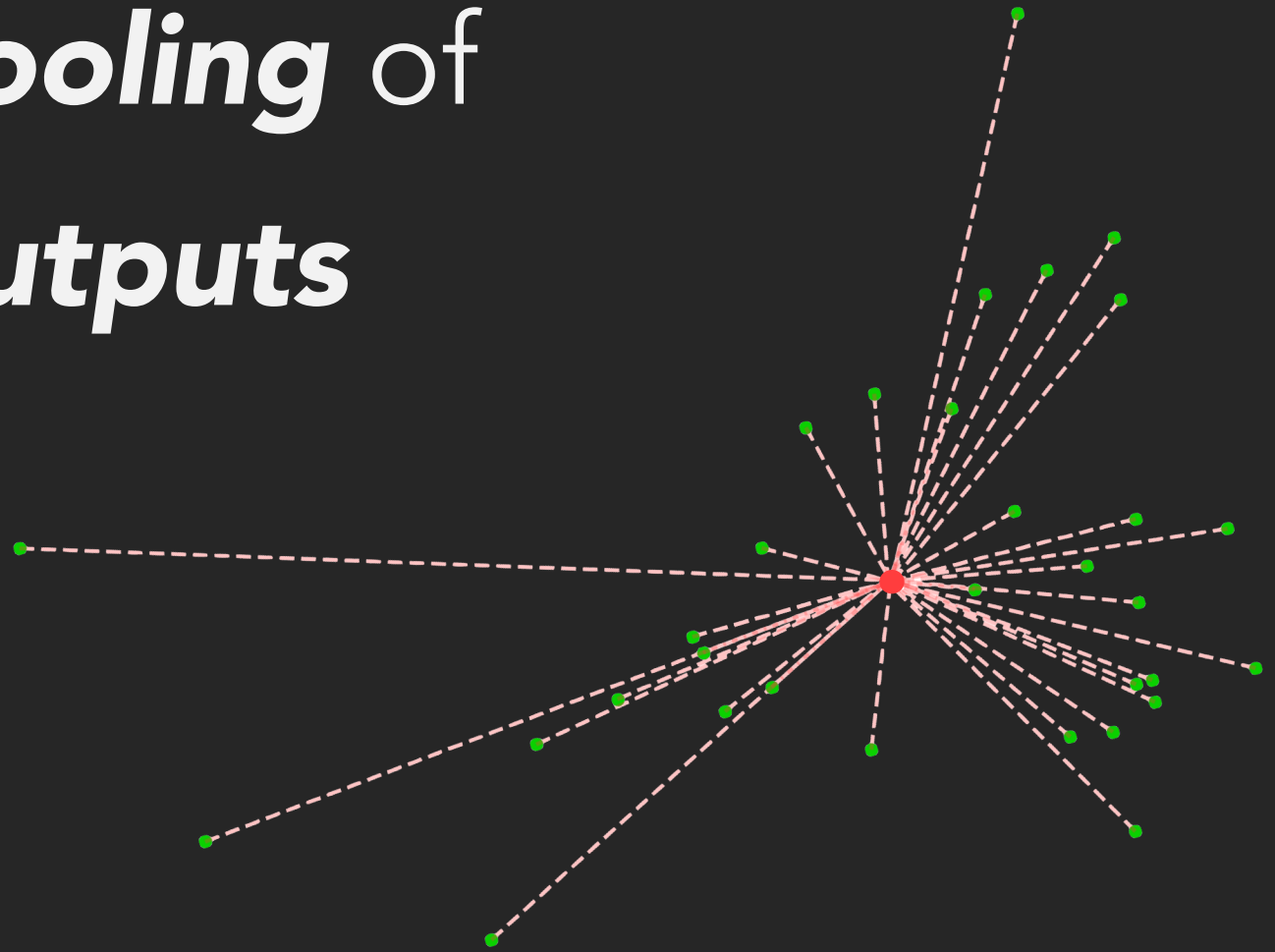
Robust Noise Attenuation via ***Adaptive Pooling*** of ***Transformer Outputs***

ICML 2025



Greyson Brothers

Johns Hopkins University

Johns Hopkins Applied Physics Laboratory

BACKGROUND

Transformer-based embedding models are broadly used to create rich representations of data in many domains

BACKGROUND

Transformer-based embedding models are broadly used to create rich representations of data in many domains

- Images
- Audio
- Text
- Graphs

BACKGROUND

1. Embed Tokens

Transformer-based embedding models are broadly used to create rich representations of data in many domains

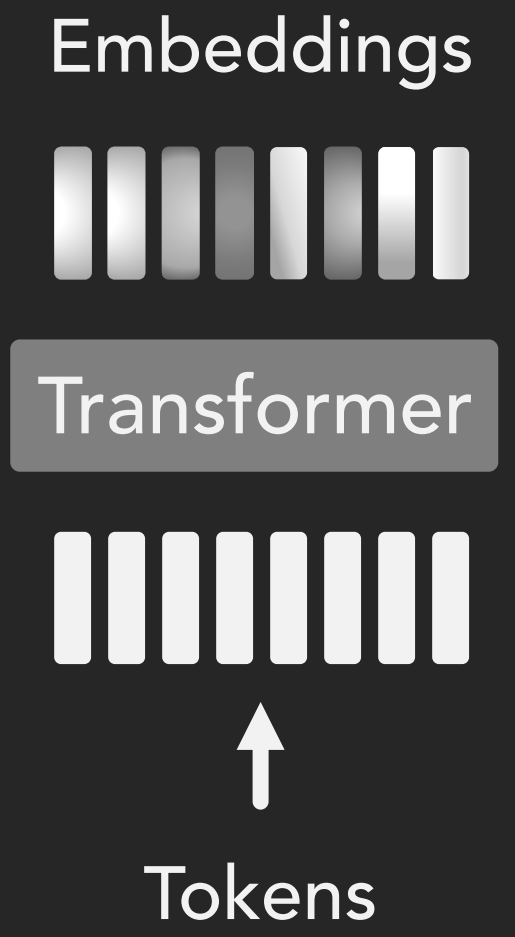
- Images
- Audio
- Text
- Graphs

BACKGROUND

1. Embed Tokens

Transformer-based embedding models are broadly used to create rich representations of data in many domains

