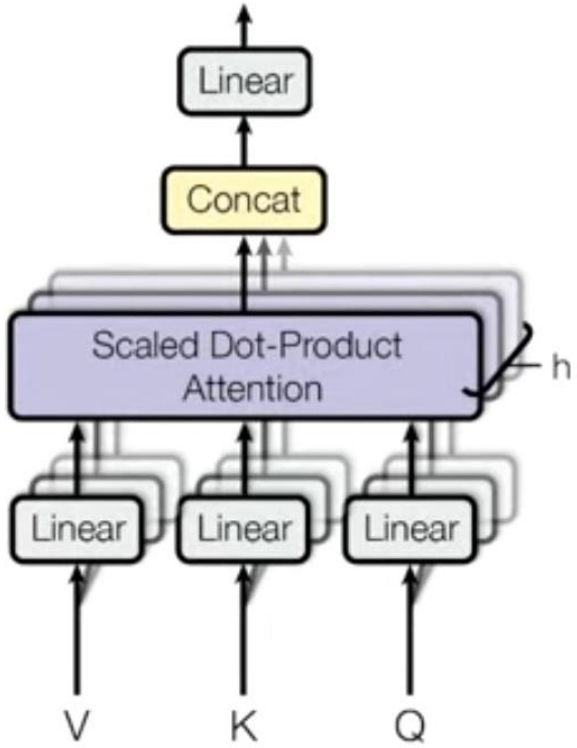
3- To provide Attention Matrices with values like probability

4- To provide word new vector by considering attention matrices

ITIC systems

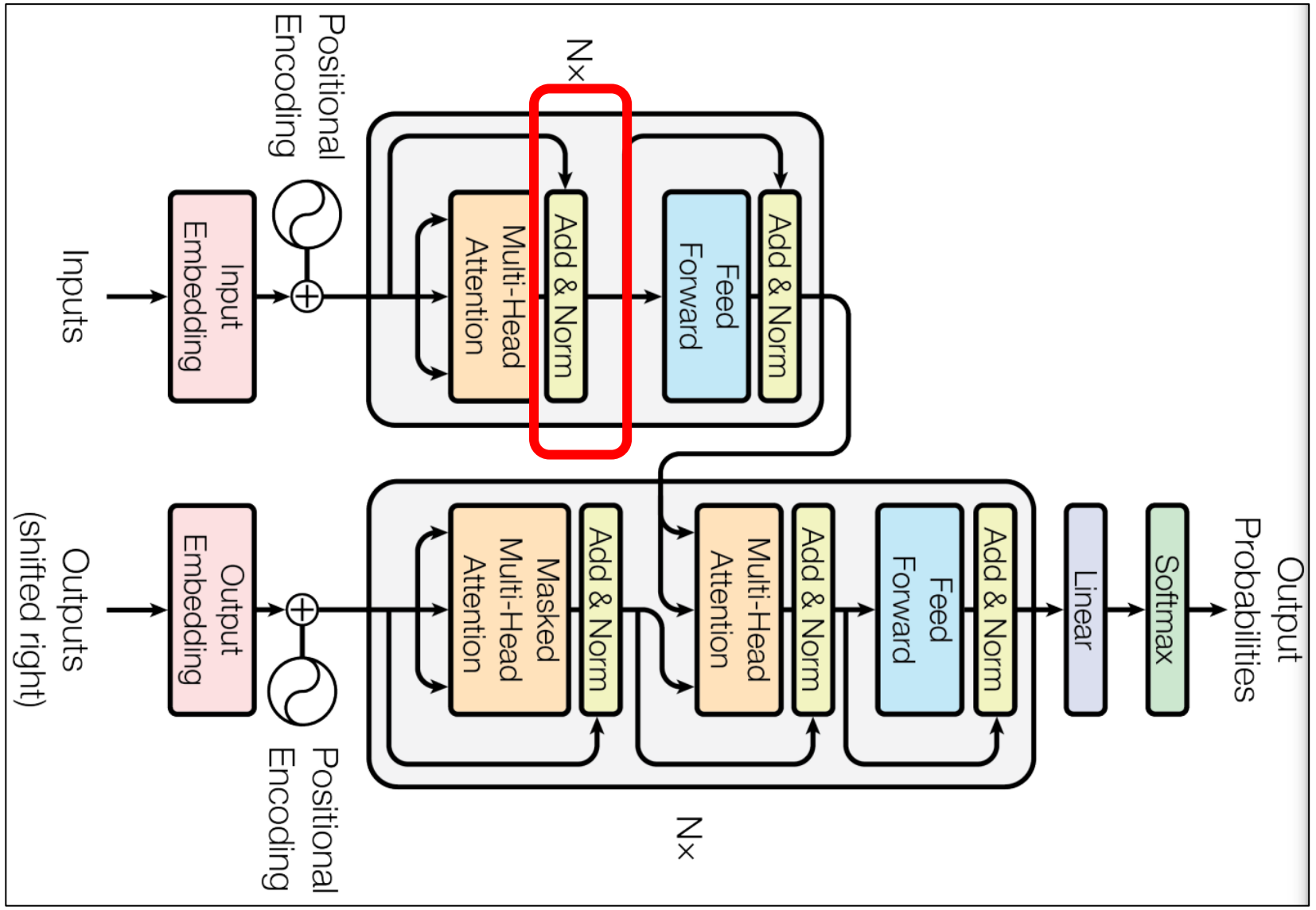
Multi Head Attention





ITIC systems

Transformer Flow – Residual Connection and Normalizing

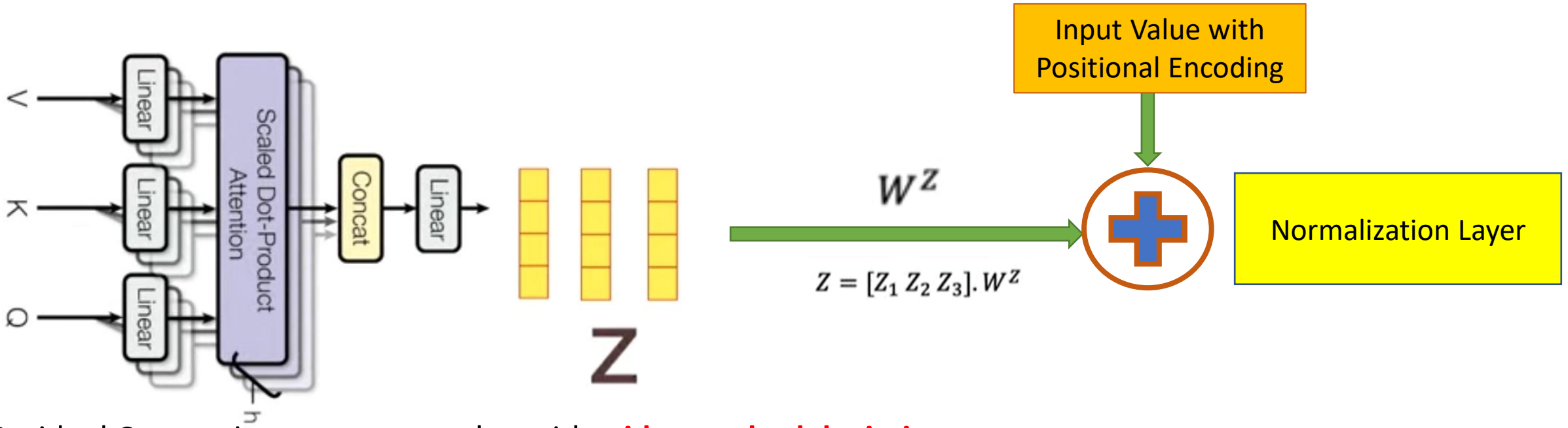


ITIC systems

Residual Connection and Normalizing

Add the result of Multi-Head attention layer with Input value alongside residual connection and perform Normalization Layer

For very deep network the back propagation eventually leads to very small gradients = > **no parameters update properly and network does not learn**



