Natural Language Processing

Week 6: Pre-trained Language Models - Simplified Edition

Joerg R. Osterrieder

The Reading Revolution: A Story

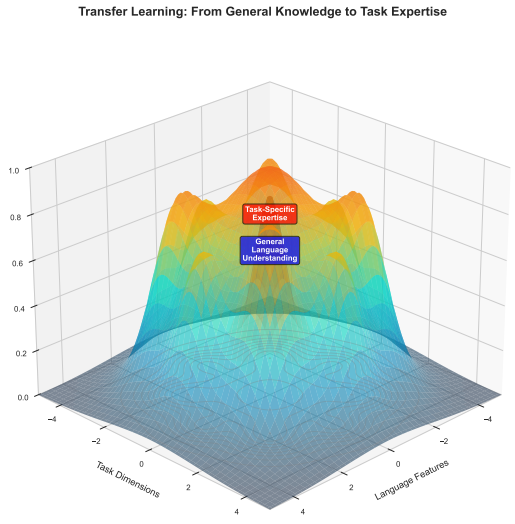
Imagine two students:

Student A: Never learned to read

- Starts medical school
- Must learn alphabet first
- Then words, grammar, sentences
- Finally can read medical texts
- Takes 10 years to become doctor

Student B: Already knows how to read

- Starts medical school
- Focuses only on medical knowledge
- Becomes doctor in 4 years



Your Learning Journey Today

By the end of this session, you will:

1. **Understand** why we pre-train models
   - The waste problem
   - Transfer learning magic
2. **Master** two approaches
   - BERT: Fill in the blanks
   - GPT: Predict what's next
3. **Apply** fine-tuning
   - Adapt to your task
   - 10x faster than training from scratch
4. **Choose** the right model
   - Decision tree for model selection
   - Trade-offs and considerations

Checkpoint

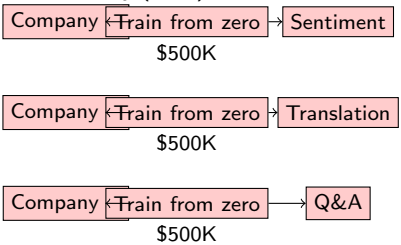
Think of pre-training like learning to drive: Once you know how, you can drive any car, not just the one you learned in!

Real World

Every time you use autocomplete on your phone, you're using a pre-trained model!

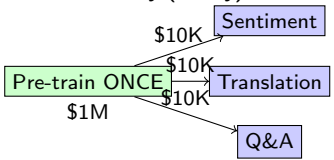
The Million Dollar Waste

The Old Way (2017):



Total waste: $1.5 million!

The Smart Way (Today):



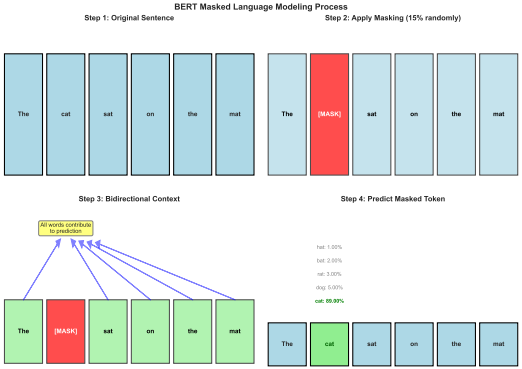
Total cost: $1.03 million Savings: $470,000!

Common Misconception

"Pre-training is just memorizing text" - No! It's learning language patterns, grammar, and world knowledge that transfers to any task.

BERT: The Fill-in-the-Blank Master

How BERT Learns:



Like a comprehension test:

1. Take a sentence
2. Hide some words (15%)
3. Guess the hidden words
4. Use ALL surrounding context

Try it yourself:
"The [MASK] sat on the mat"

- Look left: "The"
- Look right: "sat on the mat"
- Guess: cat? dog? child?

BERT's superpower: Bidirectional!

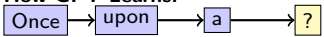
- Sees past AND future
- Like reading the whole page at once
- Not just left-to-right

Intuition

BERT is like a detective who can look at all the clues (words) around a missing piece to figure out what it should be.

GPT: The Story Continuation Expert

How GPT Learns:

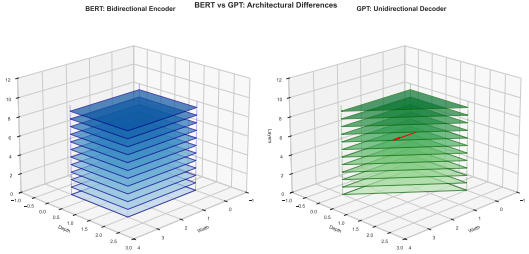


time (85%)
day (10%)
night (5%)

Like autocomplete on steroids:

1. Read text left to right
2. Predict next word
3. Use only what came before
4. Keep going to generate text

GPT vs BERT:



Real World

ChatGPT = GPT trained to have conversations. It's predicting the next word, one at a time, thousands of times per second!

