



A Sparsity Predicting Approach for Large Language Models via Activation Pattern Clustering

Nobel Dhar, Bobin Deng, Md Romyull
Islam, Xinyue Zhang, Kazi Fahim Ahmad
Nasif, Kun Suo

Presented by
Nobel Dhar

Introduction

Large Language Models

- Billions of parameters
- Huge computation and memory usage
- Operate on large data center

Introduction

Large Language Models

- Billions of parameters
- Huge computation and memory usage
- Operate on large data center

Edge AI

- Known for better latency
- Better privacy
- Non-Reliance on Network

Introduction

Large Language Models

- Billions of parameters
- Huge computation and memory usage
- Operate on large data center

Edge AI

- Known for better latency
- Better privacy
- Non-Reliance on Network

Activation Sparsity

- A subset of neurons is active
- Less neurons means less computation and memory usage

Introduction

Large Language Models

- Billions of parameters
- Huge computation and memory usage
- Operate on large data center

Edge AI

- Known for better latency
- Better privacy
- Non-Reliance on Network

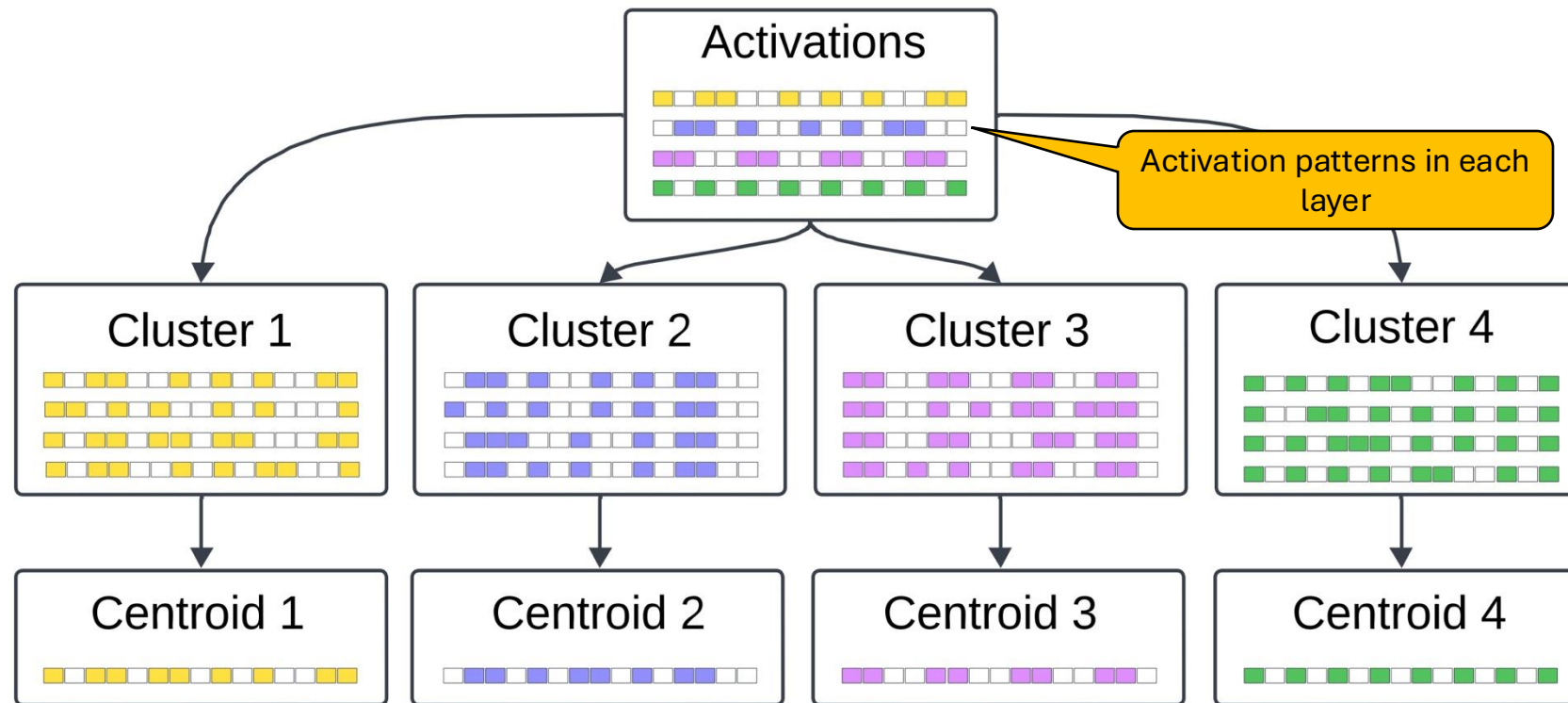
Activation Sparsity

- A subset of neurons is active
- Less neurons means less computation and memory usage

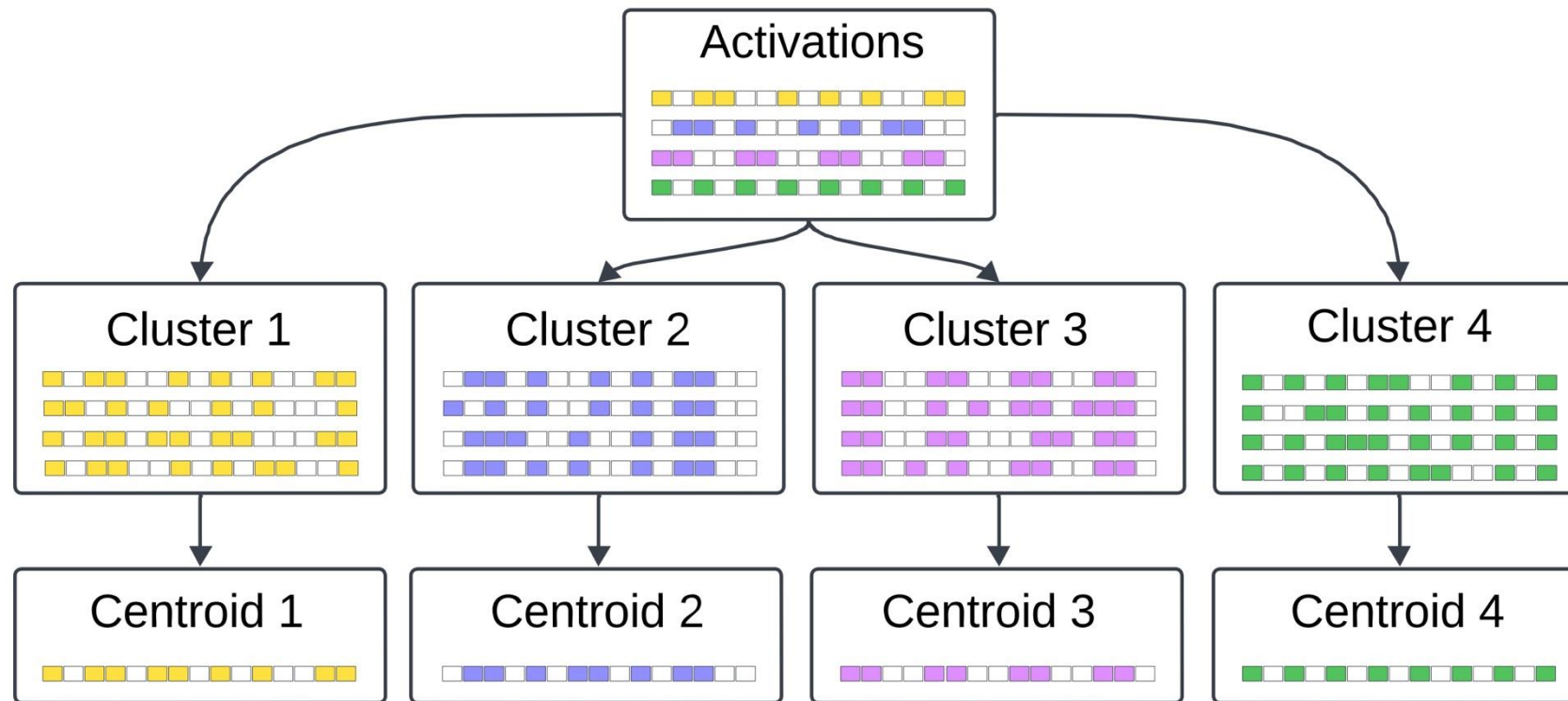
Challenges

- Different set of neurons is active for different tokens
- Need of sparsity prediction
- Neuron-level prediction is costly

