Unsupervised Learning

Anote Presented by Ben Setel

Overview

General Unsupervised Learning

RAG-based Unsupervised Learning for LLMs

Product Demo:

1. Complete answer in prompt
2. Incomplete answer in prompt
3. Relevant info but no answer in prompt

What is Unsupervised Learning

No labels/answers provided -> can’t “check our work”

BUT, can find patterns and correlations in the data

- Classification
- Outlier detection
- Dimensionality reduction

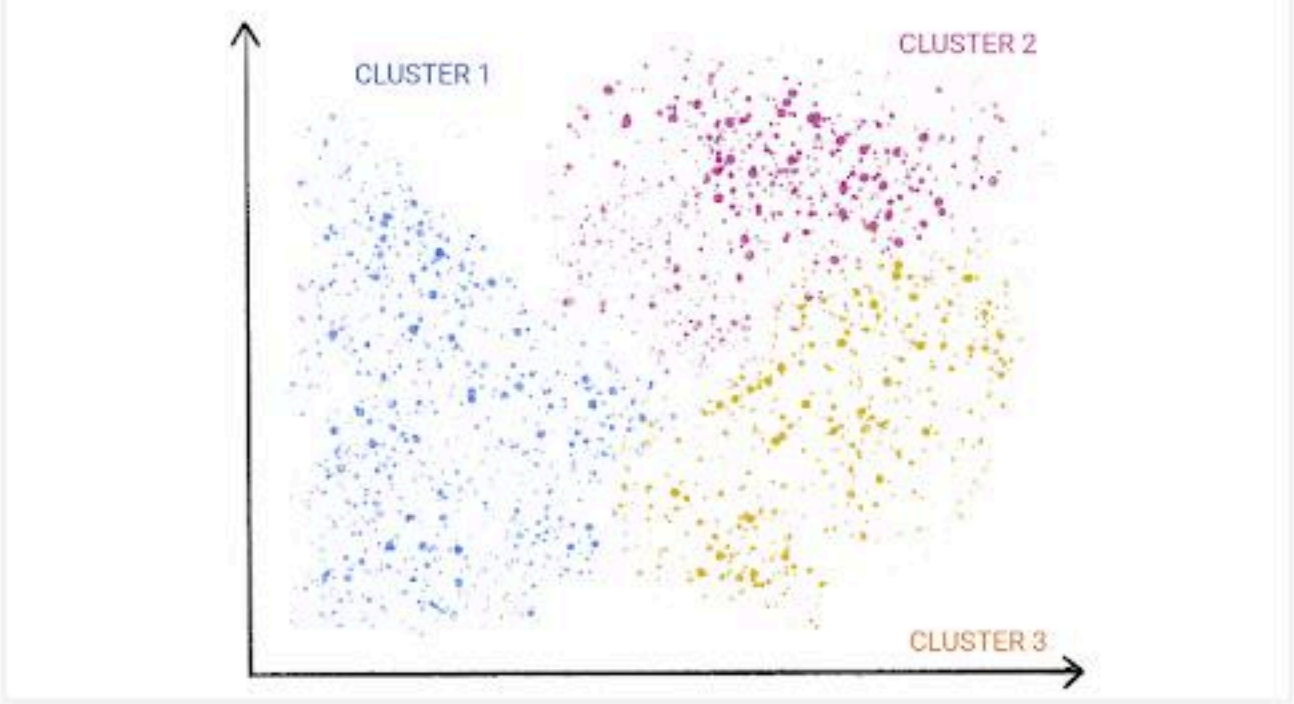


Figure 1. An ML model clustering similar data points.