- Images
- Audio
- Text
- Graphs

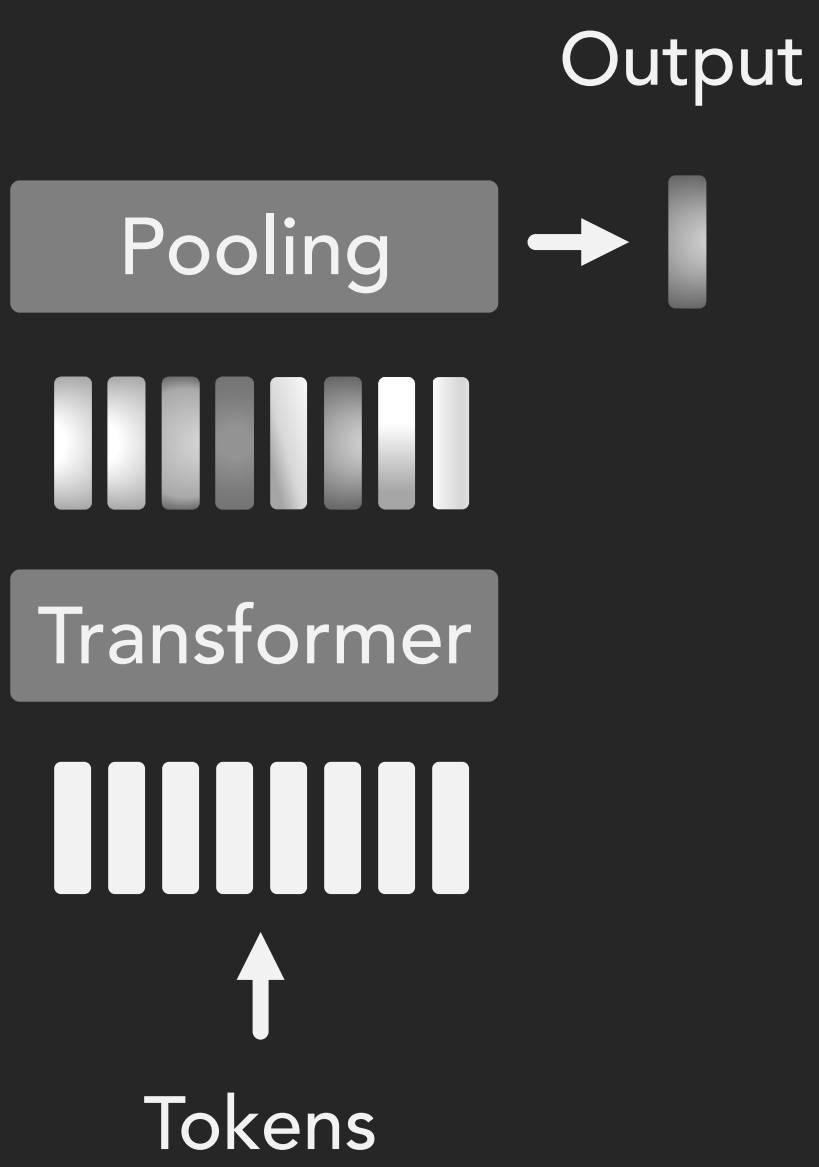


BACKGROUND

2. Summarize Embeddings

The standard methods for aggregating embeddings are:

