B.E.R.T.

Pre-training of Deep Bidirectional Transformers for Language Understanding

COLIN KIRBY – CO201115@UCF.EDU

Why BERT was Needed in NLP.

What is BERT ?

- **BERT** = Bidirectional Encoder Representations from Transfomers

Why This Problem Matters :

- NLP tasks need full context – one-way models fall short.

What Existed before BERT :

- **ELMo** : Shallow use of both directions, no deep bidirectionality.
- **Lack of context** → Poor results on QA and inference tasks.

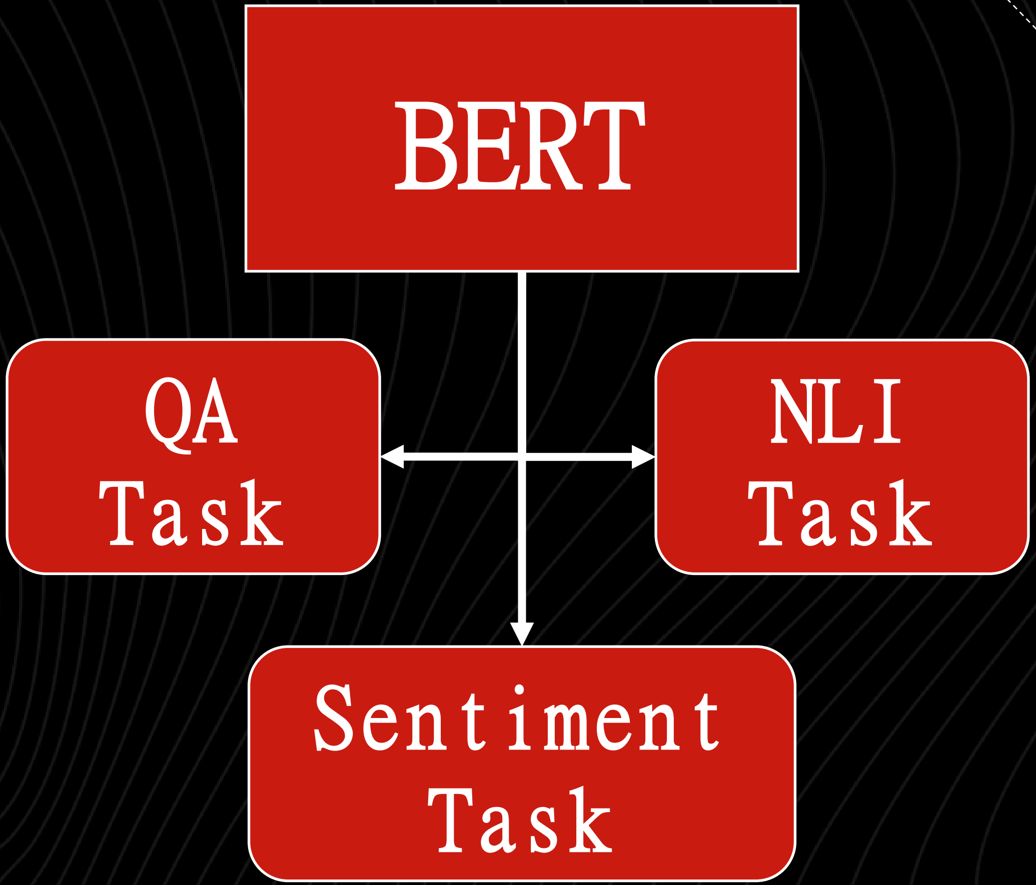
Problem With These Methods :

- Incomplete context hurts performance on key tasks like QA and NLI.

Problem Statement.

Main Goal :

