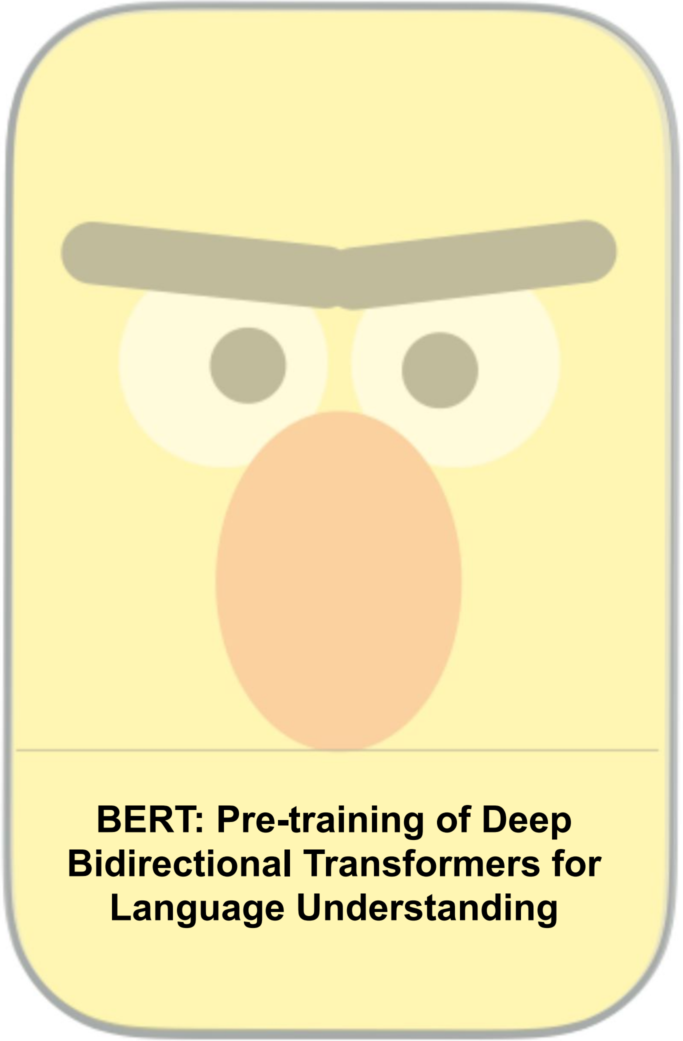


BERT

David Gamaliel Arcos Bravo

Que es BERT?

BERT (Bidirectional Encoder Representations from Transformers) es un modelo de lenguaje basado en transformers desarrollado por Google en 2018 que revolucionó el procesamiento del lenguaje natural. Su importancia radica en su capacidad para capturar relaciones contextuales y semánticas de las palabras en textos.

BERT es efectivo en una variedad de tareas de procesamiento del lenguaje, como clasificación de texto, respuesta a preguntas y análisis de sentimiento. Su capacidad para comprender el contexto y la estructura del lenguaje mejora significativamente la precisión y el rendimiento de los sistemas de procesamiento del lenguaje natural.

BERT Paper Reference

Contexto

Arquitecturas precedentes

- Redes neuronales recurrentes (RNN)
- Gated Recurrent Units (GRU)
- Bidirectional LSTM
- Embeddings from Language Models (ELMo)

Contexto

Problemas

Redes neuronales recurrentes (RNN):

- Difícil captura de dependencias a largo plazo debido a la propagación de gradientes a través del tiempo, resultando en la pérdida de información relevante en secuencias largas.

Gated Recurrent Units (GRU):

- Aunque las GRU mejoran la capacidad de las RNN para capturar dependencias a largo plazo, aún pueden tener dificultades para modelar secuencias complejas con múltiples relaciones y dependencias contextuales.