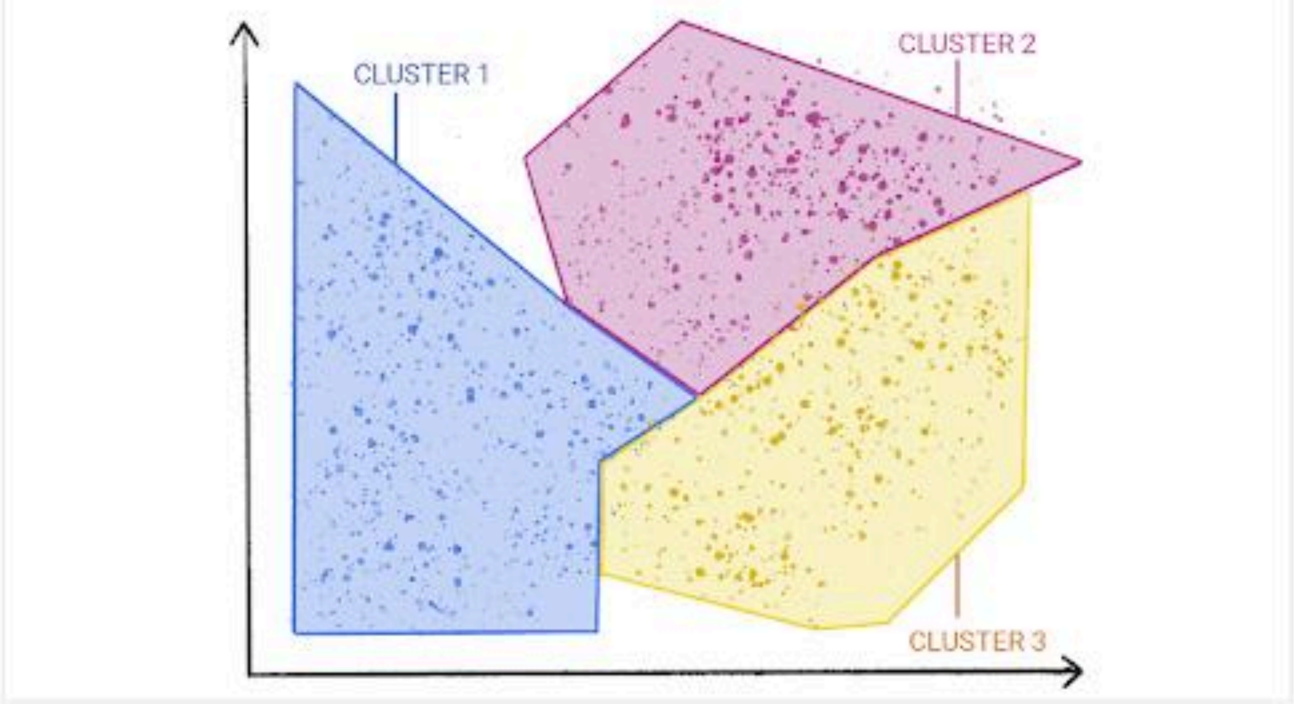


Figure 2. Groups of clusters with natural demarcations.

For language?

Traditional methods not very helpful here!

So, we make our own labels

Masked Language Modeling

“The quick brown fox jumps over the lazy dog”

- “The quick [MASK] fox jumps over the lazy dog”
- “The quick brown [MASK] jumps over the lazy dog”

For LLMs

Masked Language Modeling is a helpful start, but how do we best use it?

RAG with LLM as “Information Refiner”

Three approaches:

- Prompt contains all knowledge to answer Q; LLM has to pick it out
- Prompt has incomplete or incorrect knowledge
- Prompt has relevant knowledge, but no direct part of answer

E.g.: “What is Anote.ai and where are its headquarters?”

- “Anote.ai is a machine learning company helping users finetune LLMs with their data located in New York City”
- “Anote.ai is a technology company based in Boston”
- “Anote.ai was founded by Natan Vidra and Thomas Clifford”

Shicheng Xu, Liang Pang, Mo Yu, Fandong Meng, Huawei Shen, Xueqi Cheng, and Jie Zhou. 2024. Unsupervised Information Refinement Training of Large Language Models for Retrieval-Augmented Generation. arXiv preprint arXiv:2402.18150 (2024).