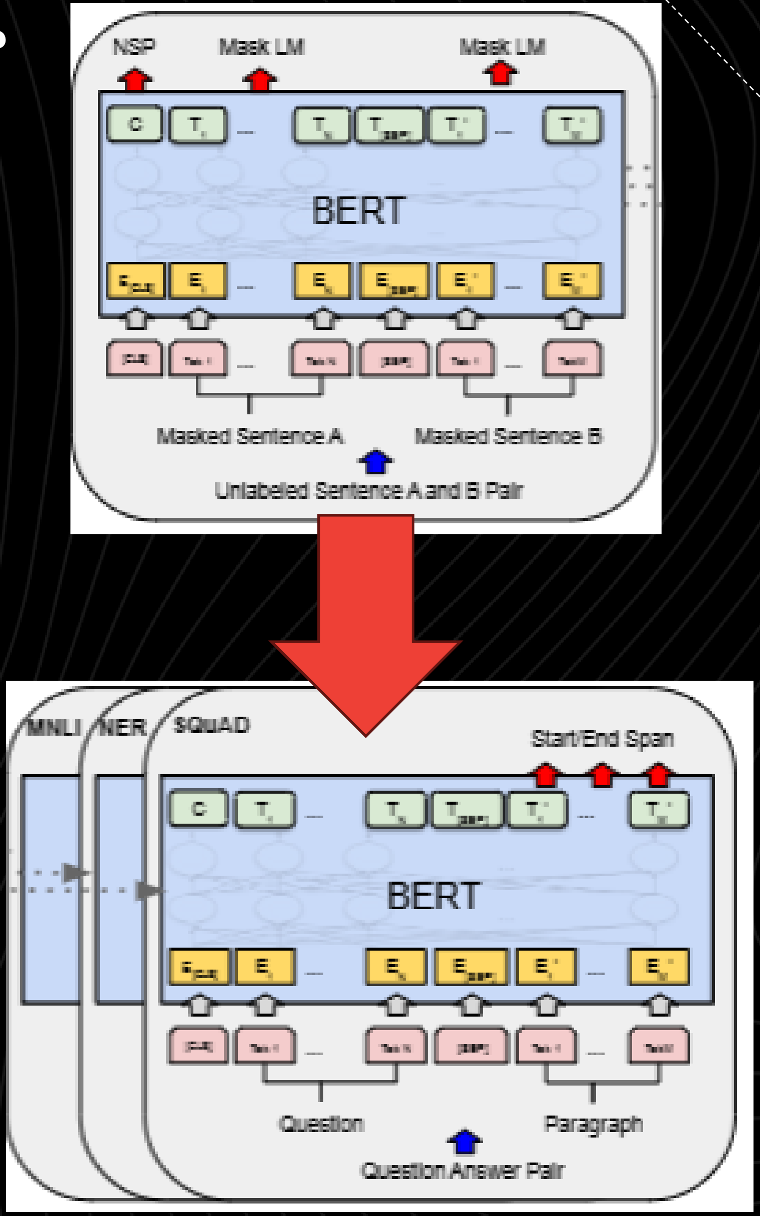
- Build a single model that works across many NLP tasks.
- Learn deep, bidirectional context representations.
- Enable fine-tuning with minimal task-specific changes.



Research Question :

Can a masked language model be pre-trained effectively to boost performance on a wide range of NLP tasks?

How BERT Learns Language.



Two-Stage Training Framework.

1. Pretraining :

- Learn from unlabeled text corpora.
- **Two Tasks :**
  - Masked Language Modeling (MLM) : Randomly Mask 15% of Tokens & Predict Them.
  - Next Sentence Prediction (NSP) : Predict if sentence B follows sentence A.

2. Fine-tuning :

- Adapt the pre-trained model to specific NLP tasks.
- Minimal architecture changes (Just adds tasks specific heads).

5 Core Contributions of BERT

Masked Language Modeling (MLM).

- Enables deep bidirectional learning from unlabeled text.

Next Sentence Prediction (NSP).

- Helps model sentence-level relationships.

One model, many tasks.

- Works on 11+ NLP tasks with minimal changes.

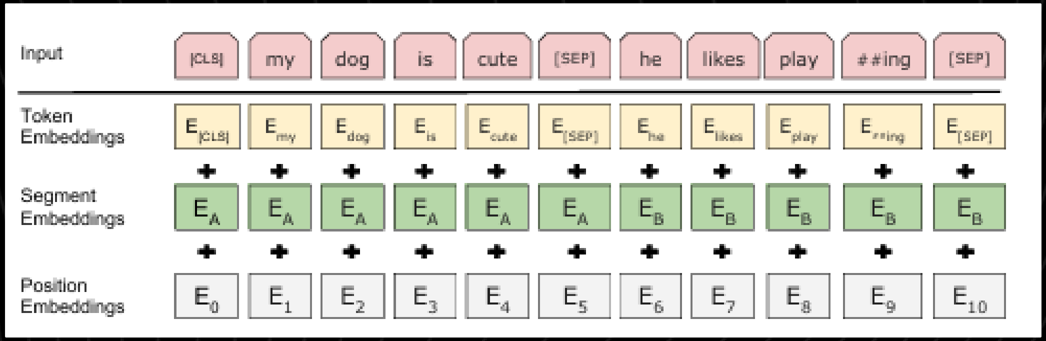
Open source + Pre-trained release.

- Made large-scale NLP accessible & reproducible.

6 Technical Details of BERT

Input = Token + Segment + Position

- Words + Sentence IDs + Word Order
- All summed → fed into Transformer Encoder.