- *ClsToken*
- *AvgPool*
- *MaxPool*



THE RESEARCH QUESTION

1. Why choose one pooling method over another?

THE RESEARCH QUESTION

1. Why choose one pooling method over another?

    > *Better empirical and/or compute performance*

THE RESEARCH QUESTION

1. Why choose one pooling method over another?

    > *Better empirical and/or compute performance*

2. What is the best possible method of summarizing embeddings?

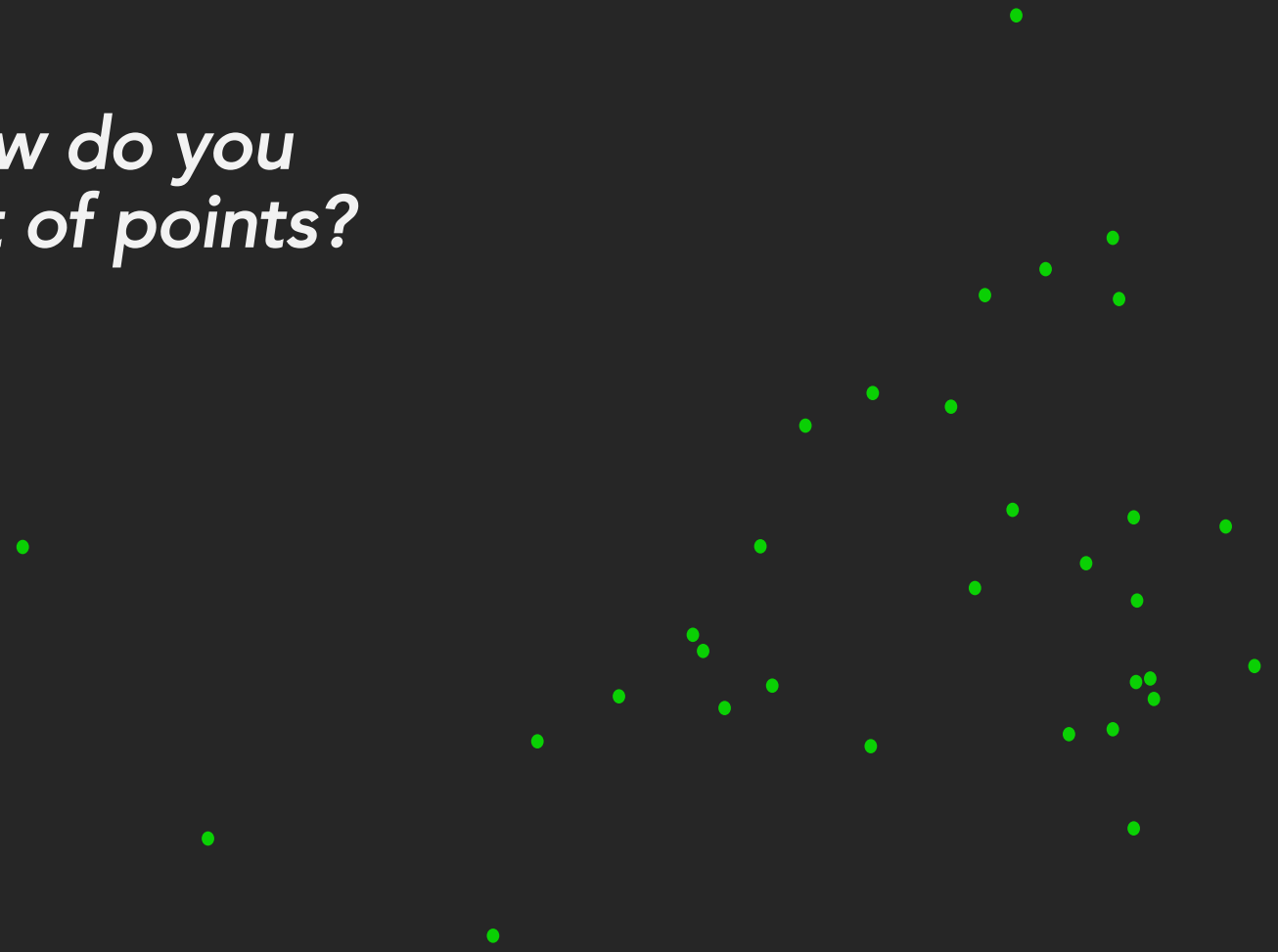
THE HYPOTHESIS

If we derive a general metric for pooling effectiveness,

then we can formally compare existing pooling methods and *potentially* define an optimal approach

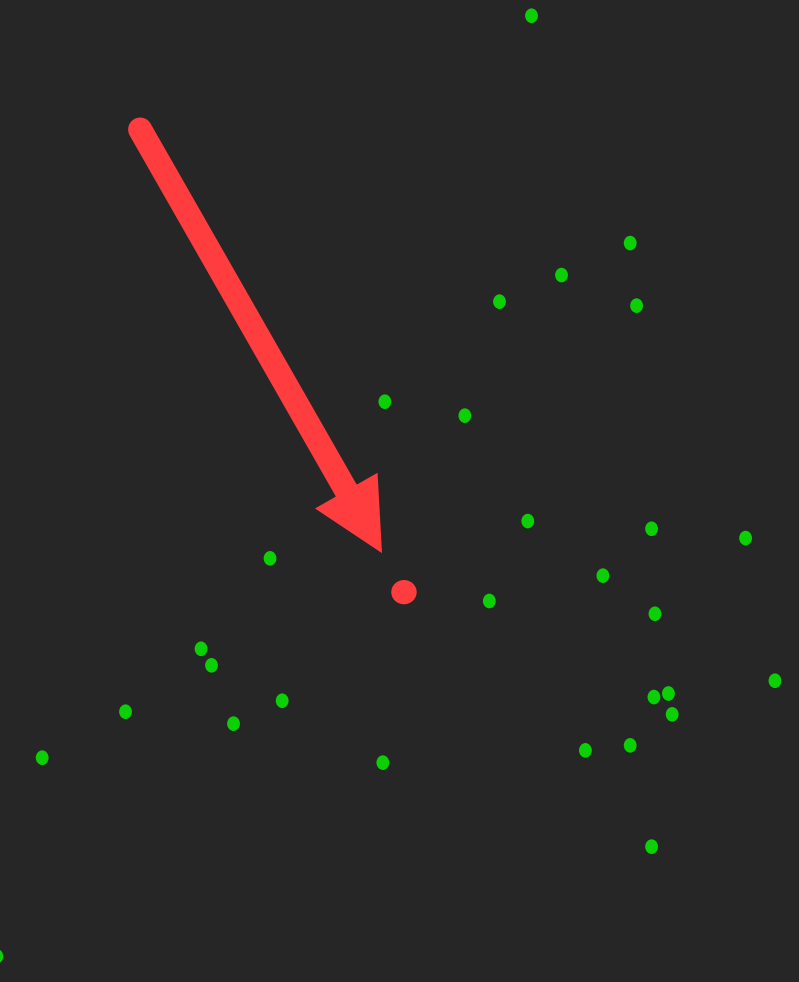
THE METRIC

Vector quantization – how do you optimally compress a set of points?



THE METRIC

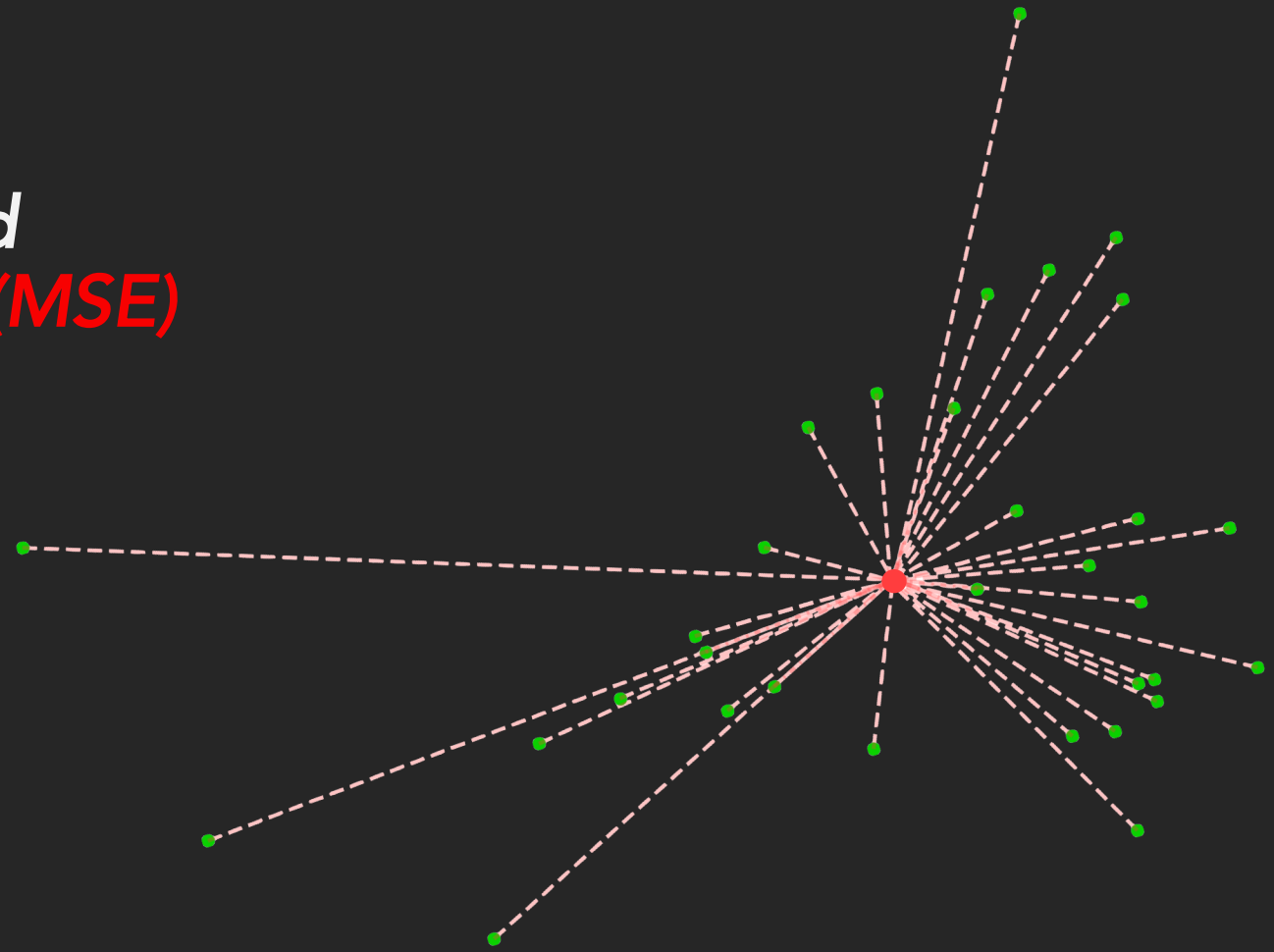
Vector quantization – how do you optimally compress a set of points?