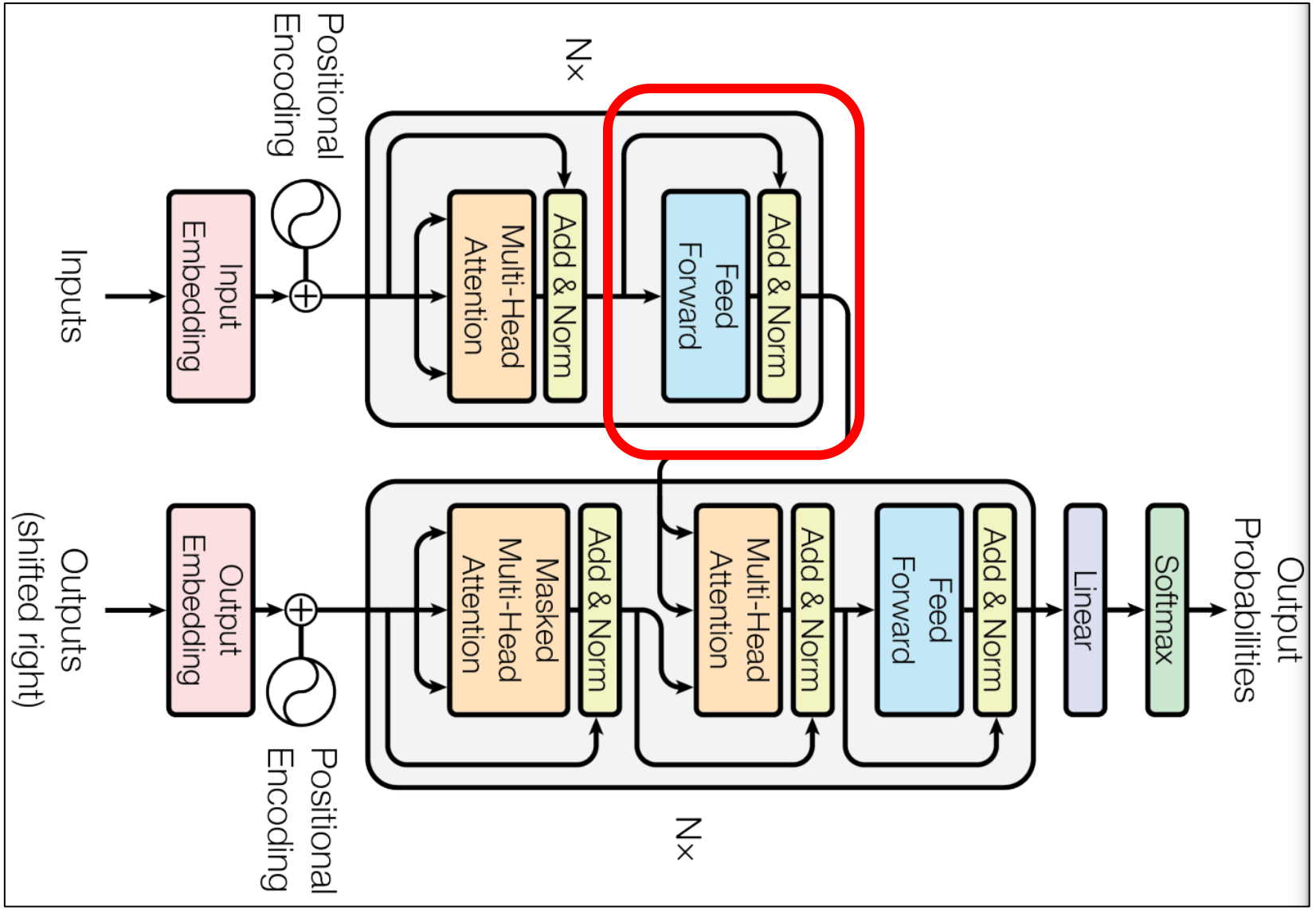
Residual Connection may cause value with **wide standard deviation**

Normalization layer is used for **perform stable training** by centralized around ZERO **with Standard deviation**

ITIC systems

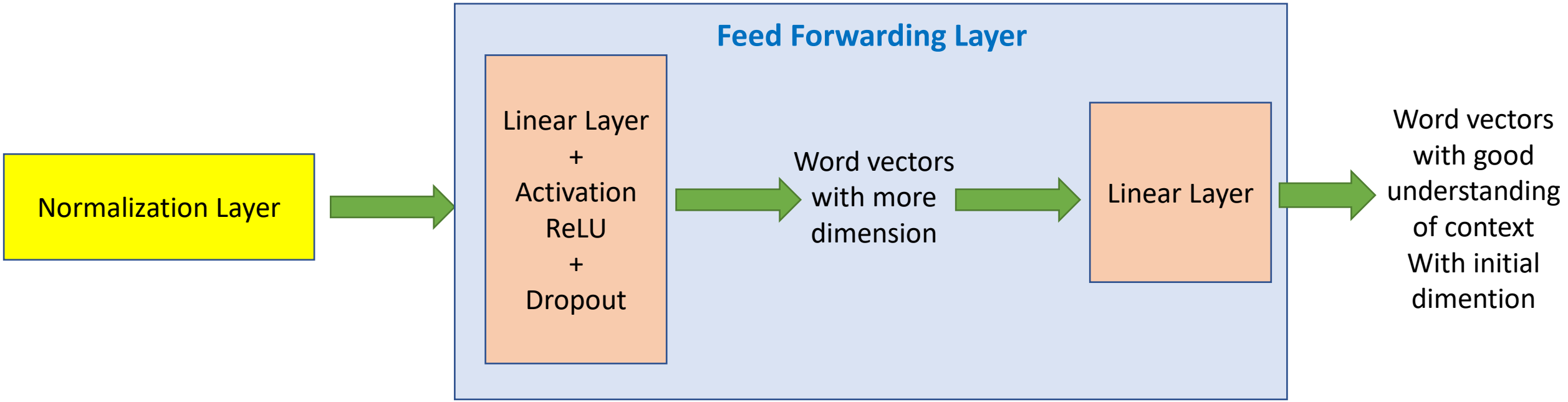
Transformer Flow – Feed Forwarding

Residual Connection and Normalizing



Transformer Flow – Feed Forwarding

Residual Connection and Normalizing



Linear Layer is going to help better interaction among Heads **because of just doing concatenation (pushing all Heads to interact with eachother cross the layers)**

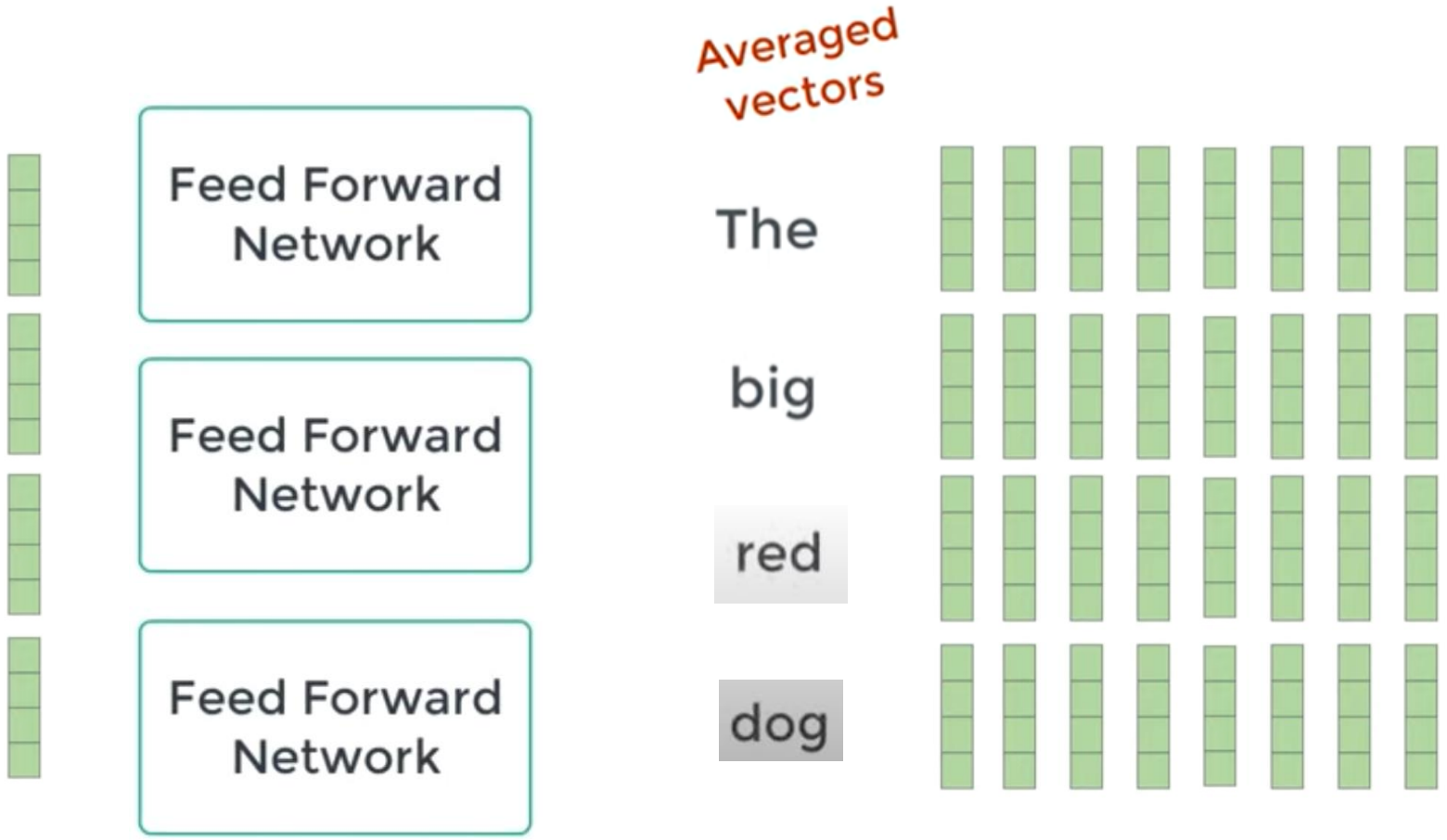
Relu activation function helps the **network understand better more complex pattern**