Motivation



Motivation



Motivation

$$C_{direct} = N \times L \times T \times C$$

$$N_{FFN} \approx (2/3) \times 7 \times 10^9 = 4.67 \times 10^9$$

$$K = 2048 \times 3 = 6144$$

$$C_{clustered} = K \times L \times T \times C$$

$$\frac{C_{direct}}{C_{clustered}} = \frac{N_{FFN}}{K} = \frac{4.67 \times 10^9}{6144} \approx 7.6 \times 10^5$$

Efficiency Gain: $\sim 760,000 \times$

- N: Total neurons
- L: Layers
- T: Tokens per sequence
- C: Per-neuron compute cost
- K: Number of clusters

Related Works

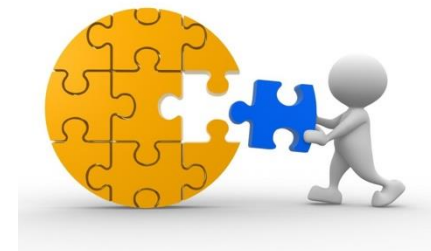
- Quantization
 - Reduce the model size, hence memory consumption
 - Doesn't provide noticeable benefit in terms of computation
 - Compromise model accuracy
- Pruning
 - Reduces both model size and computation
 - Compromise accuracy significantly
- Distillation
 - Reduces both model size and computation
 - Model loses flexibility across diverse input

Related Works

- Quantization
 - Reduce the model size, hence memory consumption
 - Doesn't provide noticeable benefit in terms of computation
 - Compromise model accuracy
- Pruning
 - Reduces both model size and computation
 - Compromise accuracy significantly
- Distillation
 - Reduces both model size and computation
 - Model loses flexibility across diverse input

Exploiting activation sparsity can provide both computation and memory benefits without losing model accuracy, if done efficiently.

Our work aims is to fill this gap.