Represent them with fewer, well-chosen points. In our case, a single point.

THE METRIC

Information lost during compression is measured via Mean Squared Error (MSE)

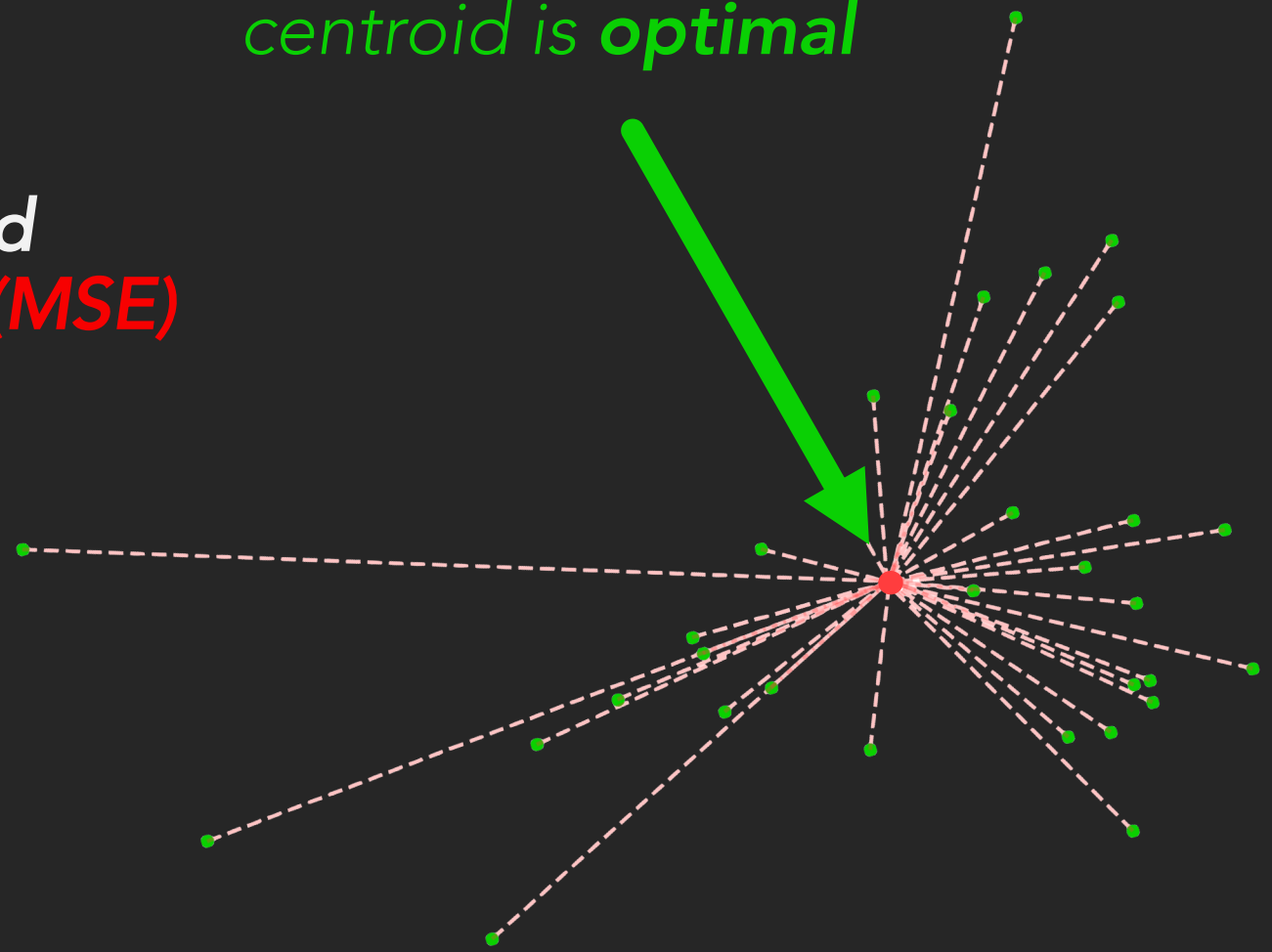


Error shown by red lines

THE METRIC

Information lost during Compression is measured via Mean Squared Error (MSE)

*For MSE, the centroid is **optimal***



Error shown by red lines

SIGNAL vs NOISE

The embeddings we want to aggregate are not equally important!

SIGNAL vs NOISE

The embeddings we want to aggregate are not equally important!