Making the data

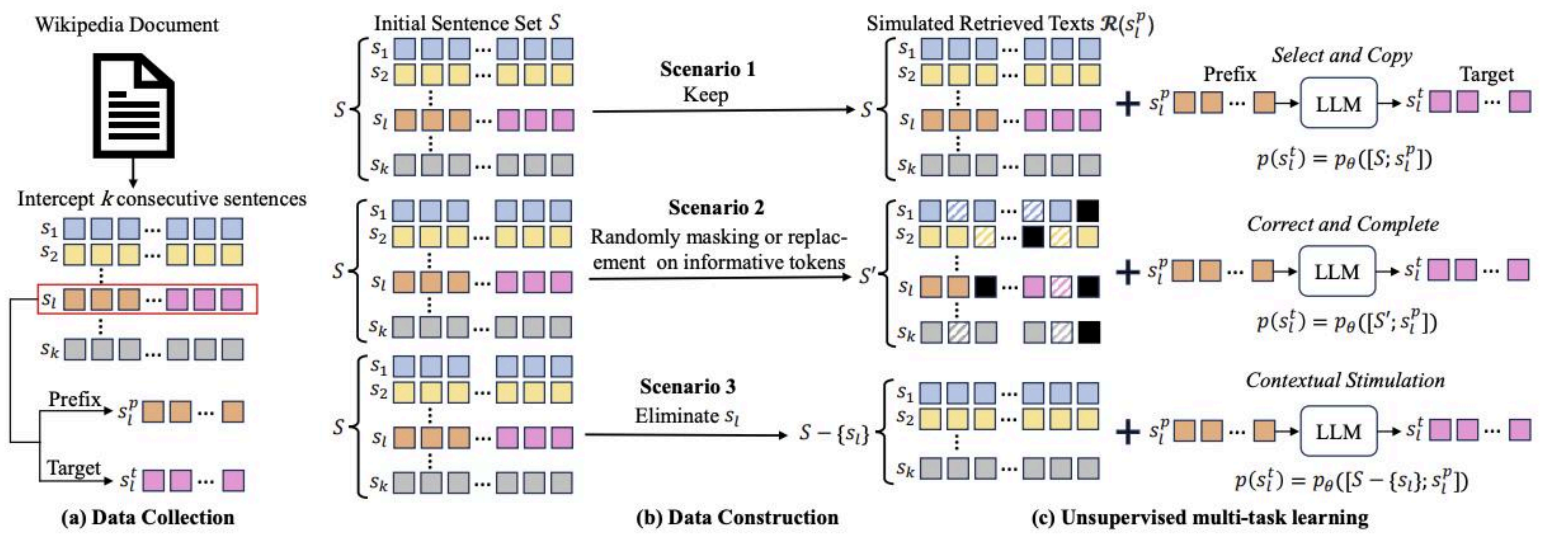

Wikipedia Document
Intercept k consecutive sentences
s_1
s_2
s_l
s_k
Prefix
s_l^p
Target
s_l^t
(a) Data Collection
Initial Sentence Set S
S
Scenario 1
Keep
Scenario 2
Randomly masking or replacement on informative tokens
S'
Scenario 3
Eliminate s_l
$S - \{s_l\}$
(b) Data Construction
Simulated Retrieved Texts $\mathcal{R}(s_l^p)$
Select and Copy
Prefix
Target
LLM
$p(s_l^t) = p_\theta([S; s_l^p])$
Correct and Complete
$p(s_l^t) = p_\theta([S'; s_l^p])$
Contextual Stimulation
$p(s_l^t) = p_\theta([S - \{s_l\}; s_l^p])$
(c) Unsupervised multi-task learning

Figure 2: Overview of our INFO-RAG. Each sample is only processed for a single scenario to avoid data leakage.

Shicheng Xu, Liang Pang, Mo Yu, Fandong Meng, Huawei Shen, Xueqi Cheng, and Jie Zhou. 2024. Unsupervised Information Refinement Training of Large Language Models for Retrieval-Augmented Generation. arXiv preprint arXiv:2402.18150 (2024).