Drpoout randomly turns off neurons within feed forward layer in order to **network better able to generalize and to avoid memorize a specific kinf of pattern**

ITIC systems

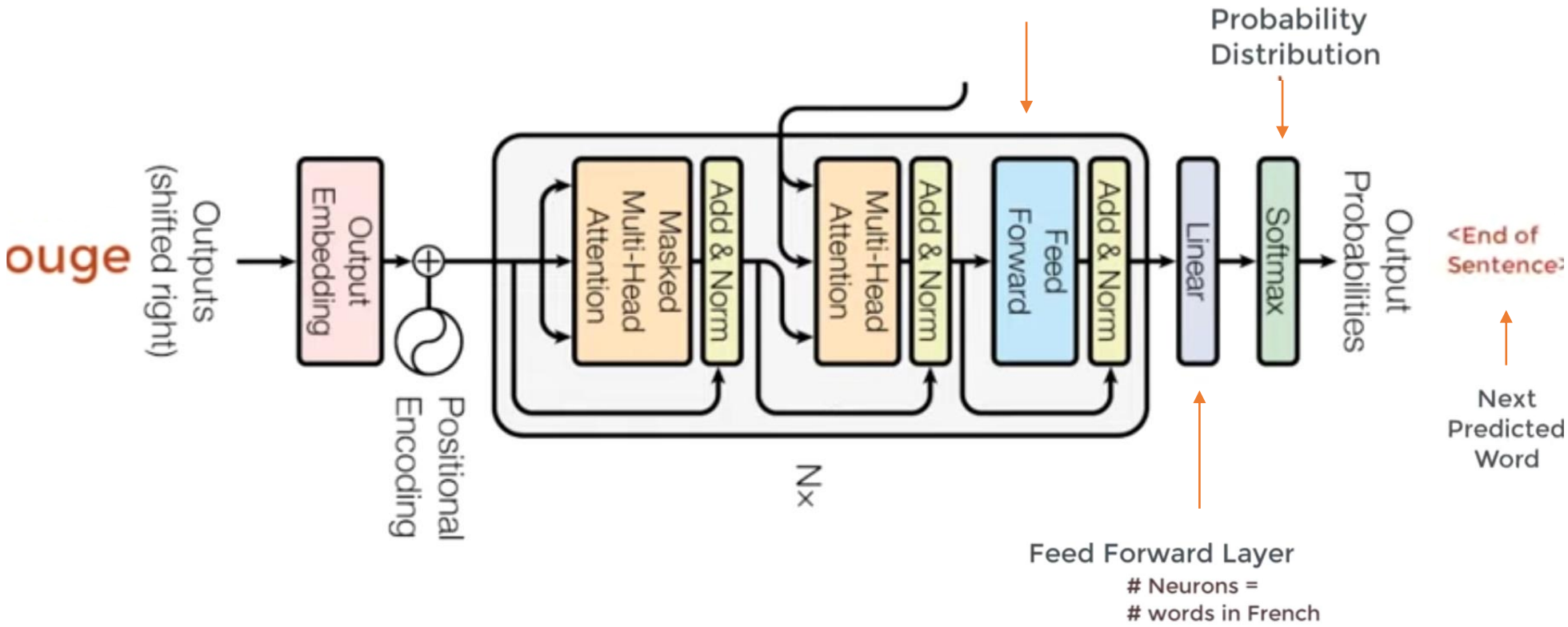
Transformer Flow – Feed Forwarding

Residual Connection and Normalizing



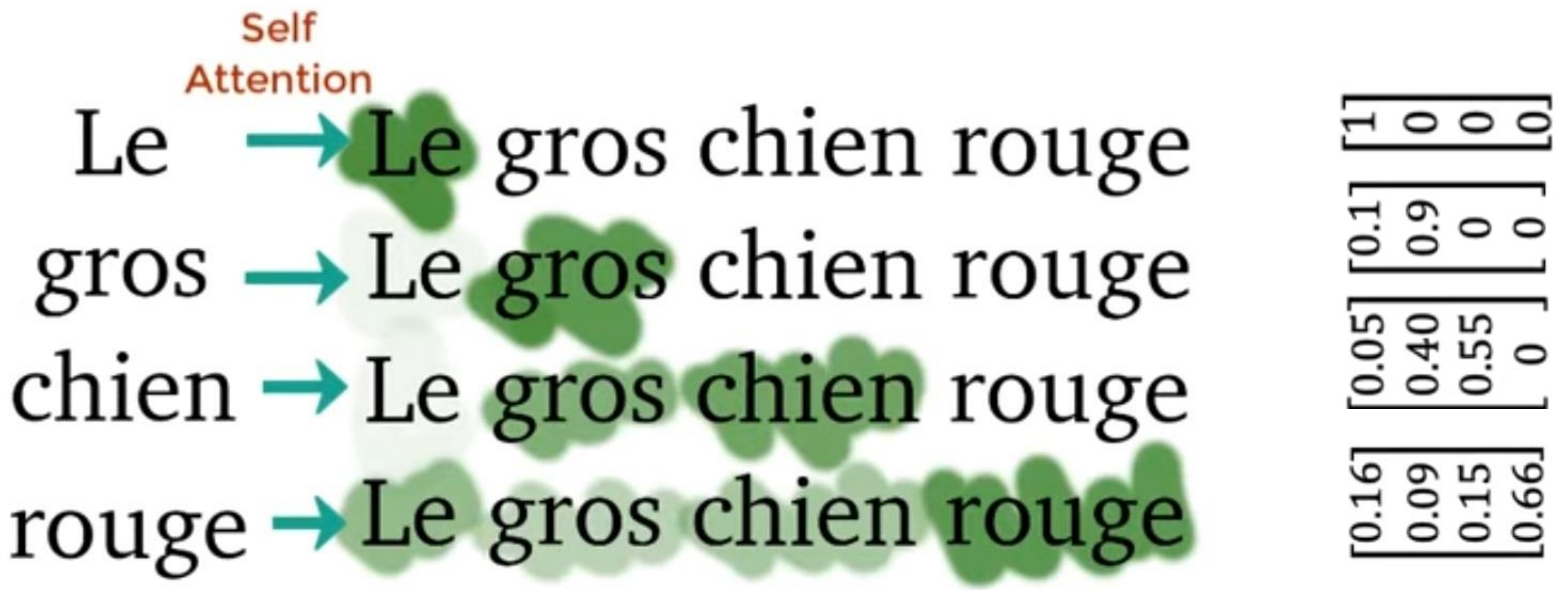
ITIC systems

Transformer Components



ITIC systems

Decoder



ITIC systems

MACQUARIE University

ITIC systems

Automatic Exam Marking Project

Phase 1: Exam Management Product

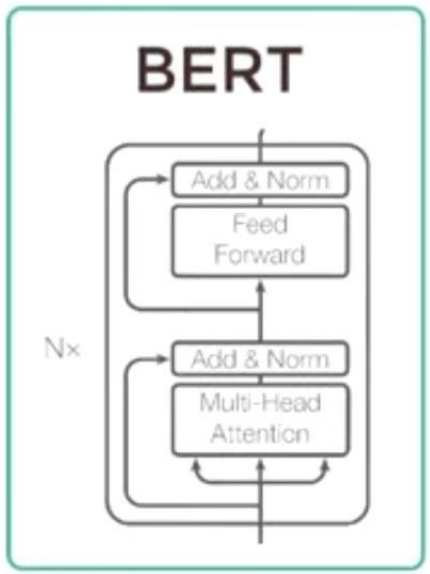
Product ML Engine

PART 3 : Bert Technical Detail

&

Fine Tuning Examples

Bidirectional Encoder Representation from Transformers

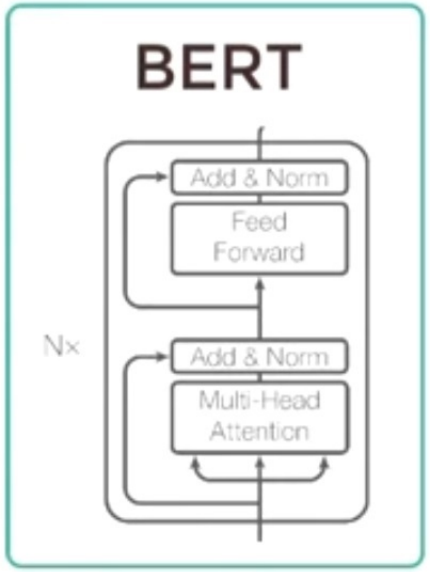


Problems to Solve

- Neural Machine Translation
- Question Answering
- Sentiment Analysis
- Text summarization