Those which are crucial for the downstream task are ***signal***

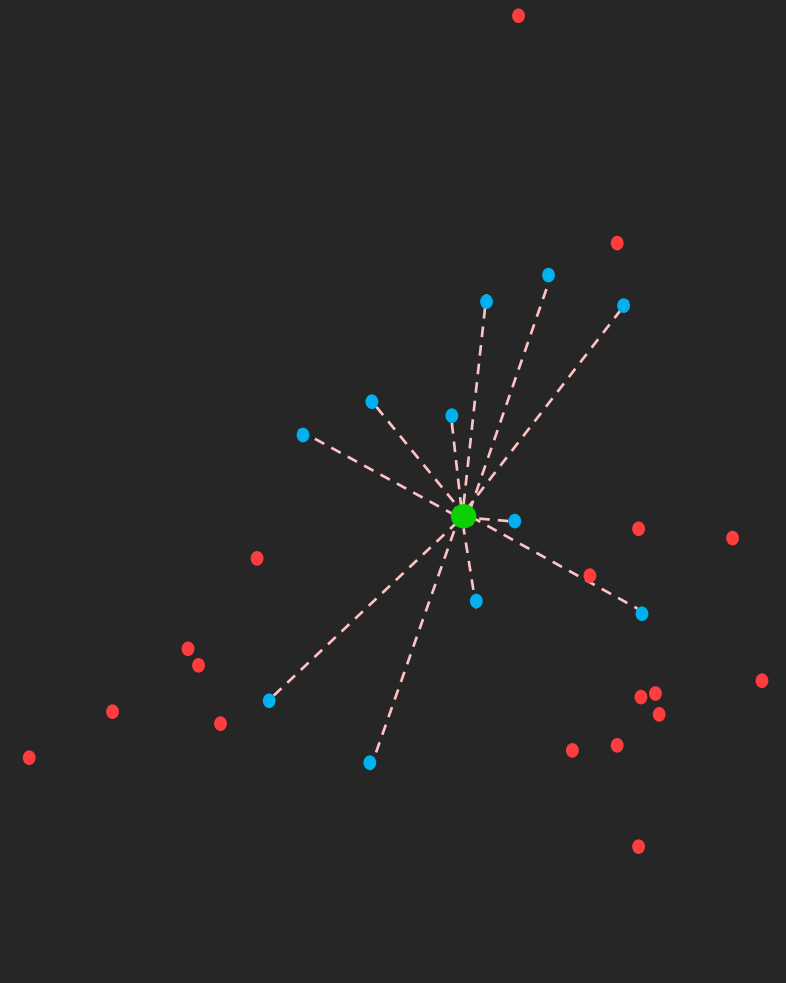
The rest are ***noise*** and should be attenuated

THE FINAL METRIC

We care only about *signal loss*, the MSE between signal vectors and the aggregate representation

THE FINAL METRIC

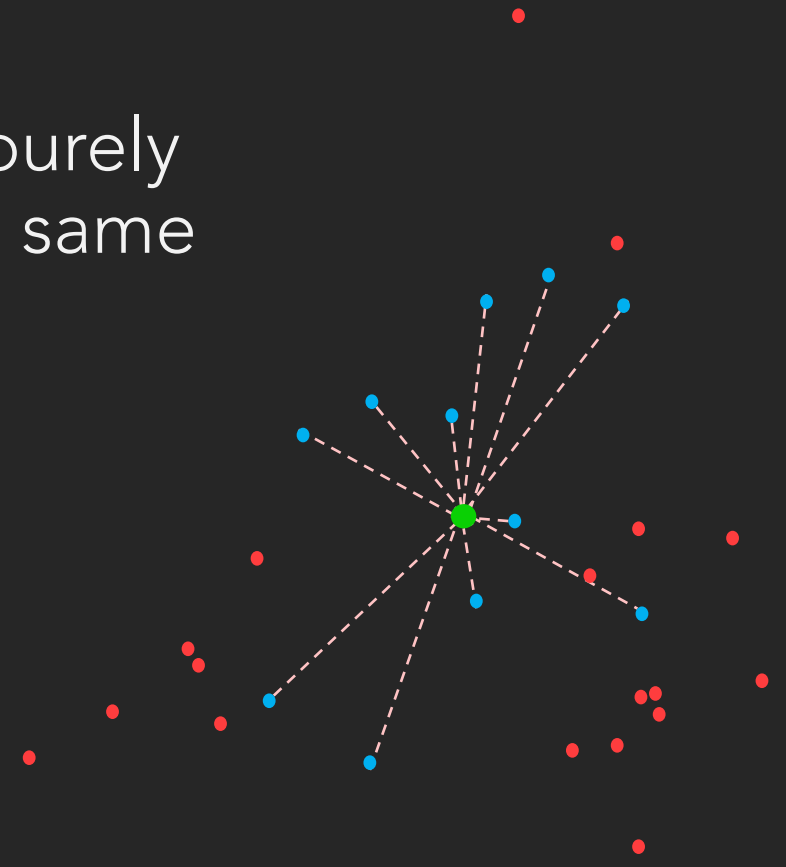
We care only about *signal loss*, the MSE between signal vectors and the aggregate representation



