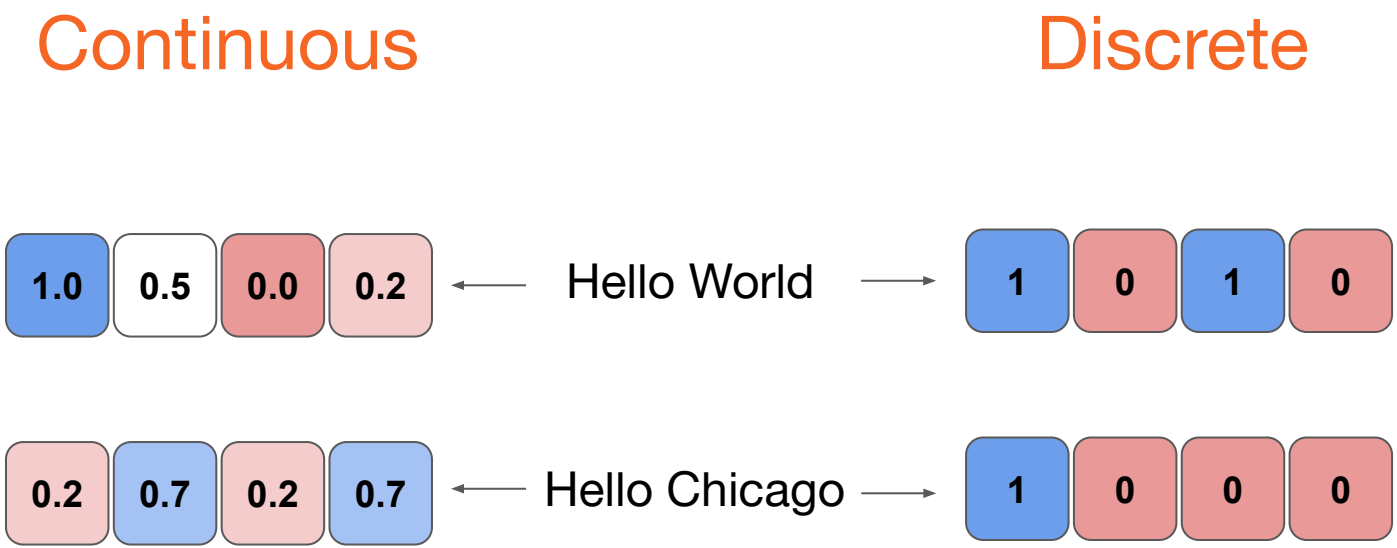
Discrete Latent Variable Representations for Low-Resource Text Classification

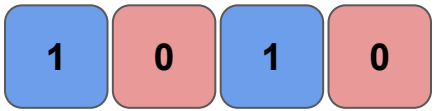
Shuning Jin[1], Sam Wiseman[2], Karl Stratos[1], Karen Livescu[2]

[1] Rutgers University [2] Toyota Technological Institute at Chicago

https://github.com/shuningjin/discrete-text-rep


Continuous
Discrete
1.0
0.5
0.0
0.2
Hello World
1
0
1
0
0.2
0.7
0.2
0.7
Hello Chicago
1
0
0
0