Low Rank Adaptation (LoRA) for PEFT

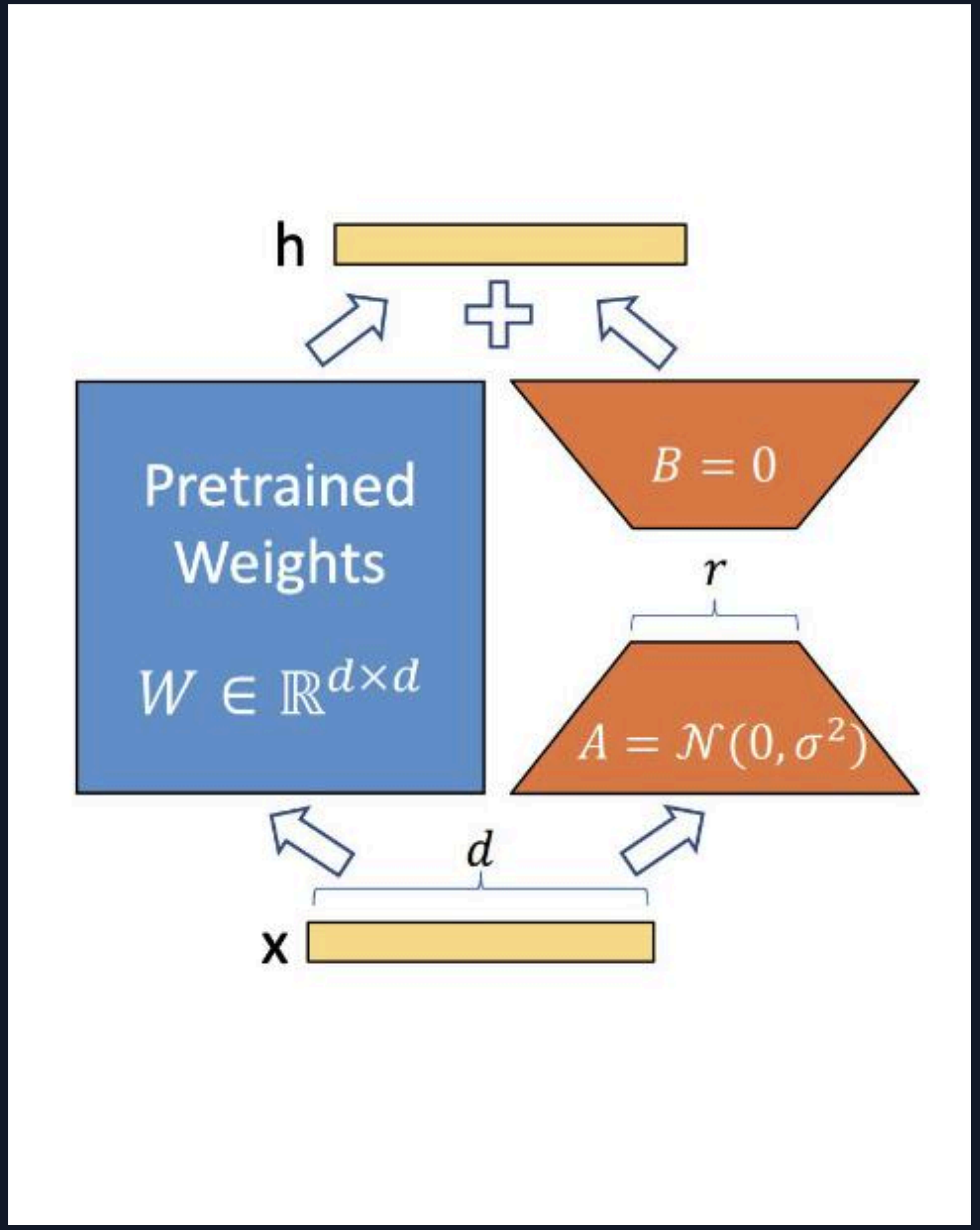
Reduces number of trainable parameters

Identifies crucial parameters for the task at hand and finetunes those

During fine-tuning, only the parameters in low-rank matrices are updated

Less chance of overfitting since only a few parameters are updated

Reduces computational and memory requirements needed to fine-tune


h