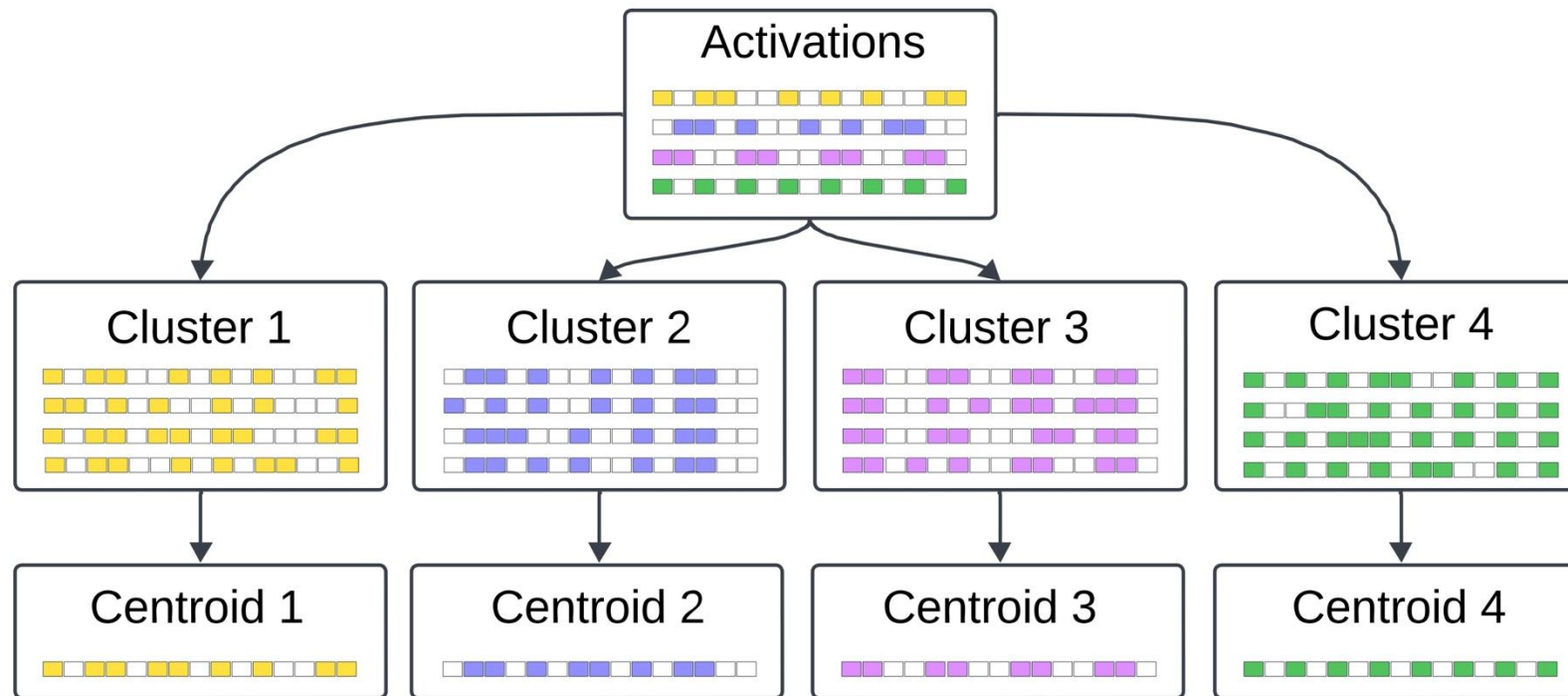
Experimental Setup

- Dataset
 - Wikitext-2
 - 333,842 number of tokens
 - 10.64 million activation patterns
- Model
 - Mistral-7B-v0.1
 - Number of FFN layers: 32
 - Each FFN layer consists of projection layers:
 - *Gate-Proj*: 14336 neuron
 - *Up-Proj*: 14336 neuron
 - *Down-Proj*: 4096 neuron
 - 50% model accuracy (PPL Score): 6.45