The optimal representation is the centroid of the signal subset

THEORETICAL RESULTS

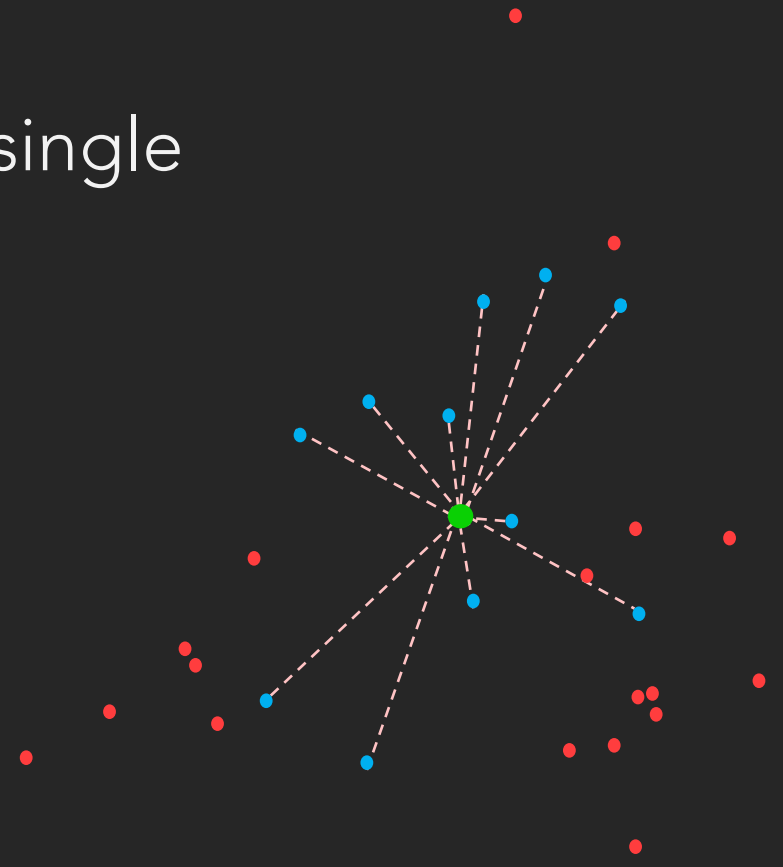
AvgPool is only optimal when the set is purely signal, or when signal and noise have the same centroid.



Signal loss tends to increases with each additional noise vector

THEORETICAL RESULTS

MaxPool is only optimal when there is a single signal vector dominating in all features



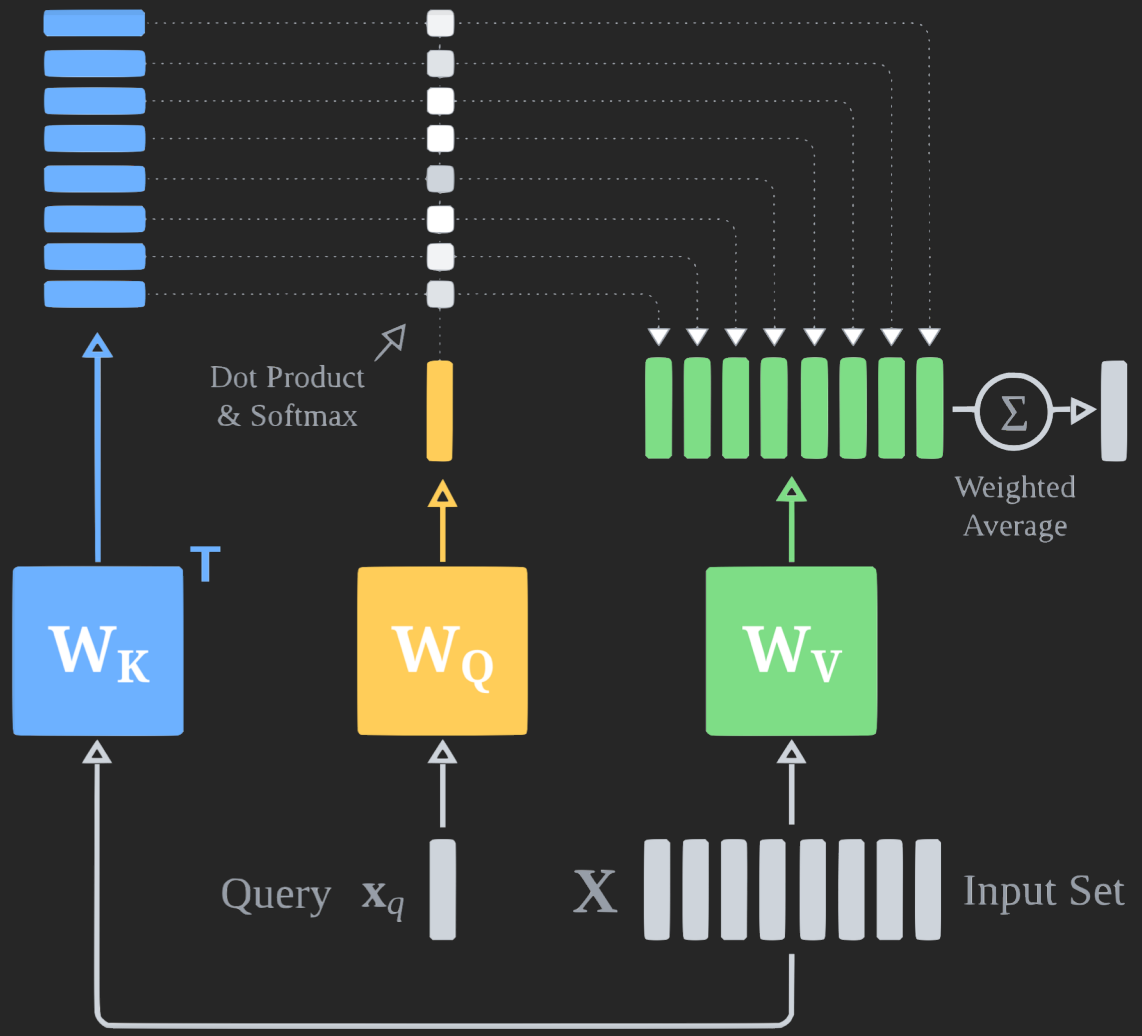
Signal loss tends to increases with each additional signal vector

THE GOLDILOCKS METHOD

Ideally, we want a pooling method that can adapt to varying quantities of signal and noise