Contexto

Problemas

Bidirectional LSTM:

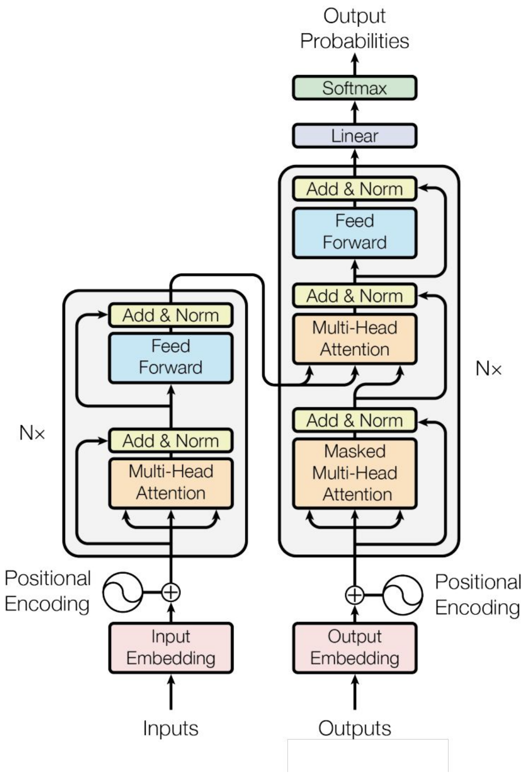
- El uso de LSTMs bidireccionales puede aumentar la complejidad computacional, lo que puede requerir más recursos y tiempo de entrenamiento en comparación con arquitecturas más simples como las RNN.

Embeddings from Language Models (ELMo):

- Aunque ELMo captura información contextualizada de una manera más efectiva que las representaciones estáticas de palabras, aún se basa en un modelo de lenguaje unidireccional, lo que puede limitar su capacidad para capturar el contexto bidireccional en el texto.