ITIC systems

Bidirectional Encoder Representation from Transformers



3 passes of explanation

Problems to Solve

- Neural Machine Translation
- Question Answering
- Sentiment Analysis
- Text summarization

Needs Language understanding

How to solve Problems (BERT Training)

- Pretrain BERT to understand langauge
- Fine tune BERT to learn specific task

ITIC systems

Bidirectional Encoder Representation from Transformers

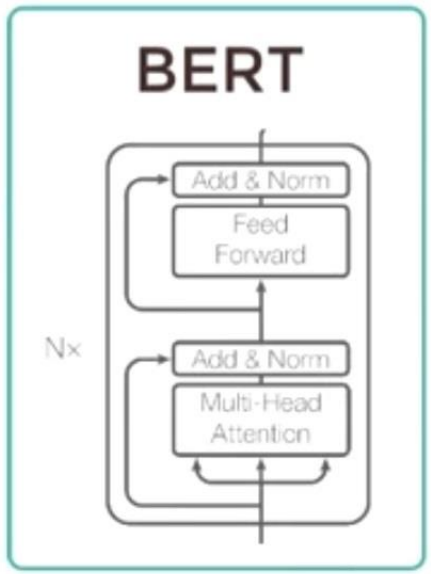
Pretraining (Pass 1) : “What is language? What is context?”

Masked Language Model (MLM)

The [MASK1] brown fox [MASK2] over the lazy dog.

Next Sentence Prediction (NSP)

A: Ajay is a cool dude.
B: He lives in Ohio



[MASK1] = quick
[MASK2]= jumped

Yes. Sentence B follows sentence A

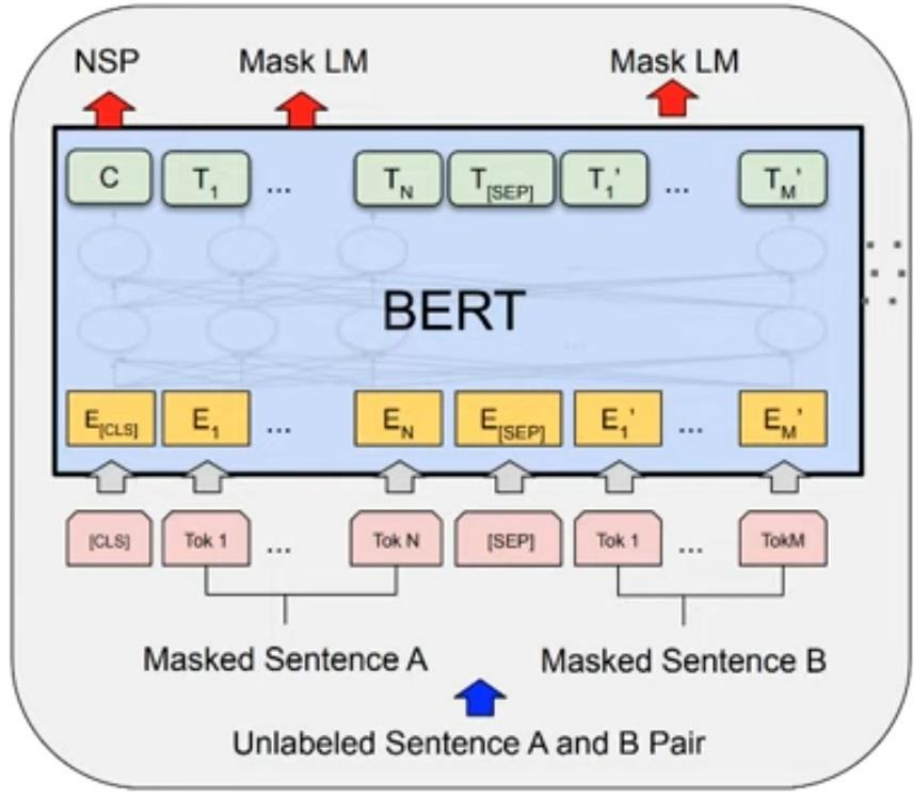
ITIC systems

Bidirectional Encoder Representation from Transformers

Pretraining (Pass 2)

Problems to train on simultaneously:

1. Masked Language Modeling (Mask LM)
2. Next Sentence Prediction (NSP)



Source: BERT: Pre-training of deep bidirectional Transformers for language understanding (Devlin et al., 2019)

ITIC systems

Masked LM

- Problem: Mask token never seen at fine-tuning
- Solution: 15% of the words to predict, but don't replace with `[MASK]` 100% of the time. Instead:
- 80% of the time, replace with `[MASK]`

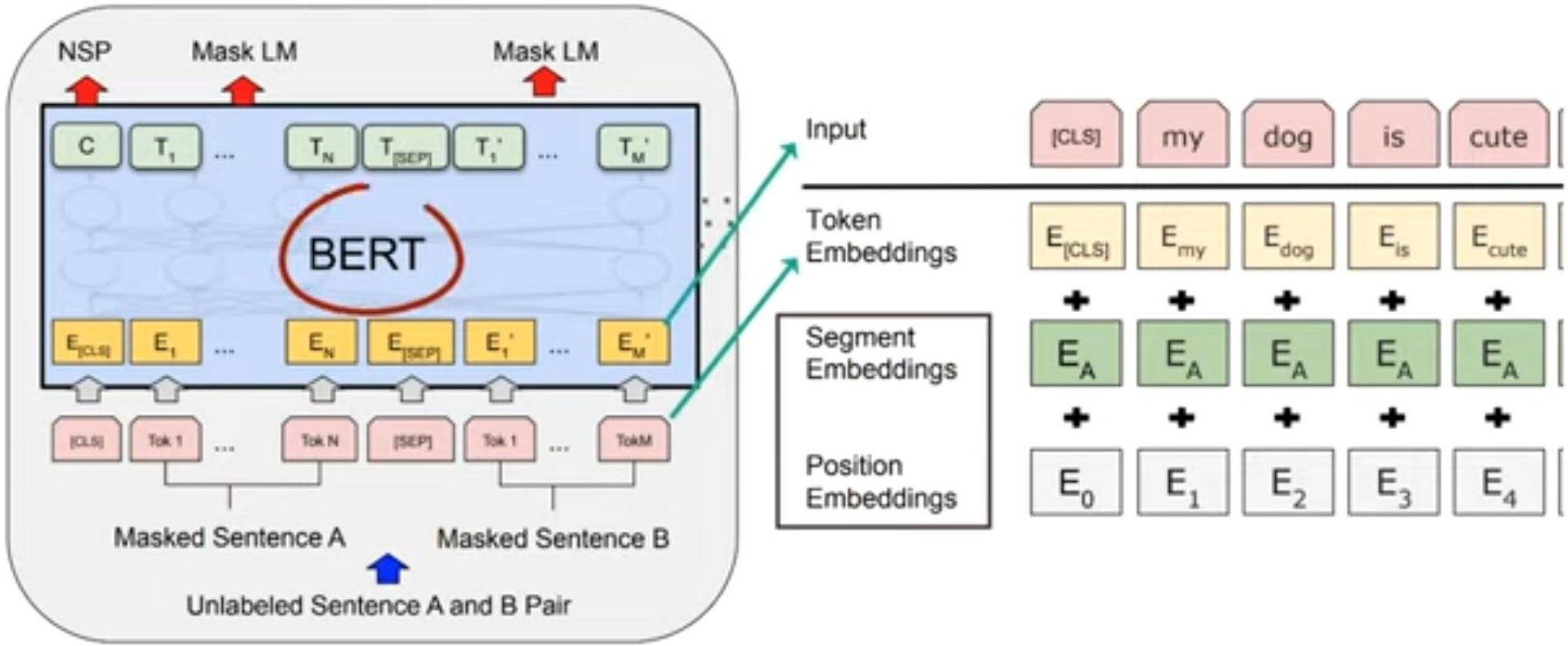
  `went to the store → went to the [MASK]`
- 10% of the time, replace random word

  `went to the store → went to the running`
- 10% of the time, keep same

  `went to the store → went to the store`

Bidirectional Encoder Representation from Transformers

Pretraining (Summary)