ATTENTION-BASED POOLING

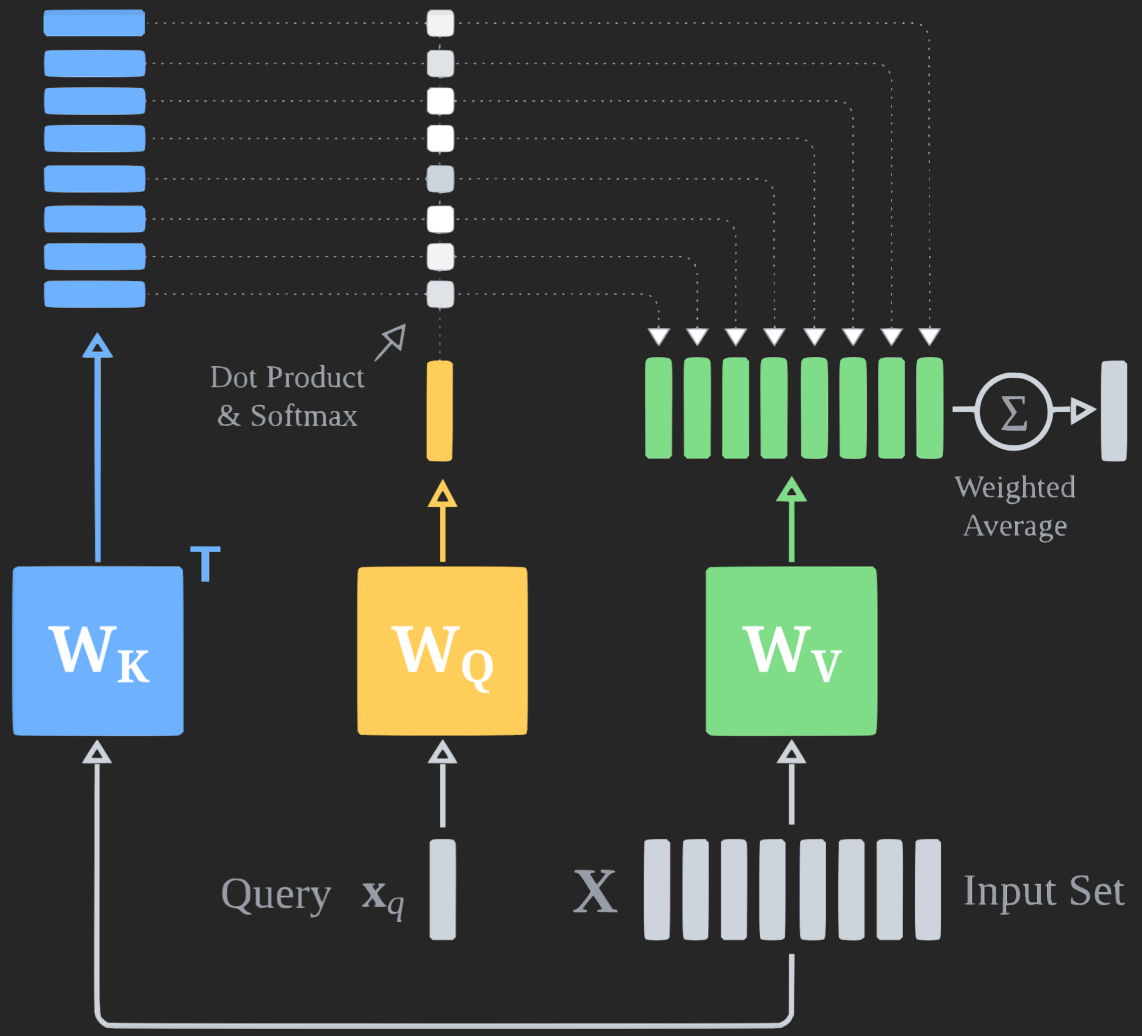
Adaptive Pooling (**AdaPool**) uses cross-attention with a single query to aggregate



ATTENTION-BASED POOLING

Adaptive Pooling (**AdaPool**) uses cross-attention with a single query to aggregate

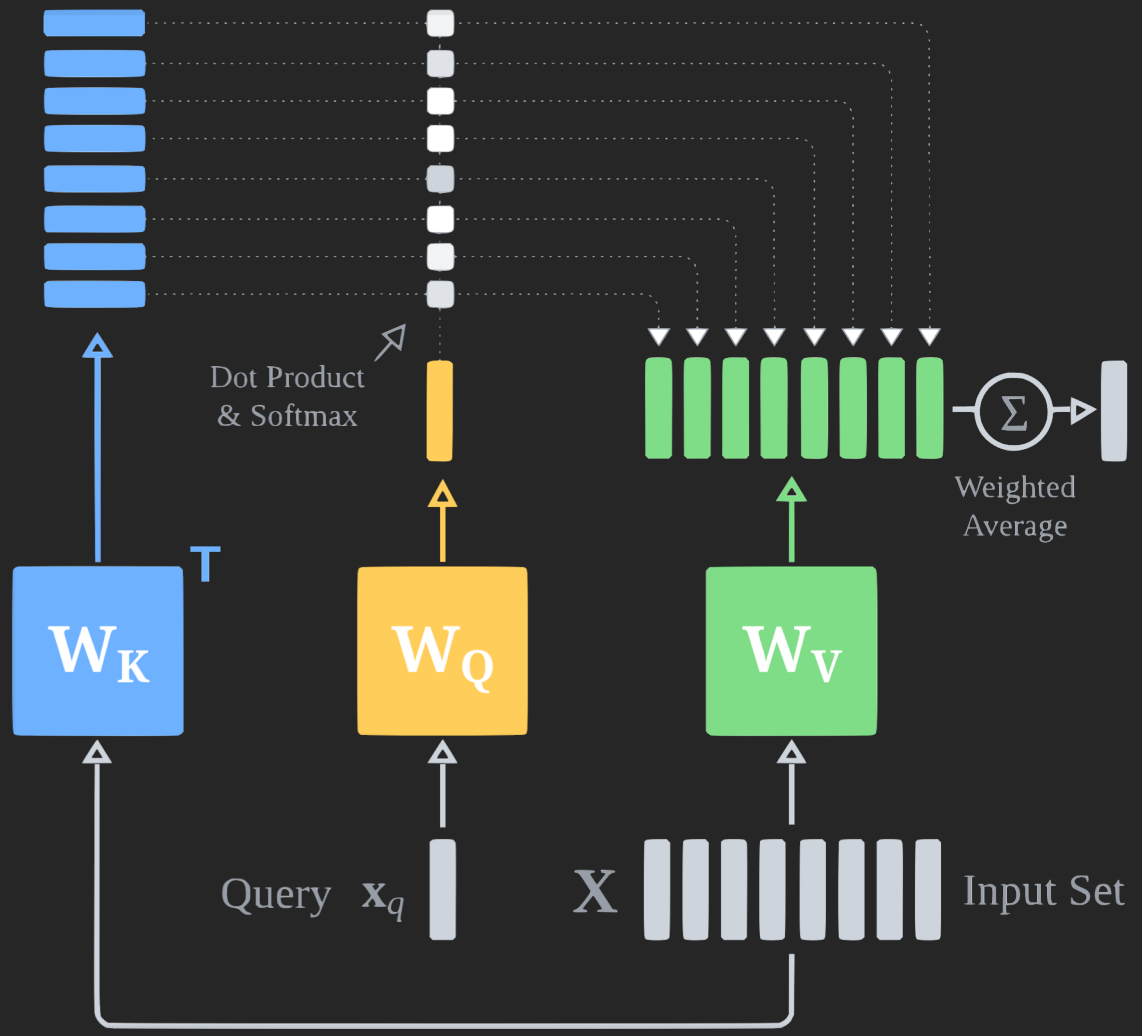
*We show that **AdaPool** approximates the optimal method within derived error bounds*