BERT

Encoder



GPT

Decoder

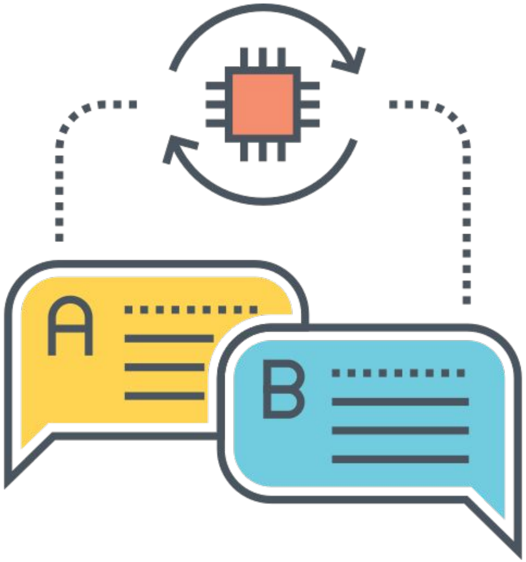
Aplicaciones

01 Respuesta a preguntas
(Question Answering)

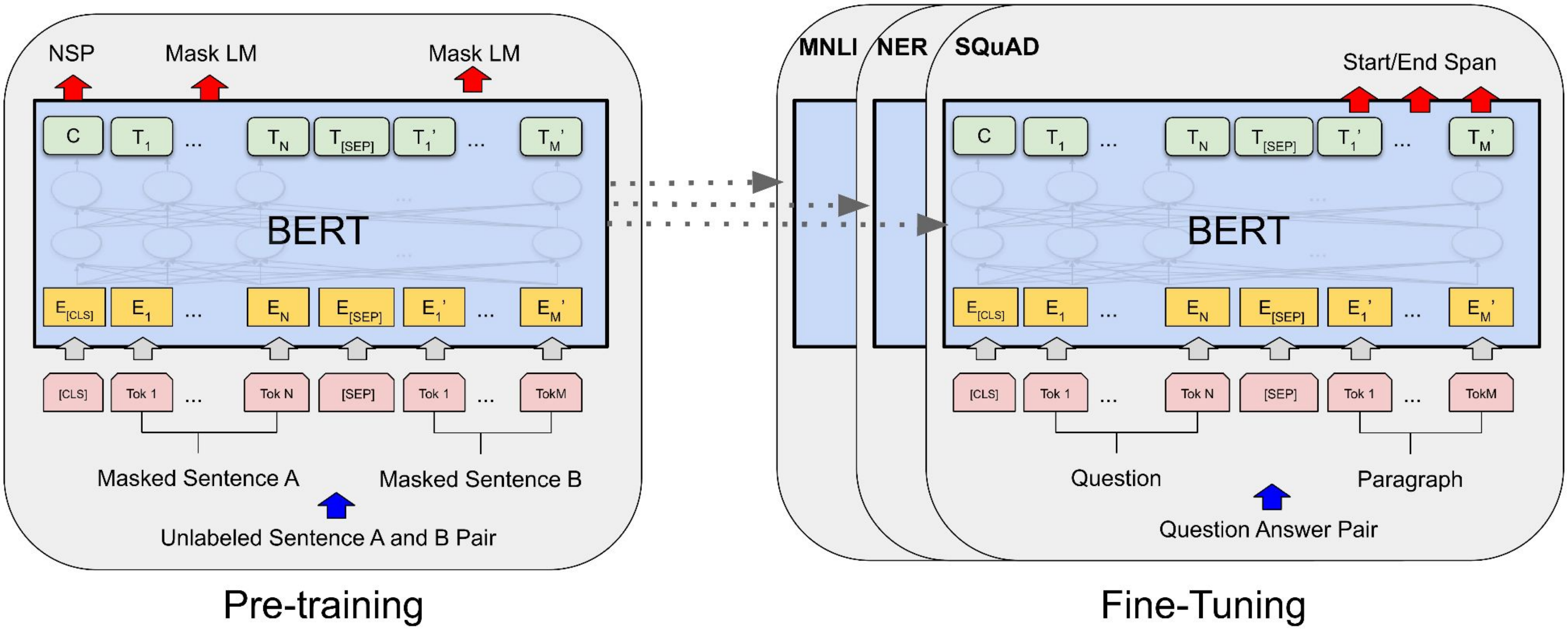
02 Inferencia de lenguaje natural
(Natural Language Inference)

03 Clasificación de texto
(Text classification)

04 Análisis de sentimientos
(Sentiment Analysis)



BERT Phases



NSP
Mask LM
Mask LM
C
T_1
T_N
$T_{[SEP]}$
T_1'
T_M'
BERT
$E_{[CLS]}$
E_1
E_N
$E_{[SEP]}$
E_1'
E_M'
[CLS]
Tok 1
Tok N
[SEP]
Tok 1
TokM
Masked Sentence A
Masked Sentence B
Unlabeled Sentence A and B Pair
Pre-training
MNLI
NER
SQuAD
Start/End Span
C
T_1
T_N
$T_{[SEP]}$
T_1'
T_M'
BERT
$E_{[CLS]}$
E_1
E_N
$E_{[SEP]}$
E_1'
E_M'
[CLS]
Tok 1
Tok N
[SEP]
Tok 1
TokM
Question
Paragraph
Question Answer Pair
Fine-Tuning

BERT Input Representation

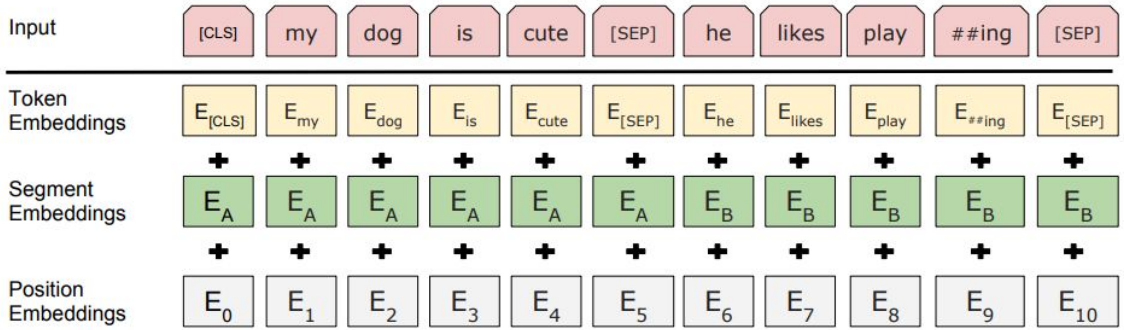


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

BERT Pretraining Tasks

Task #1: Masked LM

...