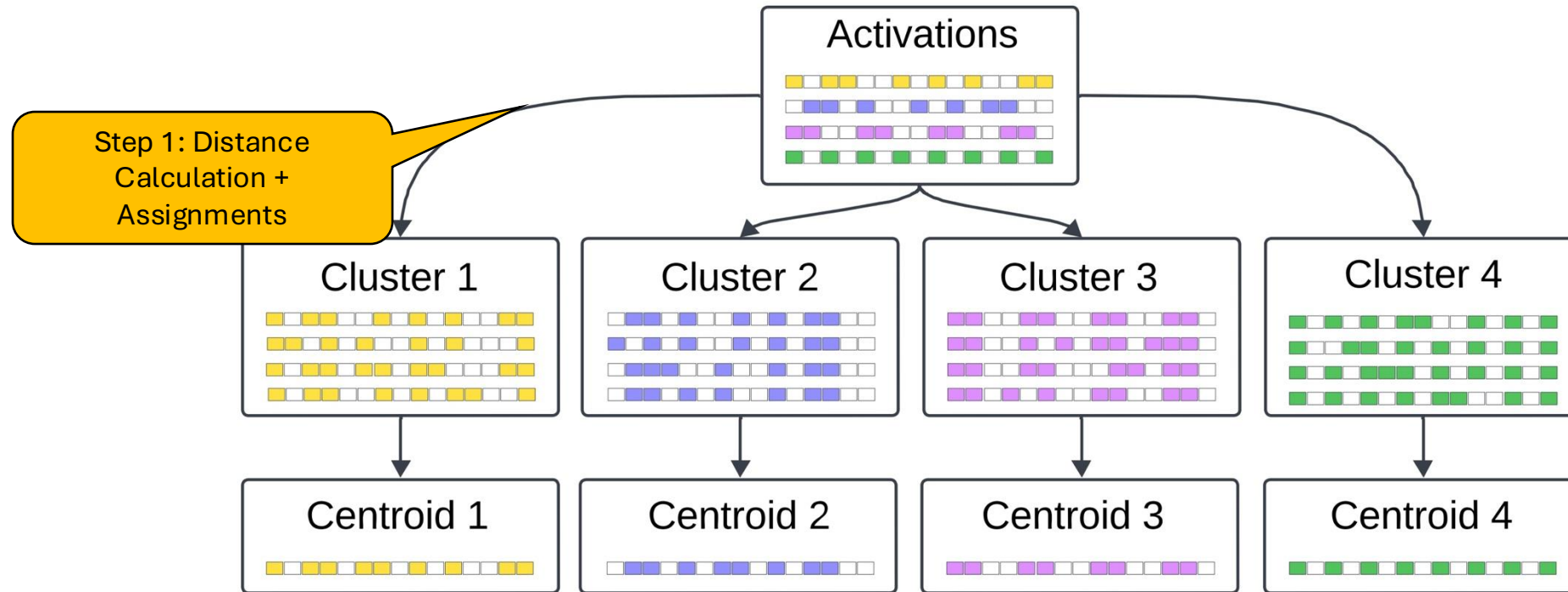
Experimental Setup

- Dataset
 - Wikitext-2
 - 333,842 number of tokens
 - 10.64 million activation patterns
- Model
 - Mistral-7B-v0.1
 - Number of FFN layers: 32
 - Each FFN layer consists of projection layers:
 - *Gate-Proj*: 14336 neuron
 - *Up-Proj*: 14336 neuron
 - *Down-Proj*: 4096 neuron
 - 50% model accuracy (PPL Score): 6.45
- Hardware used for processing:
 - 8 * Nvidia A100
 - 80GB GPU memory each

Activation-Aware Clustering (APC)