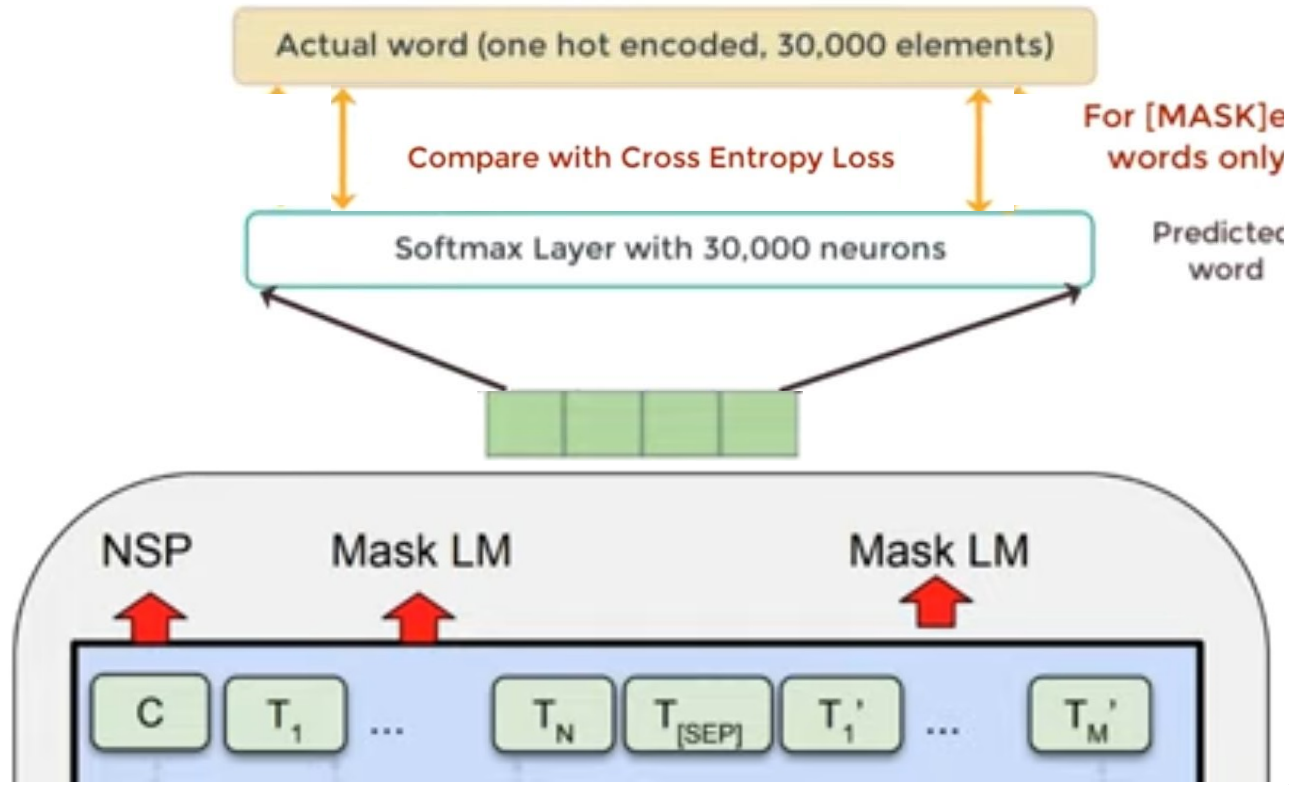
ITIC systems

Bidirectional Encoder Representation from Transformers

Pretraining (Pass 3)

Word vectors T_i have the same size.

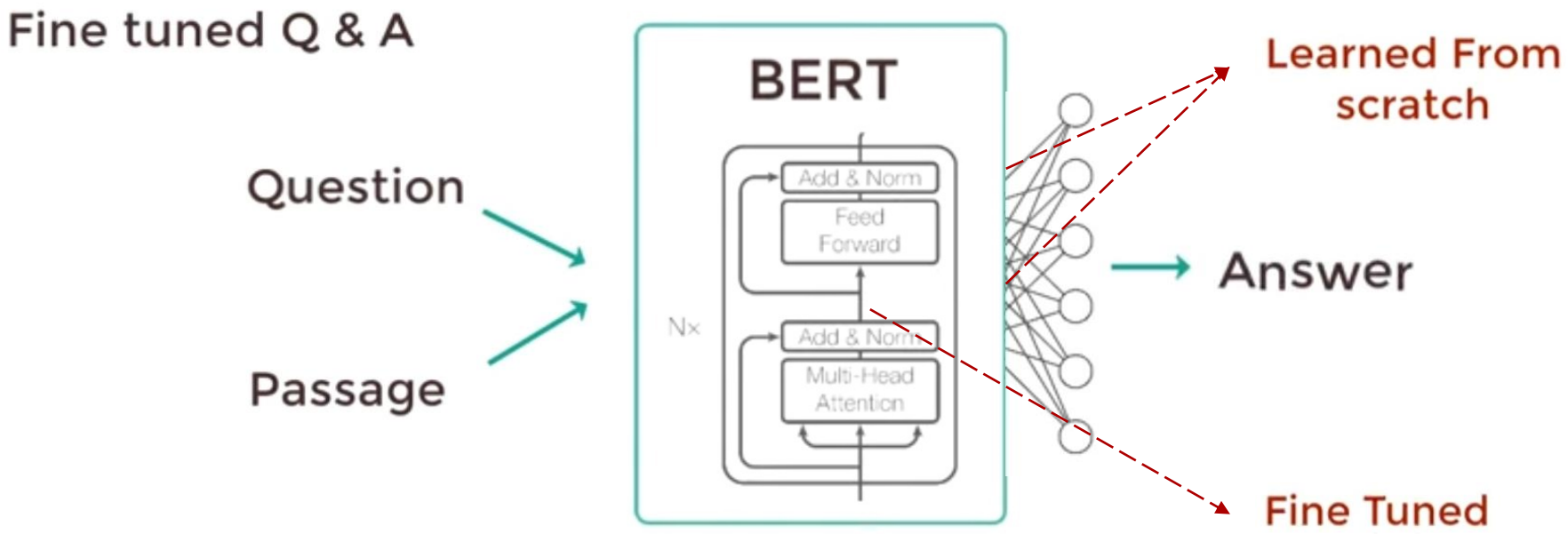
Word vectors T_i are generated simultaneously



ITIC systems

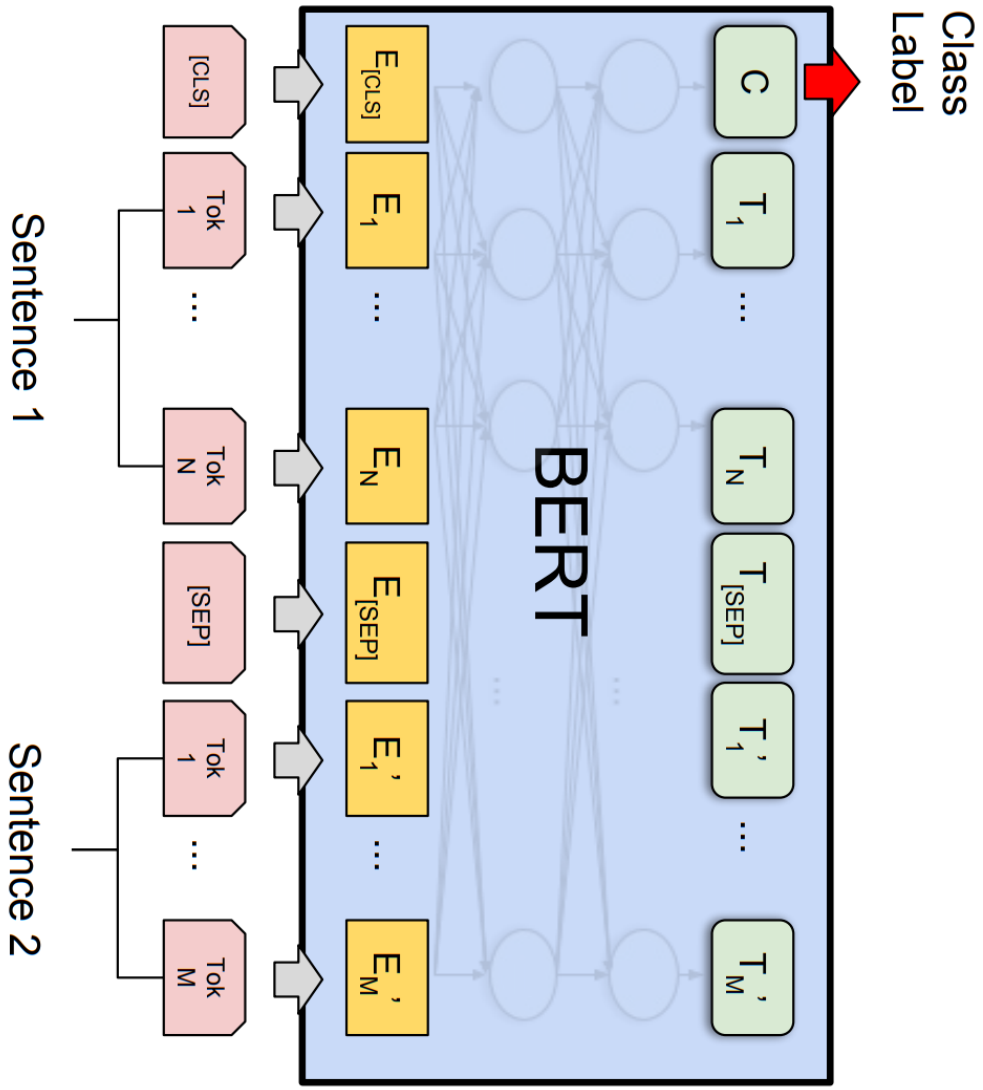
Bidirectional Encoder Representation from Transformers

Fine Tuning (Pass 1): “How to use language for specific task?”