Task #2: Next Sentence Prediction (NSP)

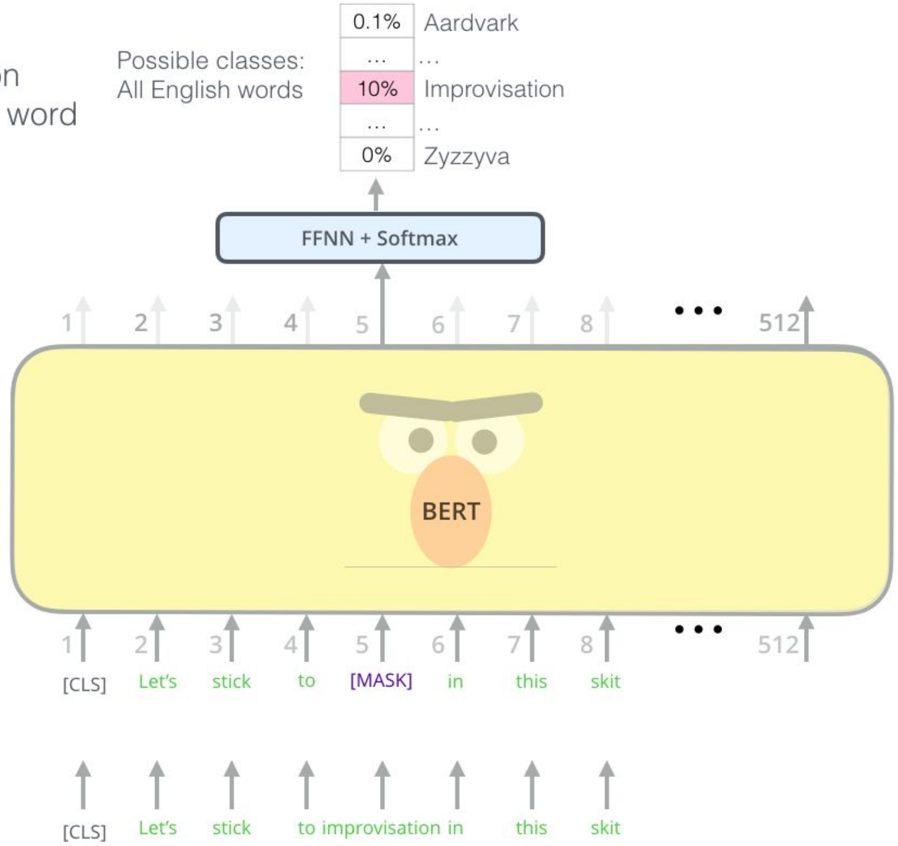
...