Pretrained Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x

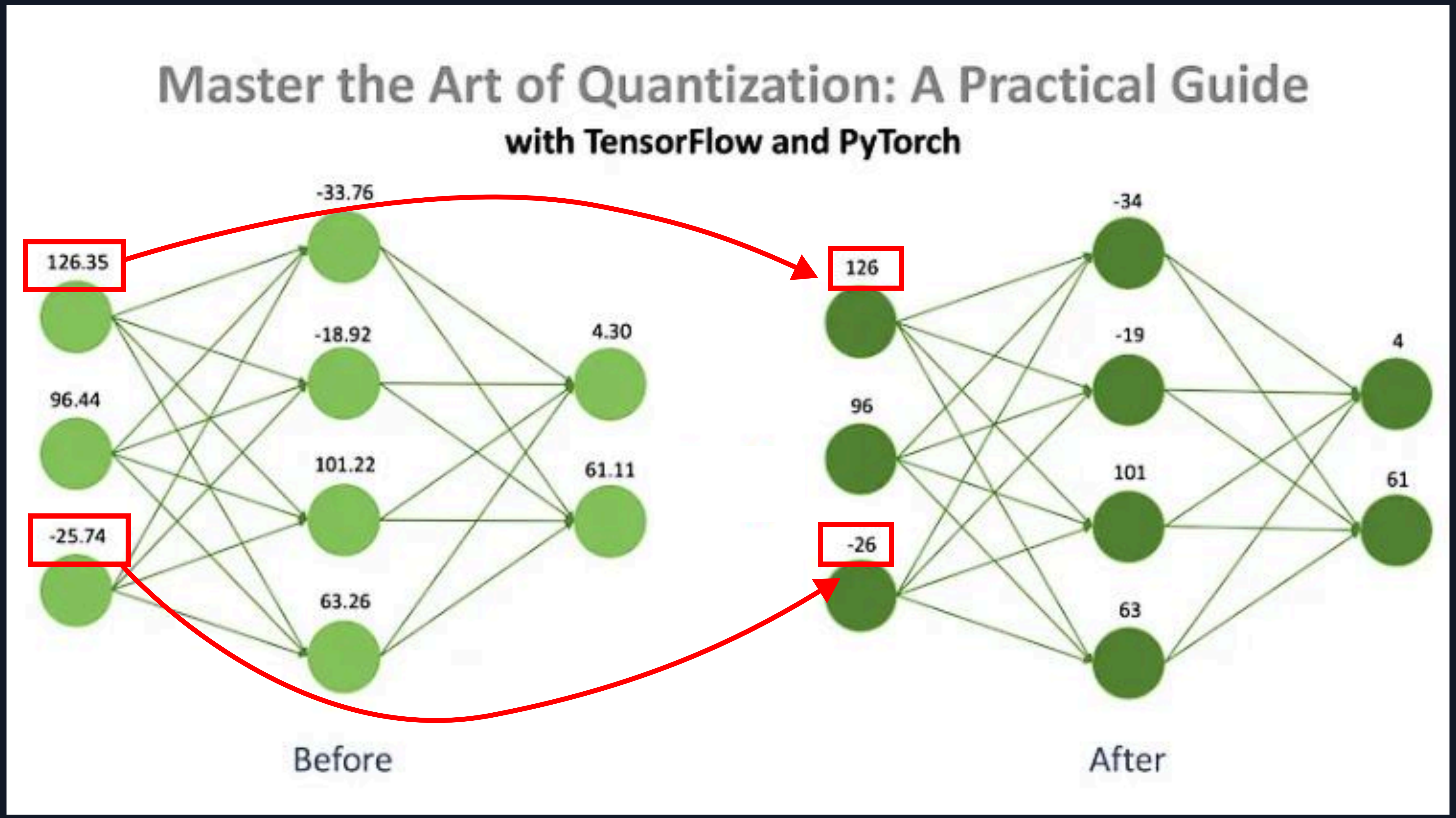
QLoRA for Quantized LLMS

Quantization: process of reducing the numerical precision of a model's tensors to make it faster and more compact

QLoRA combines quantization & low-rank adaptation

Model parameters are first quantized (usually to 4 bit precision) and then go through LoRA

Made fine tuning a lot more accessible



Product Demo