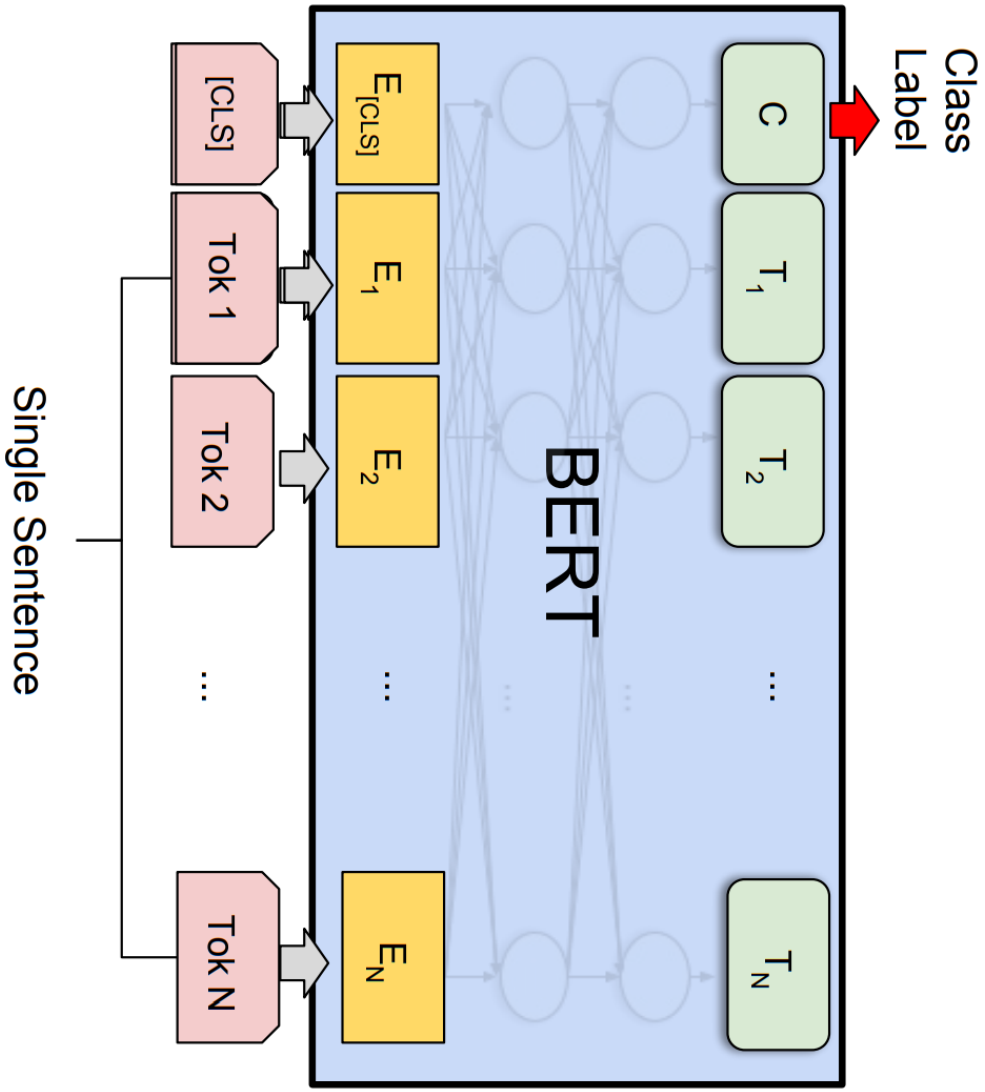
Bert Tuning For

(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG

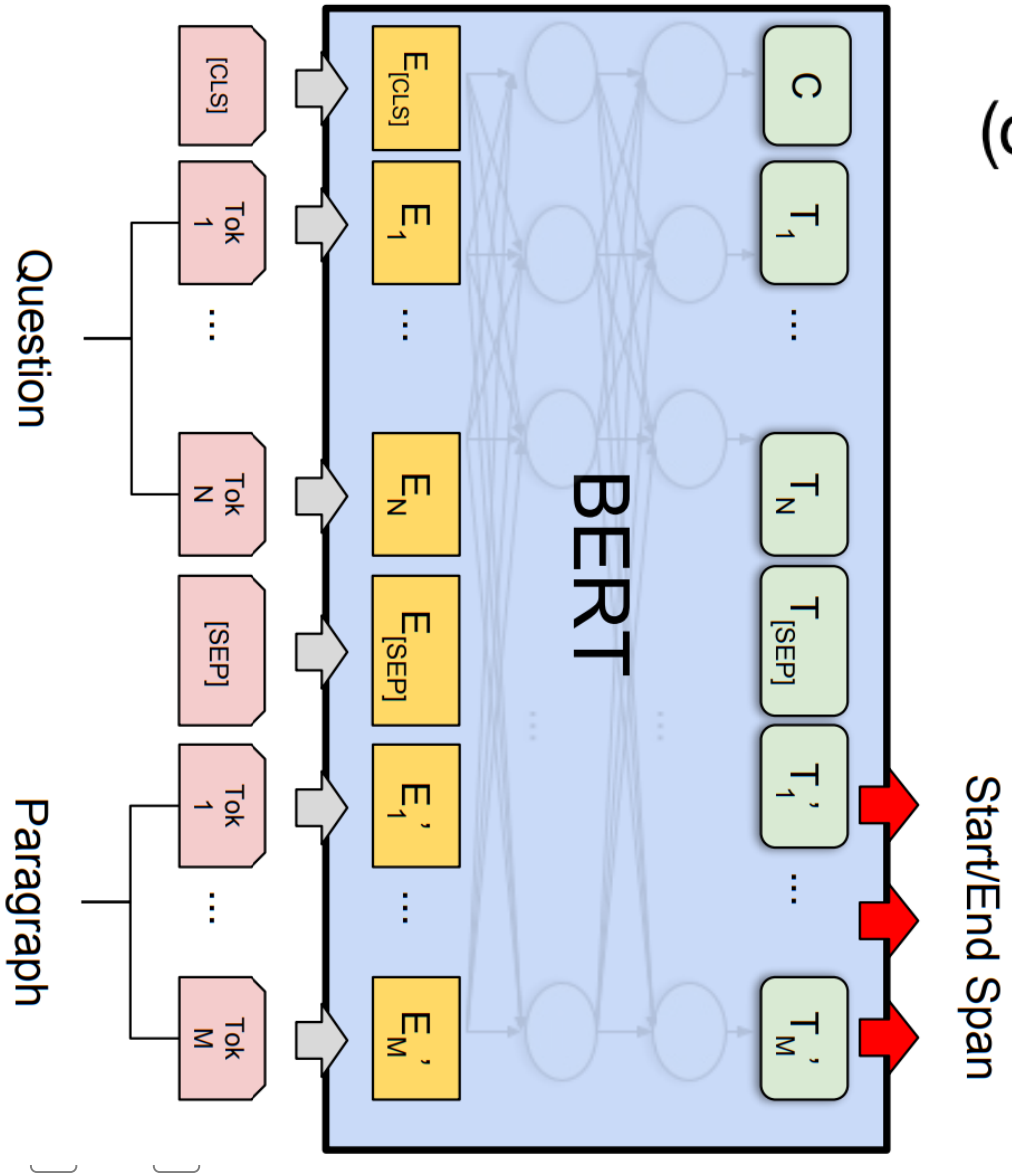


Bert Tuning For

(b) Single Sentence Classification Tasks:
SST-2, CoLA



Bert Tuning For

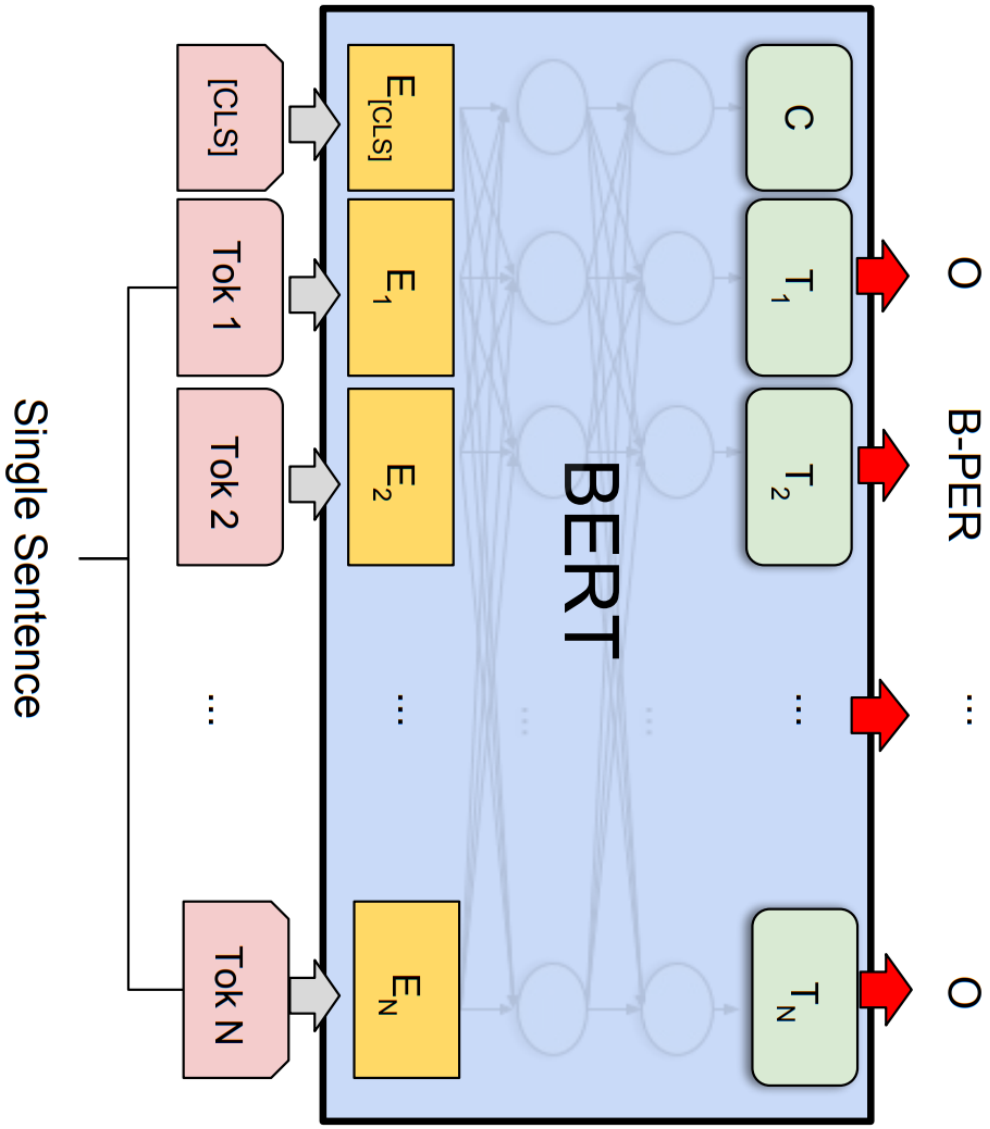


(c) Question Answering Tasks:
SQuAD v1.1

ITIC systems