BERT Pretraining Tasks – Masked LM

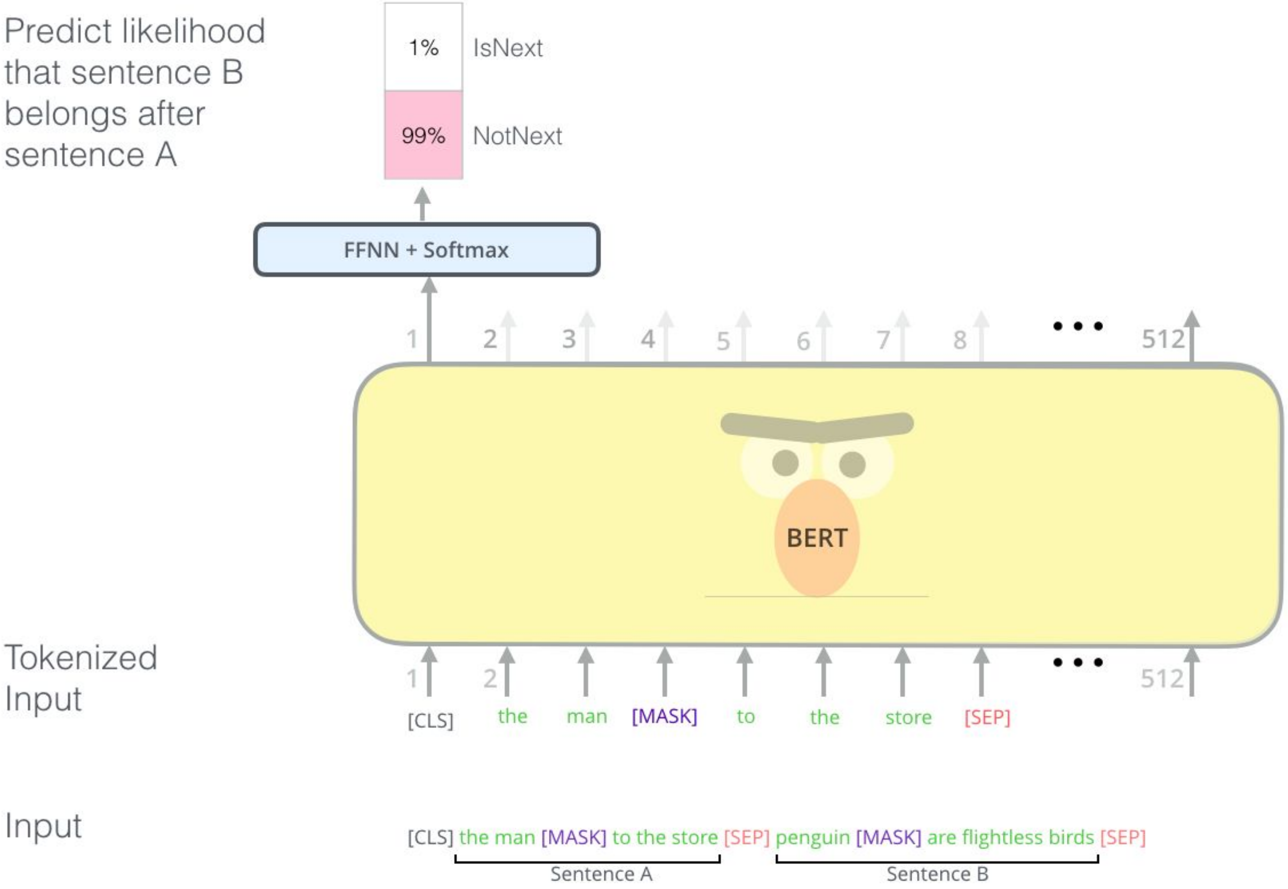
Use the output of the masked word's position to predict the masked word