Methodology



Methodology

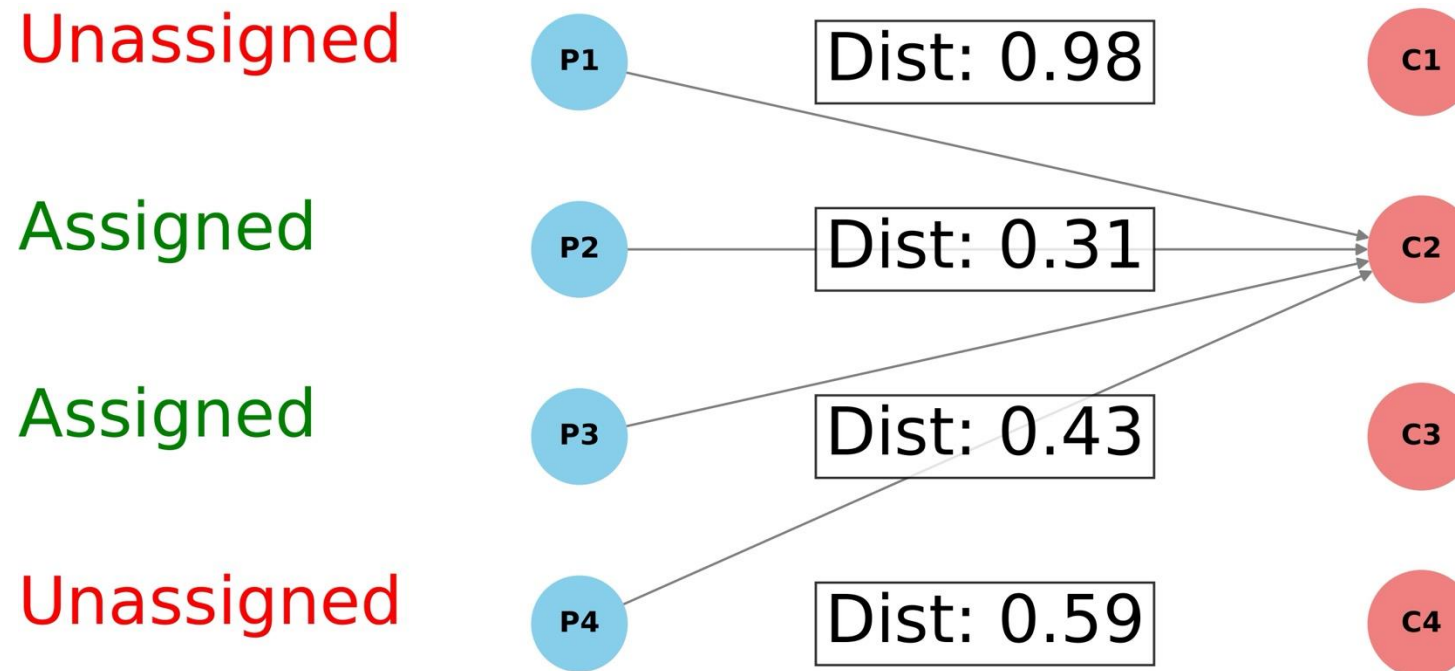
Step 1.1: Distance Calculation

Datapoint	1	0	1	1	0	0
Centroid	0	1	1	0	1	0
Considered Positions	1	0	1	1	0	0
Computed Differences	1	0	0	1	0	0