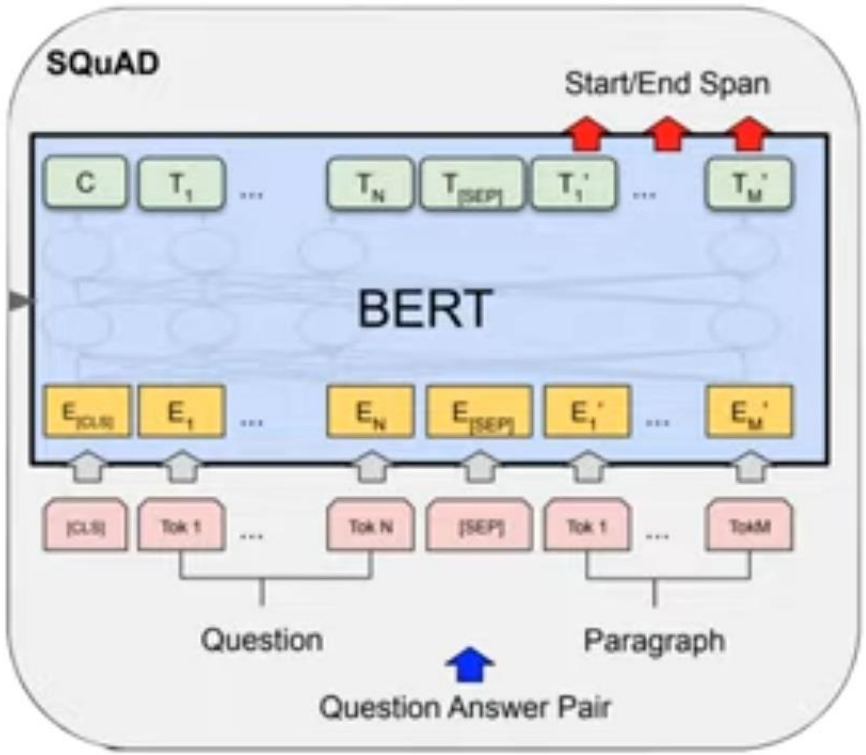
Bert Tuning For

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER



Bidirectional Encoder Representation from Transformers

Fine Tuning (Summary)



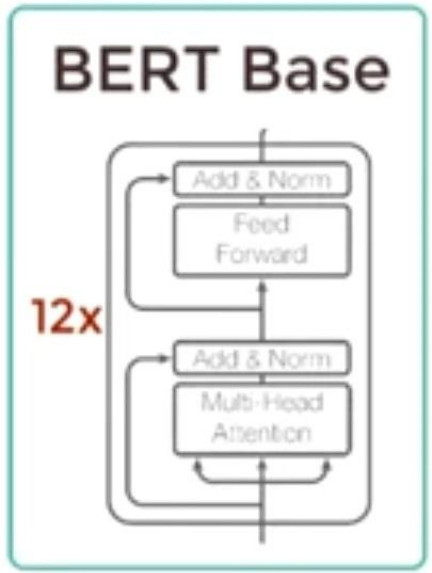
"Stanford Question & Answer Dataset"

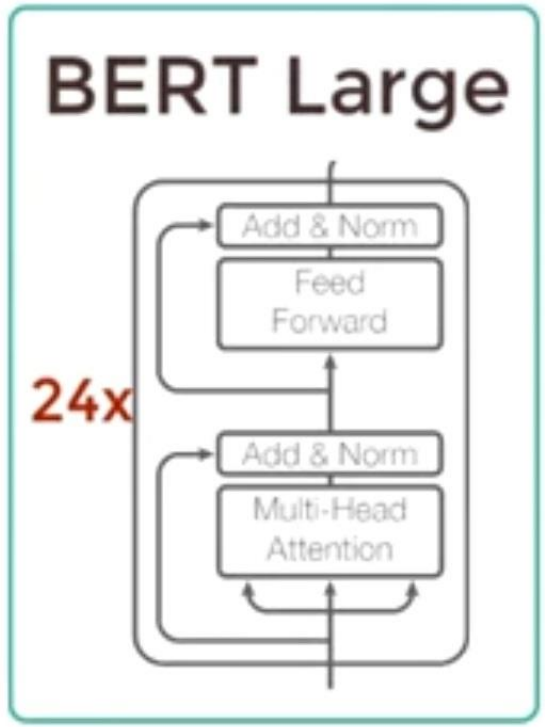
- 30 minute training on single cloud TPU with 91% F1 score.

