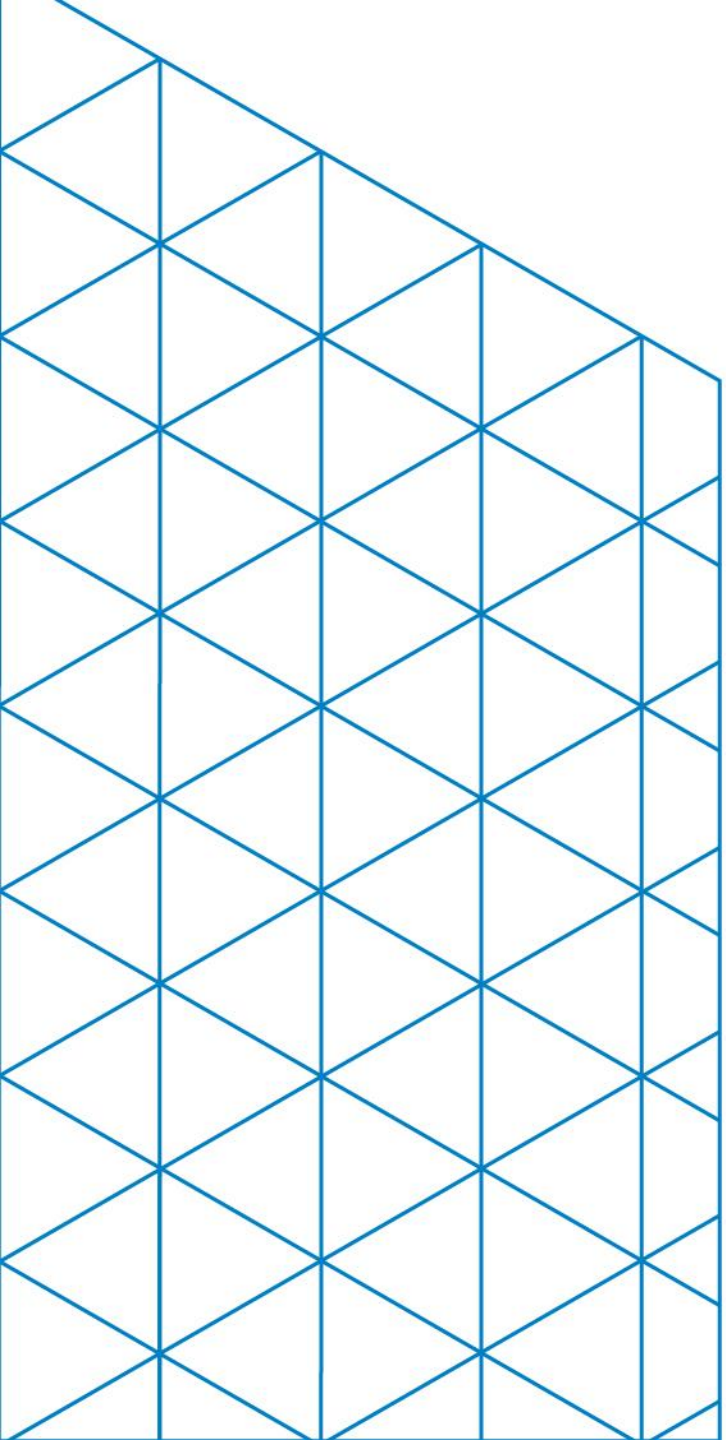


UFG
UNIVERSIDADE
FEDERAL DE GOIÁS

DEEP LEARNING
BRASIL NLP

ELECTRA: PRE-TRAINING TEXT ENCODERS AS DISCRIMINATORS RATHER THAN GENERATORS

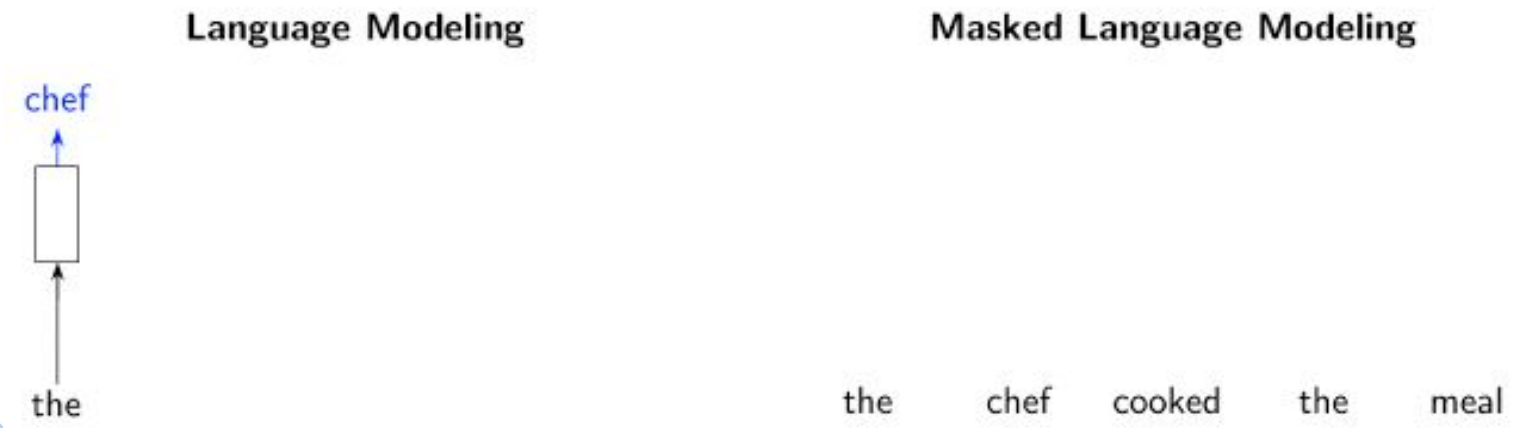
UFG
UNIVERSIDADE FEDERAL DE GOIÁS

Introdução

Os modelos atuais de estado da arte usam a abordagem de "corromper" a entrada substituindo alguns tokens por [MASK] (**MLM - Mask language modeling**) e treina a rede para reconstruir esses tokens.

Introdução

- **LMs** - Processam o texto de entrada da esquerda para a direita, prevendo a próxima palavra, considerando o contexto anterior.