Randomly mask 15% of tokens

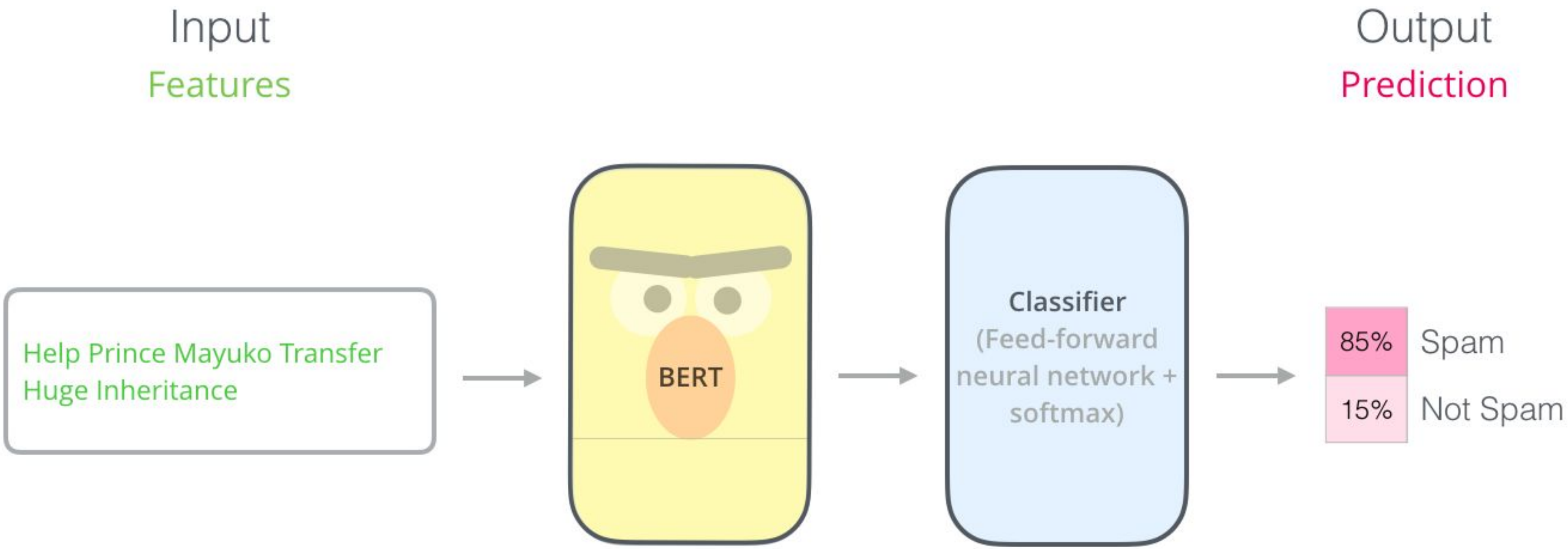
Input



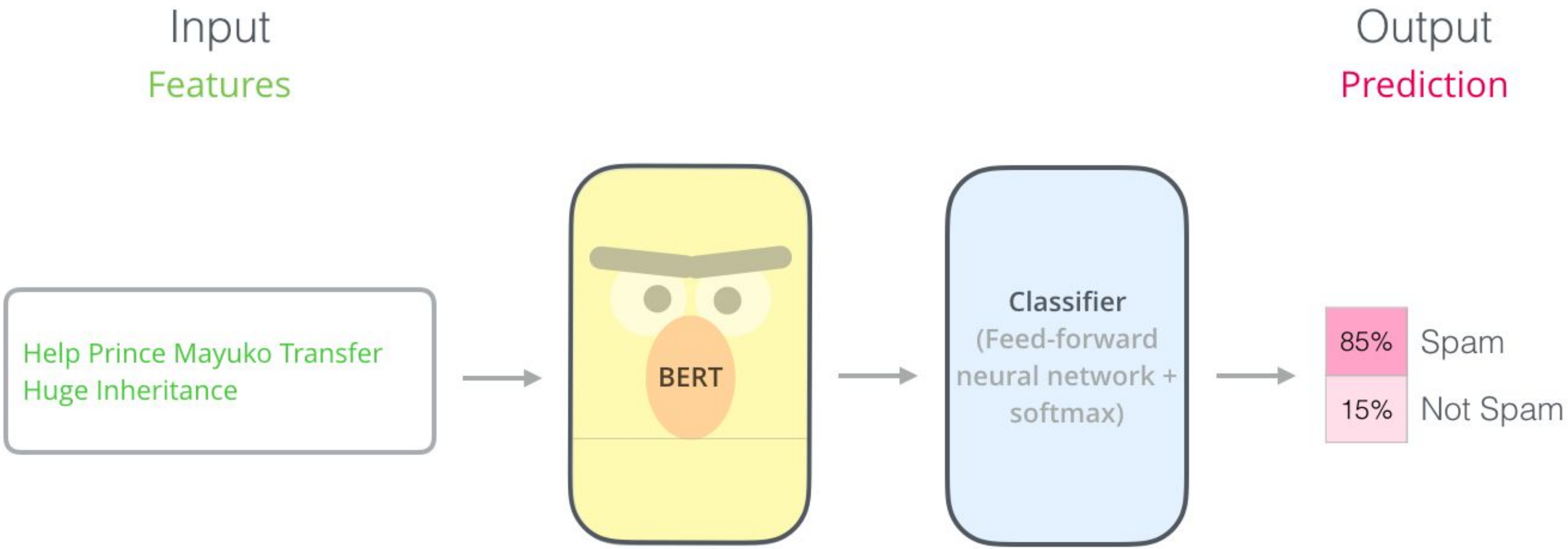
BERT Pretraining Tasks – NSP



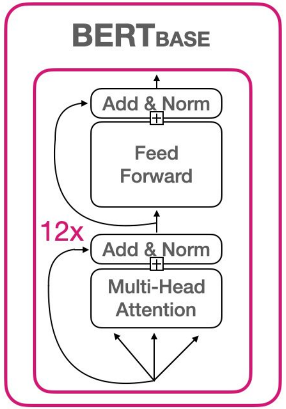
BERT Fine Tuning Examples



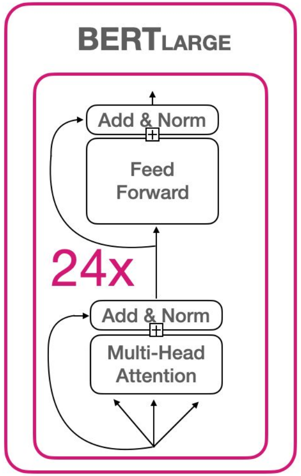
BERT Fine Tuning Examples



BERT Size & Architecture



110M Parameters



340M Parameters

Resultados

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

Table 1: GLUE Test results, scored by the evaluation server (`https://gluebenchmark.com/leaderboard`). The number below each task denotes the number of training examples. The "Average" column is slightly different than the official GLUE score, since we exclude the problematic WNLI set.[8] BERT and OpenAI GPT are single-model, single task. F1 scores are reported for QQP and MRPC, Spearman correlations are reported for STS-B, and accuracy scores are reported for the other tasks. We exclude entries that use BERT as one of their components.

Resultados

System	Dev EM	Dev F1	Test EM	Test F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
Published				
BiDAF+ELMo (Single)	-	85.6	-	85.8
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

Table 2: SQuAD 1.1 results. The BERT ensemble is 7x systems which use different pre-training checkpoints and fine-tuning seeds.