BERT vs GPT: Head-to-Head

Feature	BERT	GPT
Direction	Bidirectional $\leftrightarrow$	Left-to-right $\rightarrow$
Training	Fill blanks	Predict next
Strength	Understanding	Generation
Best for	Classification	Text creation
Example	Sentiment analysis	ChatGPT
Speed	Faster for analysis	Slower (sequential)
Context	Sees everything	Only sees past

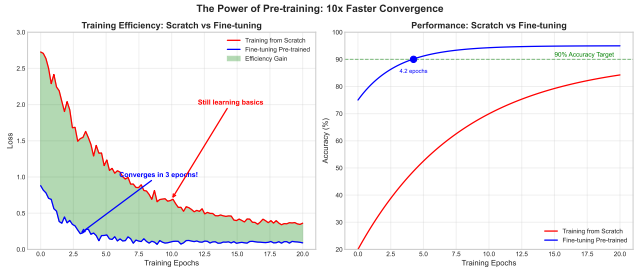
Intuition

BERT = Reading comprehension expert
GPT = Creative writing expert

Checkpoint

Remember: BERT for understanding, GPT for generating!

Fine-tuning: Teaching New Tricks



The fine-tuning advantage:

- Start with language knowledge
- Add task-specific skills
- 10x faster convergence
- Need less data

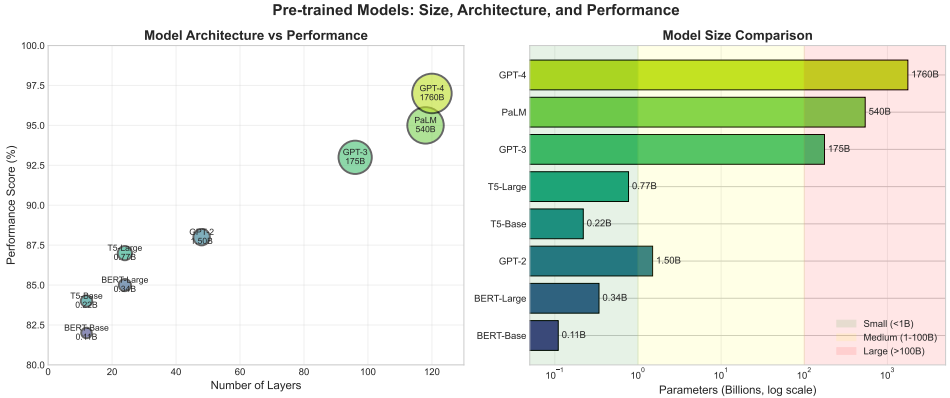
Like learning a specialty:

1. Pre-training = Medical school
2. Fine-tuning = Specialization
   - Cardiology
   - Neurology
   - Pediatrics

Common Misconception

"Fine-tuning erases pre-training" - No! It adds a thin layer of task-specific knowledge on top.

The Model Zoo: Choose Your Fighter!

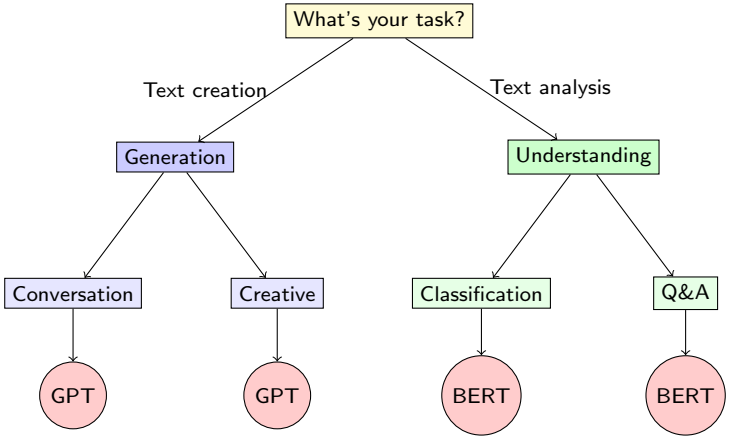


Checkpoint

Bigger isn't always better! Choose based on:

- Your task complexity
- Available computing power

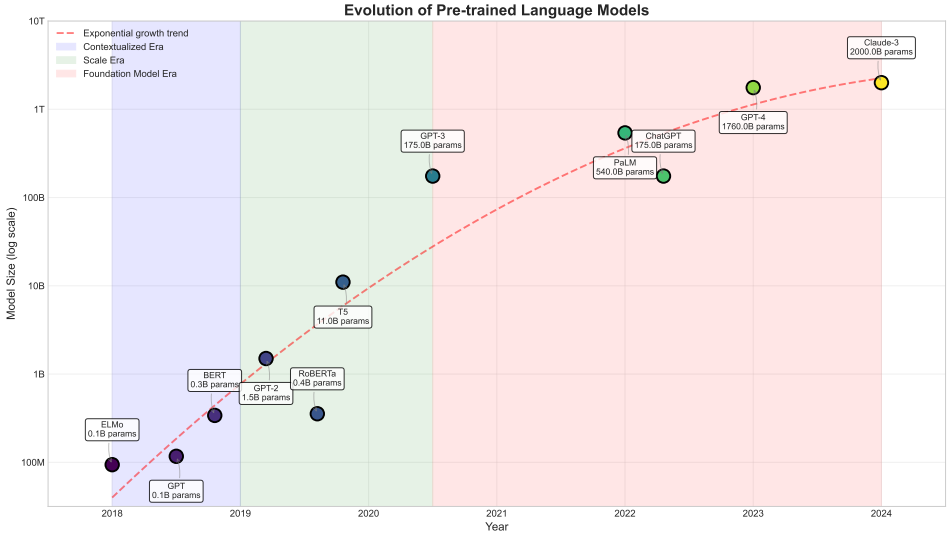
Which Model Should You Use?



Real World

Real applications: Gmail autocomplete (GPT-style), Google search understanding (BERT-style)

The Evolution: From ELMo to GPT-4



Let's Code: Using Pre-trained BERT

```python
# Super simple BERT usage
from transformers import pipeline

# Load pre-trained BERT for sentiment
classifier = pipeline("sentiment-analysis")

# That's it! Now use it:
result = classifier("I love this NLP course!")
print(result)
# Output: [{'label': 'POSITIVE', 'score': 0.99}]

# Fill-in-the-blank with BERT
fill_mask = pipeline("fill-mask")
result = fill_mask("The course is [MASK] interesting.")
print(result[0])
# Output: {'token_str': 'very', 'score': 0.85}
```

Checkpoint

3 lines of code vs 3 months of training from scratch!

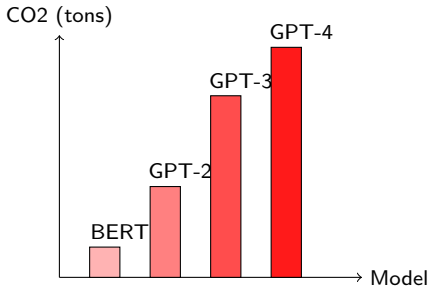
The Carbon Footprint Reality

Training costs:

- GPT-3: 1,287 MWh of electricity
- = 552 tons of CO2
- = 120 cars for a year

Why this matters:

- Use pre-trained when possible
- Don't train from scratch
- Share your models
- Consider efficiency



Real World

One GPT-3 training run = Flying from NY to SF 400 times!

Your Turn: Hands-On Exercise

Exercise: Fine-tune BERT for Movie Reviews

1. Load pre-trained BERT-base
2. Prepare IMDB movie review dataset
3. Fine-tune for 3 epochs
4. Compare with training from scratch

Expected results:

- Fine-tuned BERT: 92% accuracy in 10 minutes
- From scratch: 75% accuracy in 2 hours

Checkpoint

Use Google Colab for free GPU access!

Intuition

Fine-tuning is like teaching a professor a new course vs teaching a toddler everything from scratch.

Key Takeaways: What to Remember

The Big Ideas:

1. **Pre-training** = Learning language once
2. **Fine-tuning** = Adapting to your task
3. **BERT** = Bidirectional understanding
4. **GPT** = Sequential generation
5. **Transfer learning** = Reuse knowledge

Practical Tips:

- Never train from scratch
- Start with smallest model that works
- Use Hugging Face for easy access
- Fine-tune on your specific domain
- Monitor environmental impact

Real World

Every major AI application today uses pre-trained models: ChatGPT, Google Search, Grammarly, GitHub Copilot, and more!

The Future: What's Next?

Current Frontiers (2024):

- **Multimodal:** Text + Images + Audio (GPT-4V)
- **Efficient:** Smaller, faster models (Llama, Phi)
- **Specialized:** Domain-specific models (BloombergGPT)
- **Aligned:** Following human values (Claude)

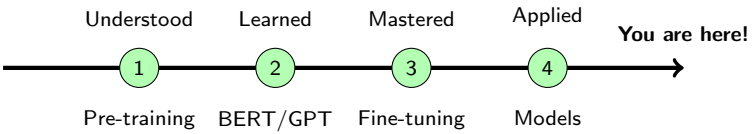
Next Week Preview:

- Advanced transformer architectures
- Efficient transformers
- Scaling laws
- Emergent abilities

Checkpoint

The field moves fast! What you learn today will evolve, but the core concepts remain.

Summary: From Zero to Hero



Intuition

You now understand the technology behind ChatGPT, Google Search, and modern AI. That's huge!