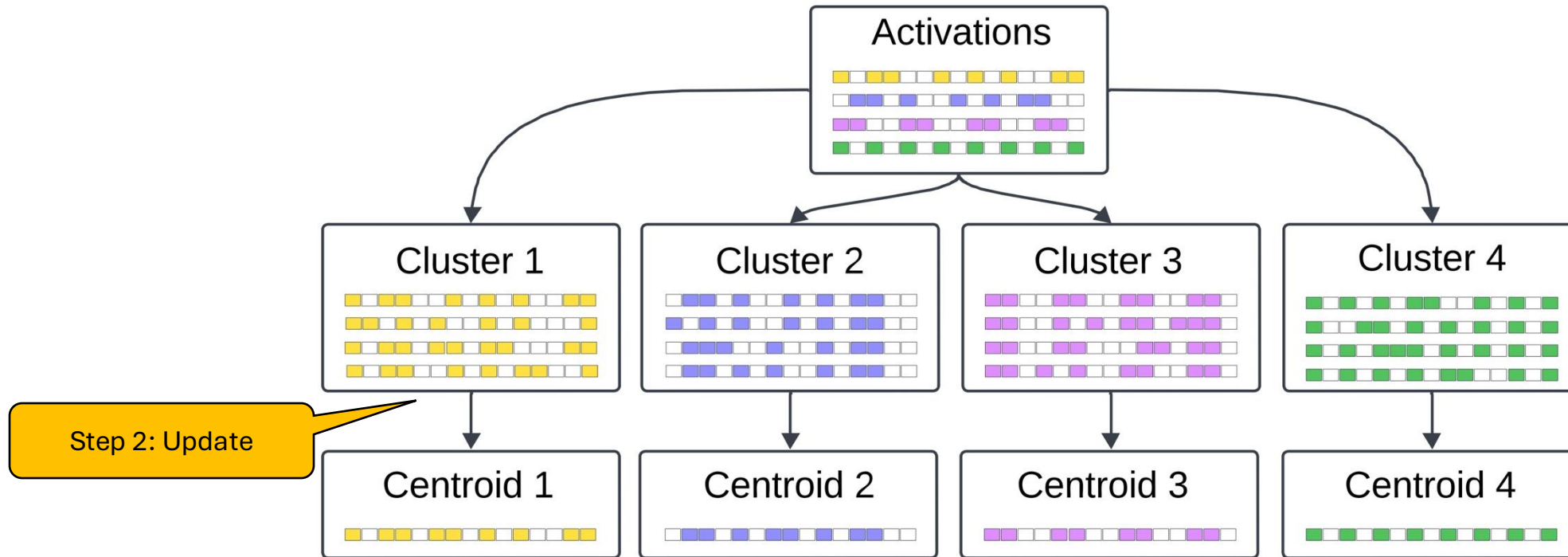
Inactive position disregarded

Methodology

Step 1.2: Assignment

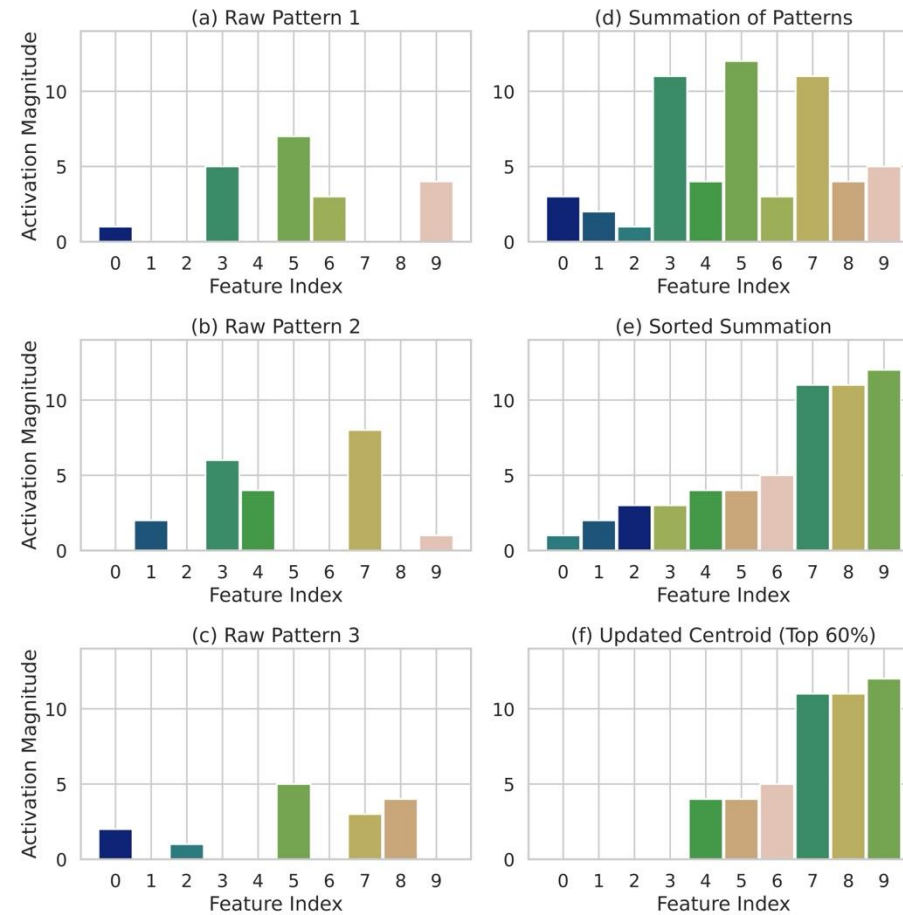


Methodology



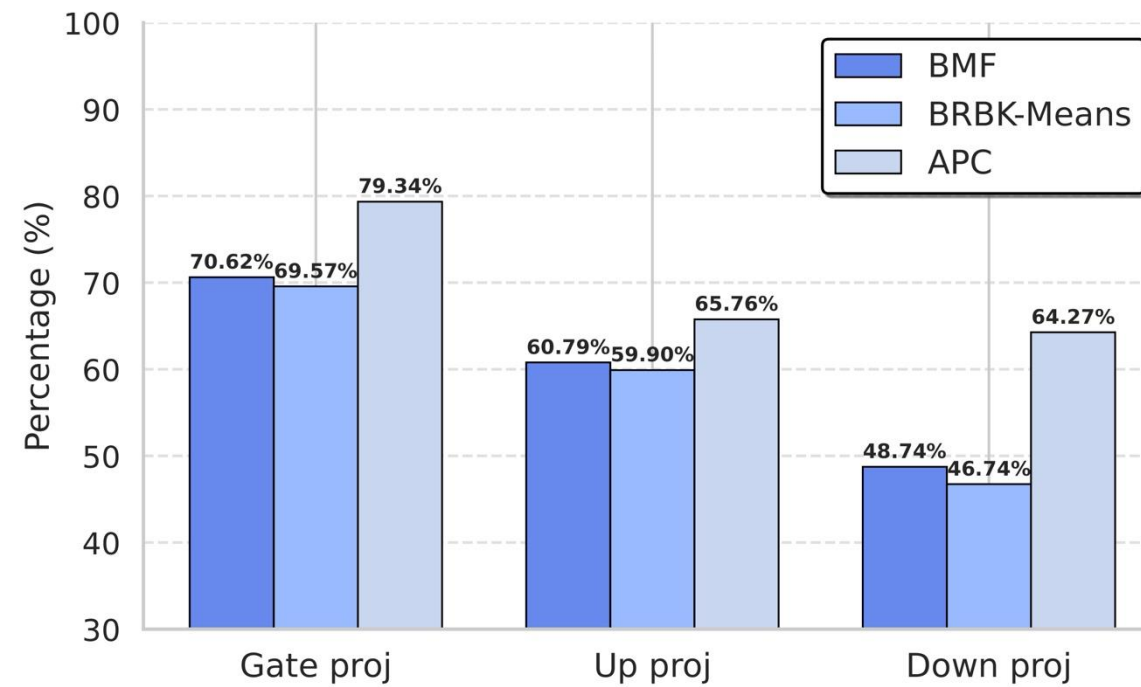
Methodology

