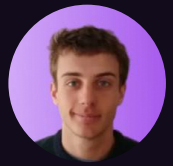
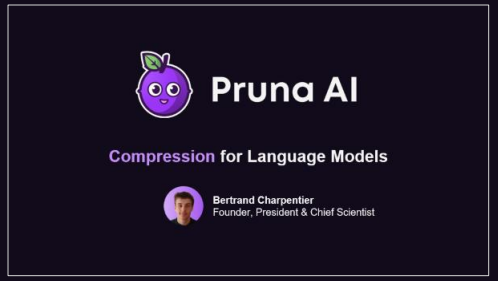
Pruna AI

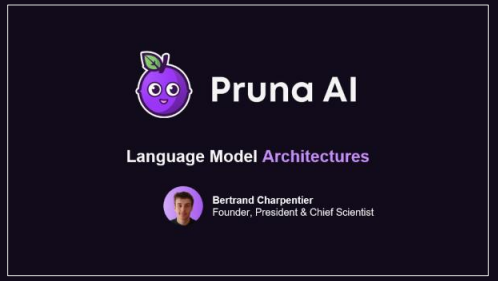
How to Make Inference of Deep Learning Models Efficient?

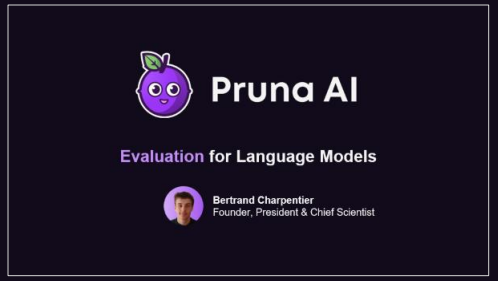


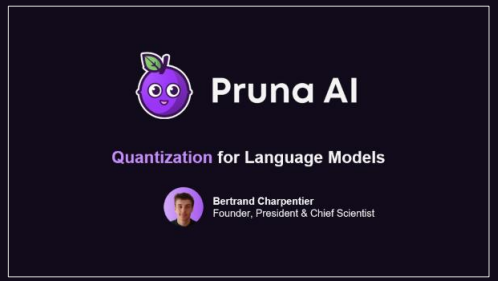
Bertrand Charpentier
Founder, President & Chief Scientist

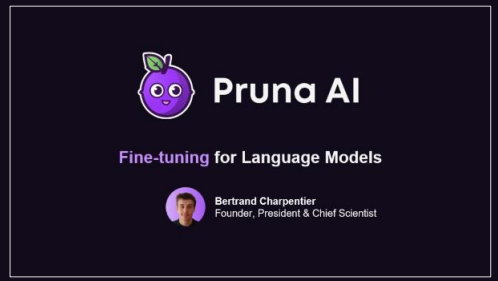
Overview


Pruna AI
Compression for Language Models
Bertrand Charpentier
Founder, President & Chief Scientist


Pruna AI
Language Model Architectures
Bertrand Charpentier
Founder, President & Chief Scientist


Pruna AI
Evaluation for Language Models
Bertrand Charpentier
Founder, President & Chief Scientist


Pruna AI
Quantization for Language Models
Bertrand Charpentier
Founder, President & Chief Scientist


Pruna AI
Fine-tuning for Language Models
Bertrand Charpentier
Founder, President & Chief Scientist

How Does This Lecture Work?

Overview of foundational concepts for AI efficiency

- Explanation of their meaning/intuitions
- Glossary/taxonomy of the terms

Overview of key algorithms for AI efficiency

How does the algorithm work?

- Practical/theoretical motivations & intuitions
- Step by step explanation of the algorithm (w/ diagram)
- Explanation of key formulas

When to use the algorithm in practice?

- Pros/cons
- Configuration requirements
- Optimized objective/final goal
- Examples of code snippets

References of This Lecture

All references are on the awesome AI efficiency Github:
https://github.com/PrunaAI/awesome-ai-efficiency

- Articles for general audience
- Reports for decision makers audience
- Technical papers for tech audience
- Also, books, lectures, people, and organizations

Give a star to the repo!