- **MLMs** - Prevêem as identidades de um pequeno número de palavras que foram mascaradas da entrada.



O problema

Esses modelos que usam **MLM** para serem efetivos utilizam um grande volume computacional para serem eficientes.

Porque consomem mais recursos?

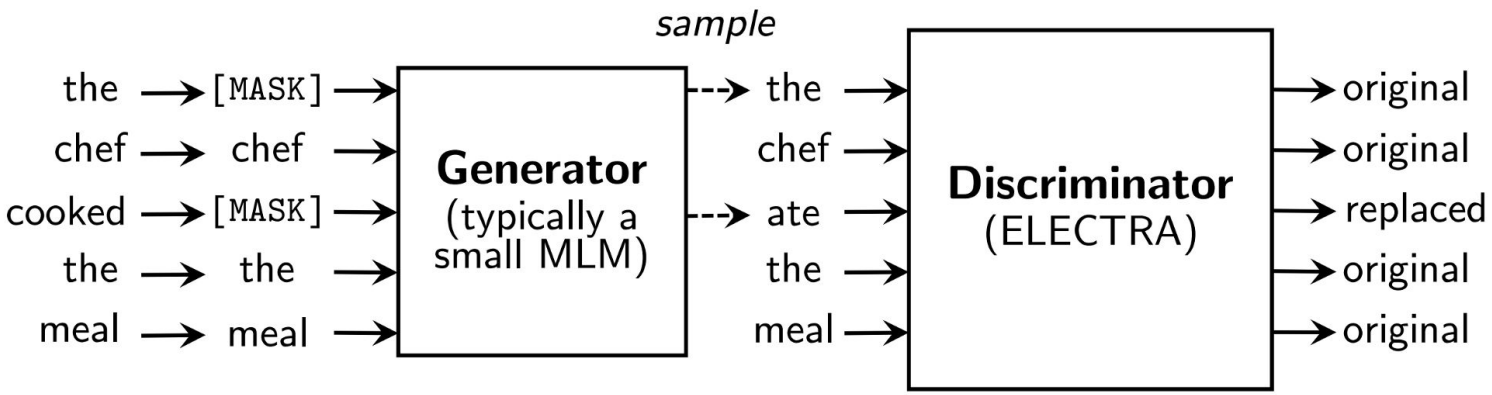
1. Abordagens MLM apenas aprendem com os tokens mascarados (normalmente 15%) de qualquer exemplo.
2. Os tokens de máscara aparecem apenas no estágio de pré-treinamento e nunca durante o ajuste fino (fine-tunning).