System	Dev EM	Dev F1	Test EM	Test F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	86.3	89.0	86.9	89.5
#1 Single - MIR-MRC (F-Net)	-	-	74.8	78.0
#2 Single - nlnet	-	-	74.2	77.1
Published				
unet (Ensemble)	-	-	71.4	74.9
SLQA+ (Single)	-		71.4	74.4
Ours				
$\text{BERT}_{\text{LARGE}}$ (Single)	78.7	81.9	80.0	83.1

Table 3: SQuAD 2.0 results. We exclude entries that use BERT as one of their components.

Resultados

Hyperparams				Dev Set Accuracy		
#L	#H	#A	LM (ppl)	MNLI-m	MRPC	SST-2
3	768	12	5.84	77.9	79.8	88.4
6	768	3	5.24	80.6	82.2	90.7
6	768	12	4.68	81.9	84.8	91.3
12	768	12	3.99	84.4	86.7	92.9
12	1024	16	3.54	85.7	86.9	93.3
24	1024	16	3.23	86.6	87.8	93.7

Table 6: Ablation over BERT model size. #L = the number of layers; #H = hidden size; #A = number of attention heads. "LM (ppl)" is the masked LM perplexity of held-out training data.

Resultados

Hyperparams				Dev Set Accuracy		
#L	#H	#A	LM (ppl)	MNLI-m	MRPC	SST-2
3	768	12	5.84	77.9	79.8	88.4
6	768	3	5.24	80.6	82.2	90.7
6	768	12	4.68	81.9	84.8	91.3
12	768	12	3.99	84.4	86.7	92.9
12	1024	16	3.54	85.7	86.9	93.3
24	1024	16	3.23	86.6	87.8	93.7

Table 6: Ablation over BERT model size. #L = the number of layers; #H = hidden size; #A = number of attention heads. "LM (ppl)" is the masked LM perplexity of held-out training data.

Gracias! :)

Dudas?

0223826@up.edu.mx
+52 449 999 1194