ITIC systems

Bidirectional Encoder Representation from Transformers

Performance





340M Parameters

Model Details

- Data: Wikipedia (2.5B words) + BookCorpus (800M words)
- Batch Size: 131,072 words (1024 sequences * 128 length or 256 sequences * 512 length)
- Training Time: 1M steps (~40 epochs)
- Optimizer: AdamW, 1e-4 learning rate, linear decay
- `BERT-Base`: 12-layer, 768-hidden, 12-head
- `BERT-Large`: 24-layer, 1024-hidden, 16-head
- Trained on 4x4 or 8x8 TPU slice for 4 days

MACQUARIE University

ITIC systems

Automatic Exam Marking Project

Phase 1: Exam Management Product

Product ML Engine

PART 4 : New Task

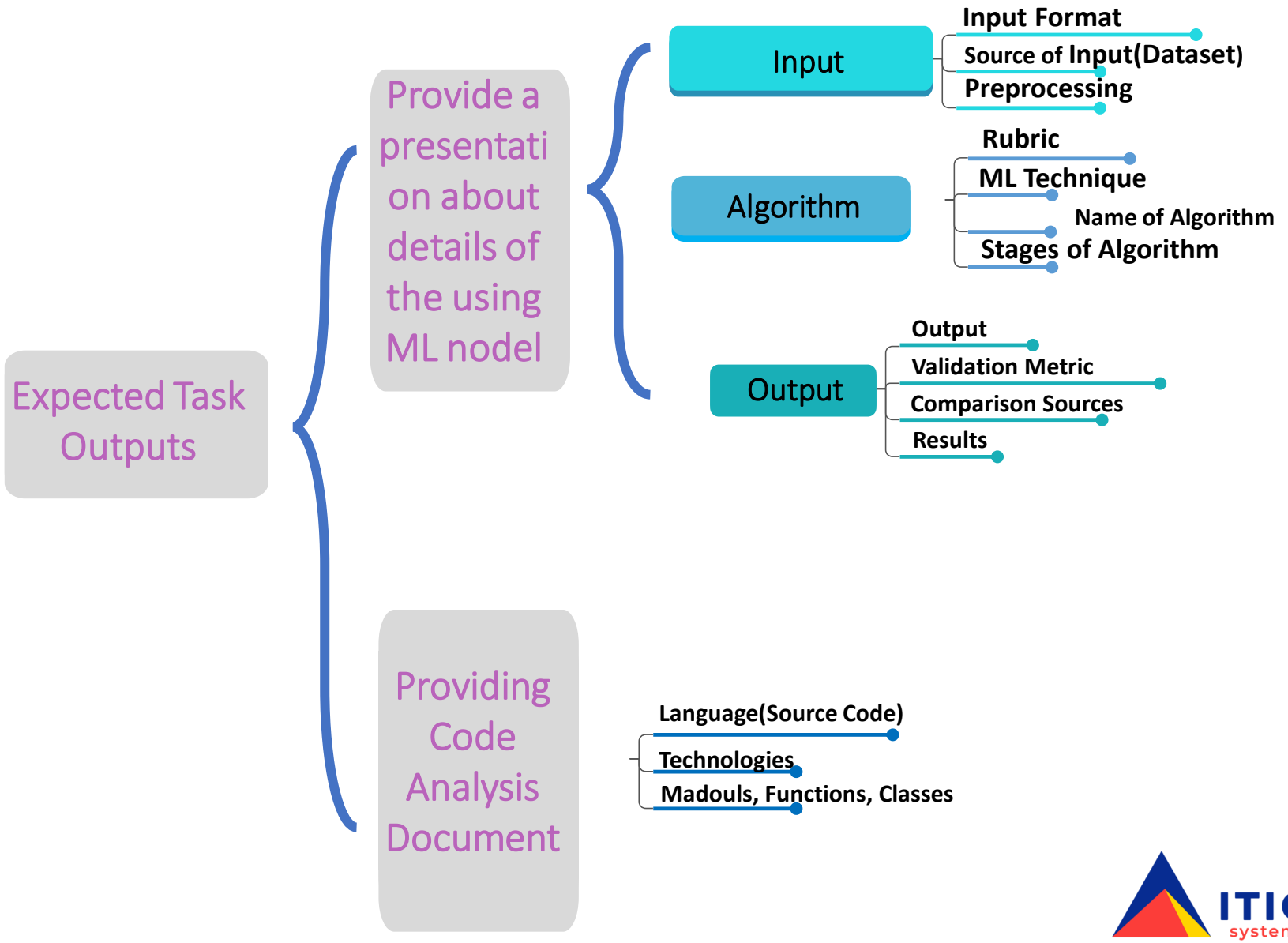
&

Code Analysis Definition

Task 4: Existing GAN For Generation Essay DataSet

Research & Code Analysis

Park, Y. H., Choi, Y. S., Park, C. Y., & Lee, K. J. (2022). EssayGAN: Essay Data Augmentation Based on Generative Adversarial Networks for Automated Essay Scoring. *Applied Sciences*, *12*(12), 5803.



Task 5: Bert Pre-traing and Fine Tuning Code Analysis

Expected Task Outputs

Providing Code Analysis Document For Bert Pre-training

Python Code:

queries. https://github.com/jina-ai/clip-as-service

Providing Code Analysis Document For Existing Bert Fine Tuning Phases for Specific Task

Test Summarization
Question Generating
Answers Marking

Task 5 & 7 & 8 & 9: Bert Pre-traing and Fine Tuning Code Analysis

Expected Task Outputs

Providing Code Analysis Document For Bert Pre-training

Python Code:

queries. https://github.com/jina-ai/clip-as-service

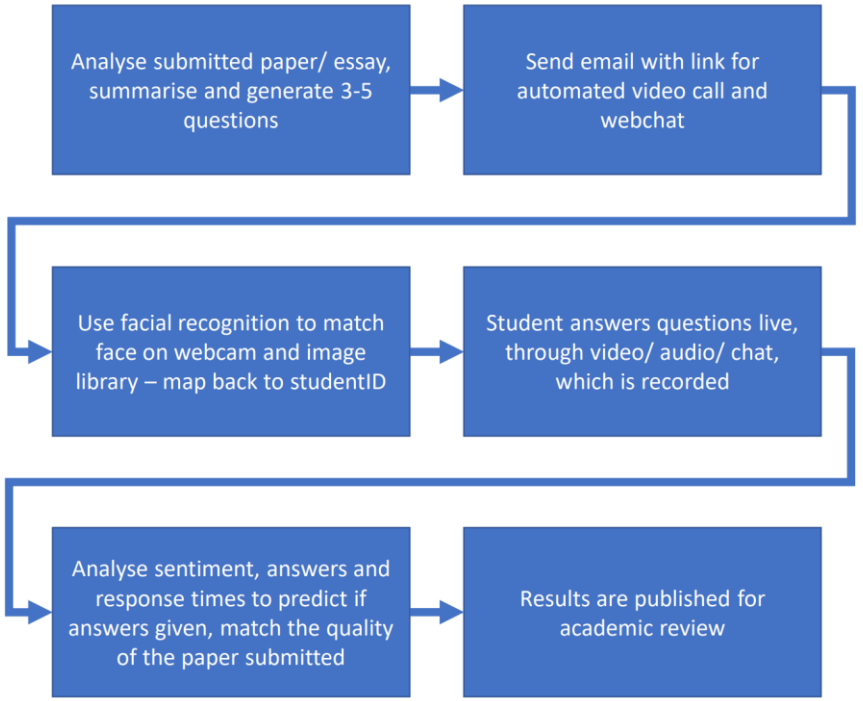
Providing Code Analysis Document For Existing Bert Fine Tuning Phases for Specific Task

Test Summarization
Question Generating
Answers Marking

ITIC systems

Task 10: Analyzing Existing methods for plagiarism

– Find Similarities between existing source of question



ITIC systems

Code Analysis – Required Output

ITEM	Description per Line	Input	Output	Parameters	Variable Description	Using in which Part of Code	Used Item		
Function									
Classes									
Libraries									

Data Base	Data Structure	Relations & Keys	Data Type	Data Scale					

MACQUARIE University

ITIC systems

Automatic Exam Marking Project

Phase 1: Exam Management Product

Product ML Engine

PART 5 : References

References

Bert Overview
https://www.youtube.com/watch?v=xI0HHN5XKDo

Bert Paper Overview
https://www.youtube.com/watch?v=-9evrZnBorM

Transformer Neural Networks - EXPLAINED! (Attention is all you need)
https://www.youtube.com/watch?v=TQQlZhbC5ps

Stanford Bert
https://www.youtube.com/watch?v=knTc-NQSjKA

Stanford Transfomer and self attention
https://www.youtube.com/watch?v=5vcj8kSwBCY

Self attention in Transformer with code
https://www.youtube.com/watch?v=QCJQG4DuHT0&list=PLTl9hO2Oobd97qfWC40gOSU8C0iu0m2l4

Multi Head Attention in Transformer Neural Networks with Code!
https://www.youtube.com/watch?v=HQn1QKQYXVg

References

Architecture connection & Normalization layer
https://www.youtube.com/watch?v=G45TuC6zRf4&list=PLTl9hO2Oobd97qfWC40gOSU8C0iu0m2l4&index=5

Positional Encoding in Transformer Neural Networks Explained
https://www.youtube.com/watch?v=ZMxVe-HK174&list=PLTl9hO2Oobd97qfWC40gOSU8C0iu0m2l4&index=4

Blowing up the Transformer Encoder!
https://www.youtube.com/watch?v=QwfuoNhjbkI&list=PLTl9hO2Oobd97qfWC40gOSU8C0iu0m2l4

Blowing up Transformer Decoder architecture
https://www.youtube.com/watch?v=ekg-hoob0SM&list=PLTl9hO2Oobd97qfWC40gOSU8C0iu0m2l4&index=8

NLP Project in 20 Minutes Using BERT
https://www.youtube.com/watch?v=iiwEW-sg9KE

Sample Topic Modeling problem with using bert
https://www.youtube.com/watch?v=TLPmlVeEf1k