A proposta

- O **ELECTRA** usa a técnica **RTD** - Replaced Token Detection, que treina um modelo bidirecional (como um MLM) enquanto aprende com todas as posições de entrada (como um LM).
- Inspirado pelas redes adversárias generativas (**GANs**), o ELECTRA treina o modelo para distinguir entre dados de entrada "reais" e "falsos".

Arquitetura

ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately)



Passos do pré-treino

1. Para uma determinada sequência de entrada, substitua aleatoriamente alguns tokens por um token **[MASK]**.
2. O gerador prevê os tokens originais para todos os tokens mascarados.
3. A sequência de entrada para o discriminador é criada substituindo os tokens **[MASK]** pelas previsões do gerador.
4. Para cada token na sequência, o discriminador prevê se é um original ou se foi substituído pelo gerador.

Detalhes da Arquitetura

- O modelo do gerador é treinado para prever os tokens originais para tokens mascarados.
- O modelo discriminador é treinado para prever quais tokens foram substituídos.
- A perda do discriminador pode ser calculada sobre todos os tokens de entrada à medida que executa a previsão em cada token.

Detalhes da Arquitetura

- Com o MLM, a perda do modelo é calculada apenas sobre os tokens mascarados.
- O modelo discriminador é usado para as tarefas específica (fine-tunning) e o gerador é descartado após o pré-treinamento.

[MASKS] Pre-trainning x Fine-Tunning

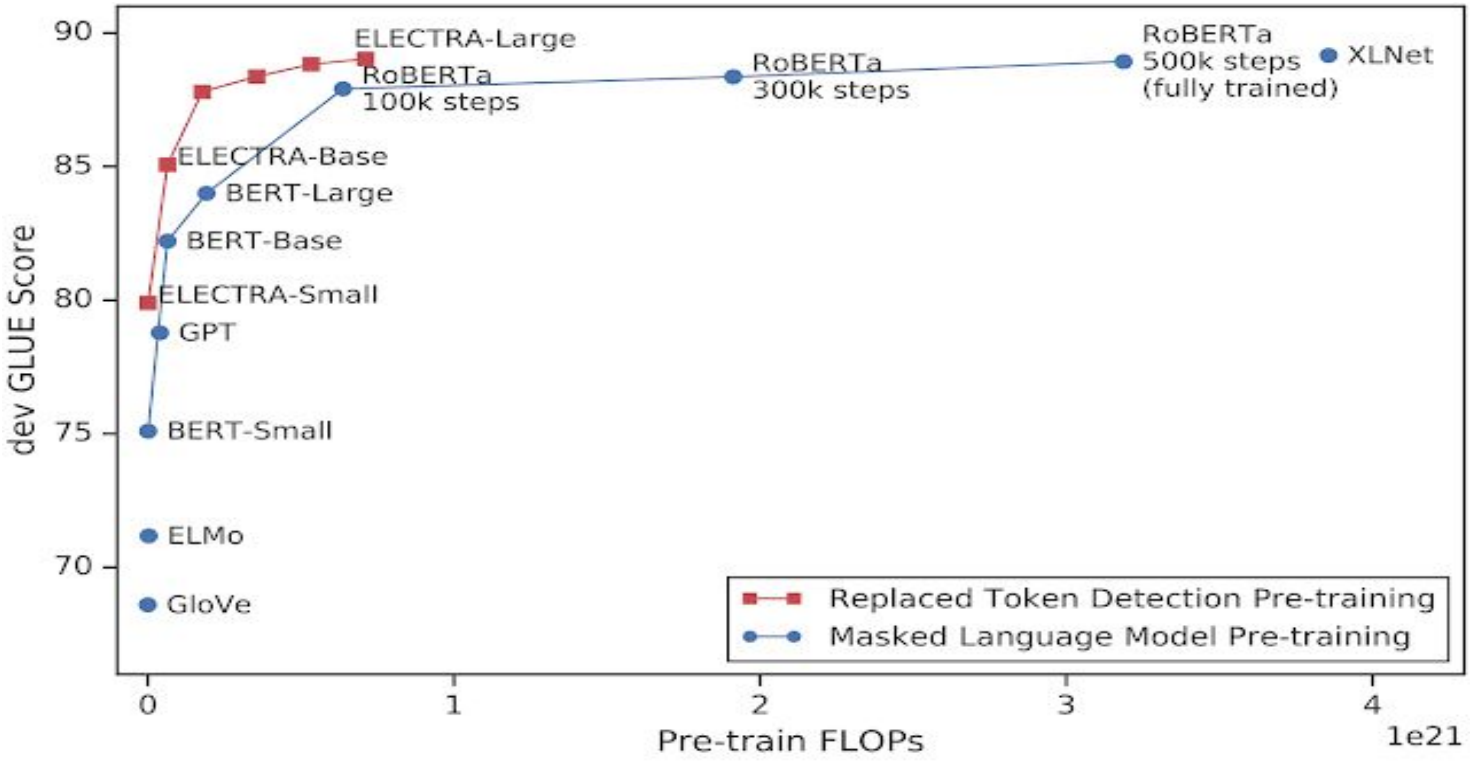
- O modelo **MLM** de substituição pontua **82,4**, superando ligeiramente o desempenho do **BERT 82,2**. Essa diferença indica que a discrepância entre o pré-treinamento e o ajuste fino prejudica levemente o desempenho do **BERT**.