OTHER THEORETICAL RESULTS

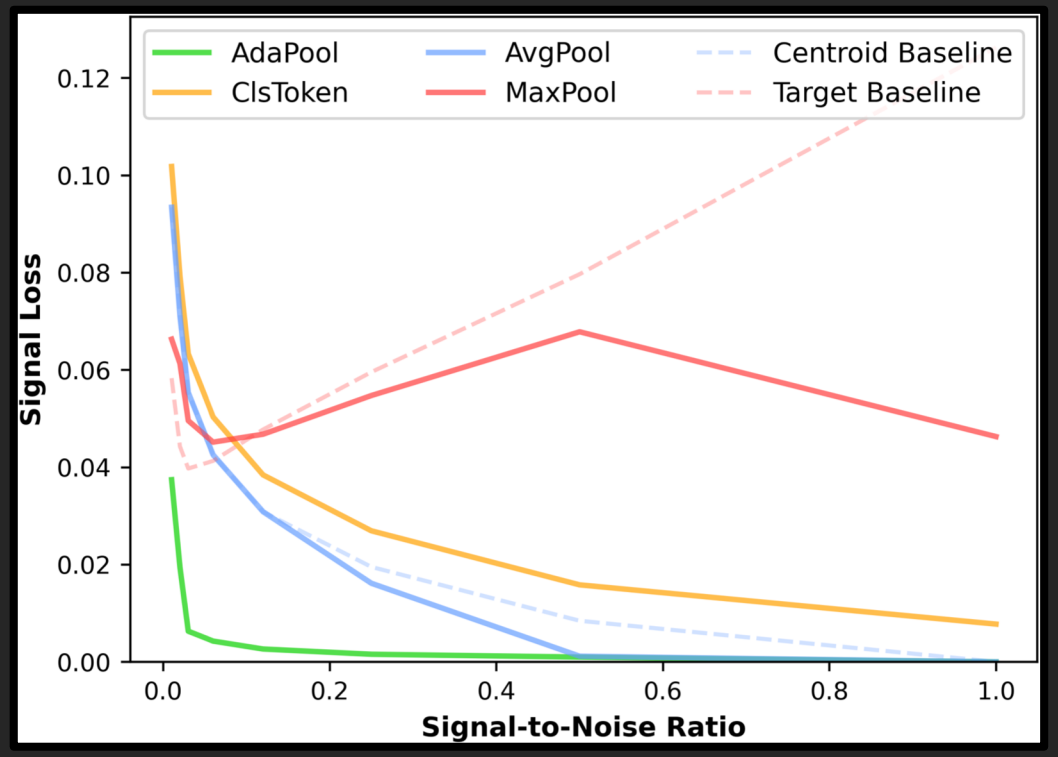
AvgPool & **MaxPool** are special cases of **AdaPool**

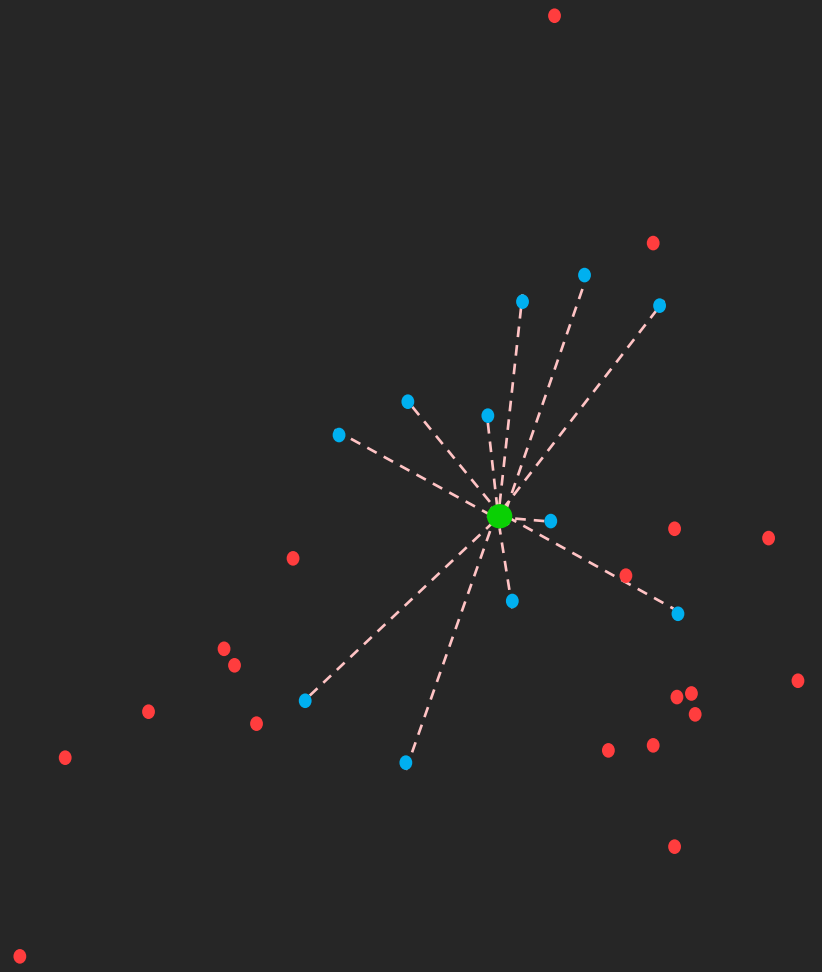
ClsToken is very similar to a case of **AdaPool** where the query vector is learned and embedded via transformer



EMPIRICAL RESULTS

kNN-Centroid Task





THANK YOU!

Please refer to the paper for more detail and additional experiments