Pre-training = MLM + NSP

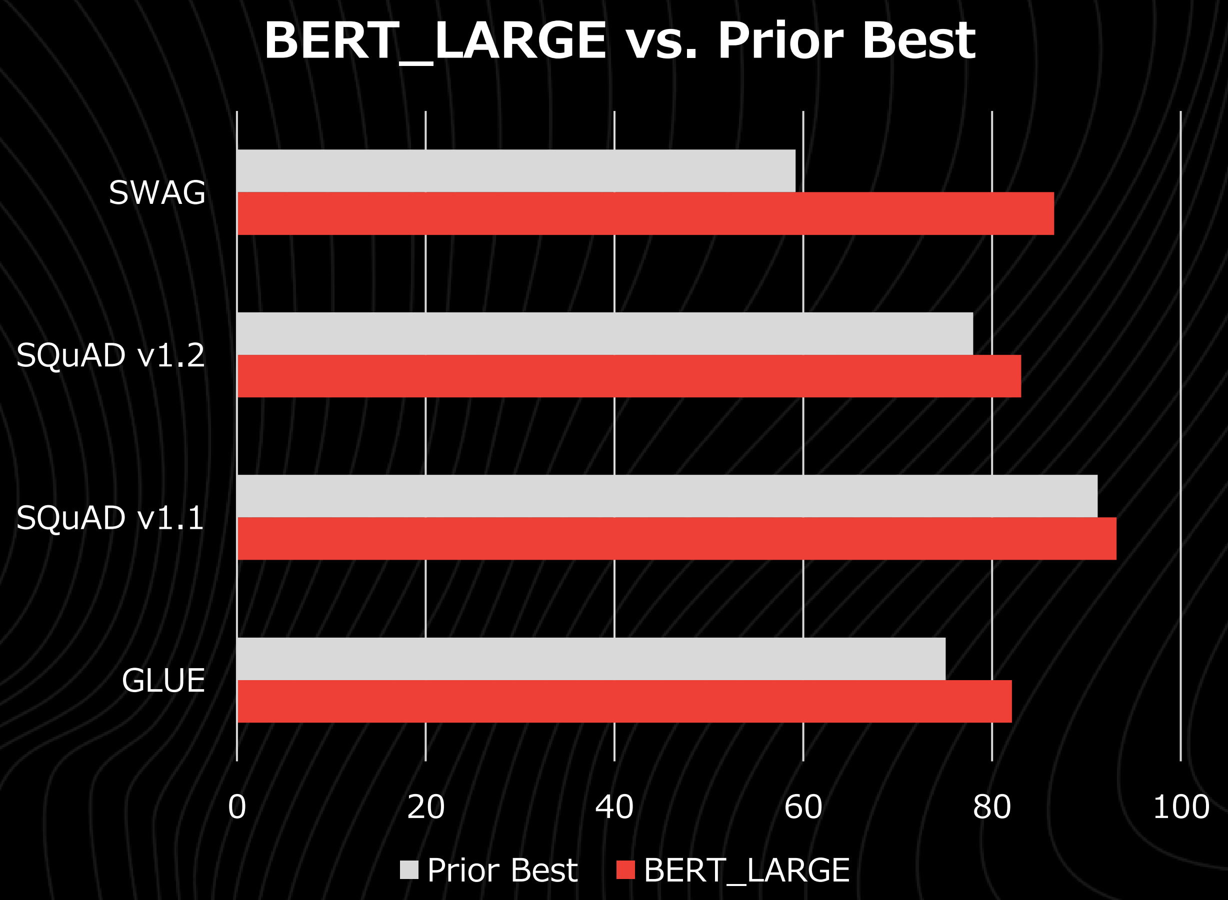
- **15% masked token :**
  - 80% [MASK]
  - 10% Random
  - 10% Unchanged

Model	Layers	Hidden Size	Heads	Params
BERT_BASE	12	768	12	110M
BERT_LARGE	24	1024	16	340M

- **Next Sentence Prediction :** 50% Actual Next Sentence, 50% Random Pairing

Technical Details of BERT

1. **GLUE Benchmark (** 8 NLP Tasks **)**
   - **BERT_LARGE Avg. :** 82.1
   - **GPT Avg :** 75.1 (+7 pt gain across diverse tasks)
2. **SQuAD Question Answering**
   - **v1.1 F1 Score :**
     - **Single :** 90.9
     - **Ensemble :** 93.2 (better than human-level)
3. **SWAG (** Commonsense Inference **)**
   - BERT_LARGE outperformed ELMo by +27.1%



ELMo vs. GPT vs. BERT

ELMo : Bidirectional LSTM

← the ← cat ← sat ← (Reads R→L) && → the → cat → sat → (Reads L→R)

- Words see context only from one side at a time. No integration of both dirs at once.

GPT : Unidirectional Transformer Decoder

→ the → cat → sat → on → the → mat →

- Each word can only “see” the words before it, not after.

BERT : Deep Bidirectional Transformer Encoder

↔ the ↔ [MASK] ↔ sat ↔ on ↔ the ↔ mat

- It uses both left and right context at the same time.

ELMo vs. GPT vs. BERT

Model	Context Type	Pretraining Mask	Used For	Limitations
ELMo	LTR + RTL Concat	LM (No Mask)	Feature Extraction	Shallow fusion, no fine-tuning
GPT	LTR Only	LM (Next Word)	Task-specific Fine-tuning	Can’t use right context
BERT	Deep Bidirectional	MLM + NSP	Task-specific Fine-tuning	Higher training cost, slower inference

Why BERT Wins :

BERT was the first to use deep two-way context and support easy fine-tuning — helping it outperform past models on tasks like question answering and sentiment analysis.

10 Conclusion & Key Takeaways

What BERT did.

- Introduced deep bidirectional pretraining using Masked LM and Next Sentence Prediction.

Impact.

- Achieved SOTA on 11 NLP Tasks.
- Outperformed humans on some benchmarks (e.g. SQuAD)

Legacy.

- Inspired successors like **RoBERTa**, **ALBERT**, **DistilBERT**, and **T5.**

Key Insight.

- Pretraining on full context + minimal fine-tuning = strong general NLP performance