Step 2: Update Centroid



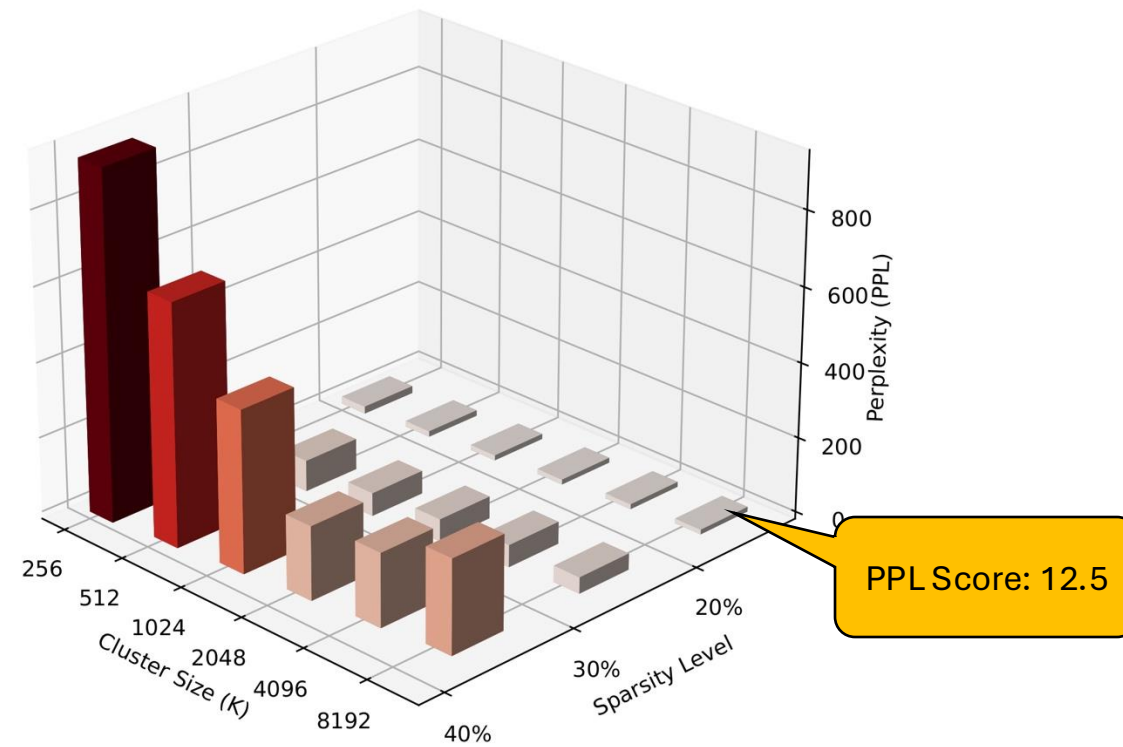
Result

1: Clustering Precision



Result

2: Clustering Effect On The Model Accuracy



Result

2: Clustering Effect On The Model Accuracy