Model	ELECTRA	All-Tokens MLM	Replace MLM	ELECTRA 15%	BERT
GLUE score	85.0	84.3	82.4	82.4	82.2

UFG
UNIVERSIDADE FEDERAL DE GOIÁS

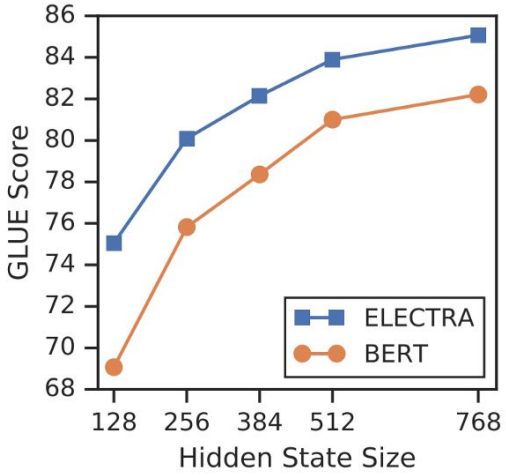
Comparação de Resultados

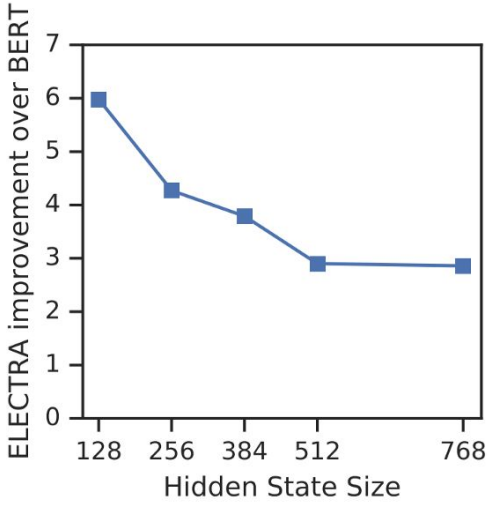
- O ELECTRA foi comparado com outros modelos e apresentou um desempenho comparável ao RoBERTa e XLNet, enquanto utiliza menos de 25% da computação.

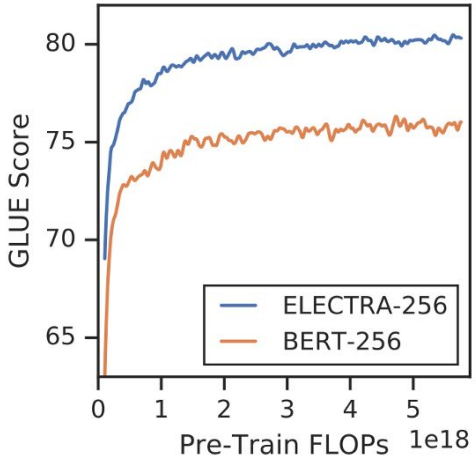


Electra X BERT (Smaller Models)

- Os ganhos de desempenho do ELECTRA sobre o BERT são maiores em tamanhos de modelo menores.







Conclusões

- O resultados do ELECTRA são excelentes.
- ELECTRA deve reduzir significativamente a barreira dos recursos de computação para treinar seus próprios modelos de linguagem.
- A possibilidade de treinar em uma única GPU, resolve o problema de poucos recursos.

Obrigado(a)!

UFG
UNIVERSIDADE
FEDERAL DE GOIÁS

Manoel Veríssimo
verissimo.manoel@gmail.com