Why Discrete



Compression: store integer

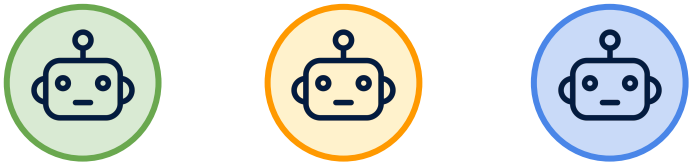
Interpretability: meaningful clusters

This Work

Learn discrete representation

- Compare three discrete VAEs

VQ-VAE CatVAE Hard EM



Use discrete representation

- Classification & Document retrieval

This Work

Learn discrete representation

- Compare three discrete VAEs

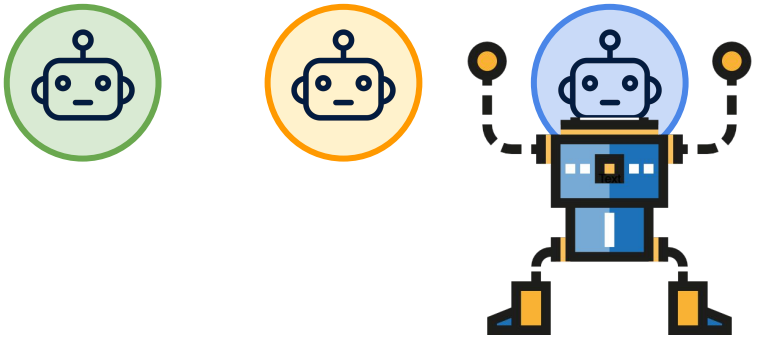
Hard EM is strong

Use discrete representation

- Classification & Document retrieval

Discrete matches continuous

VQ-VAE CatVAE Hard EM

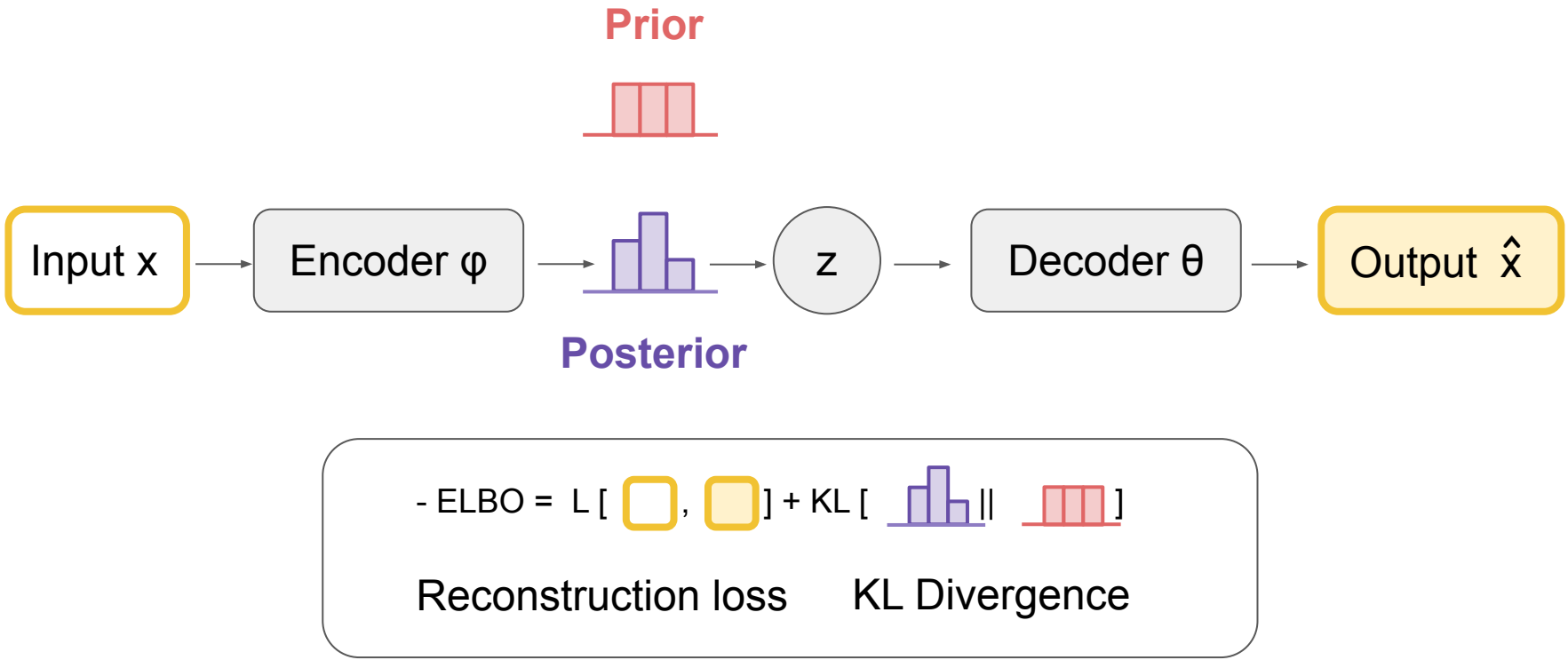


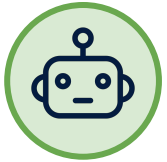
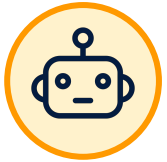
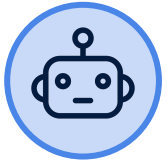
Goal

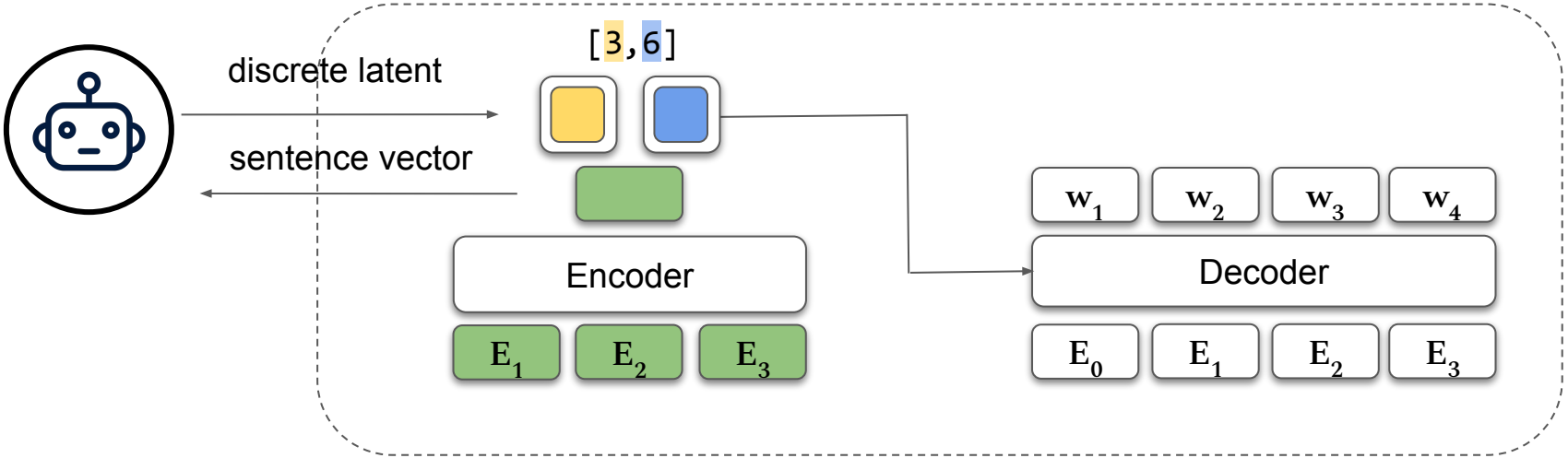
→ How to get discrete representation?

How to use discrete representation?

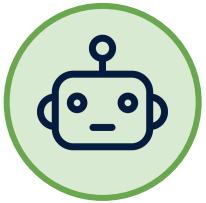
Variational Autoencoder (VAE)

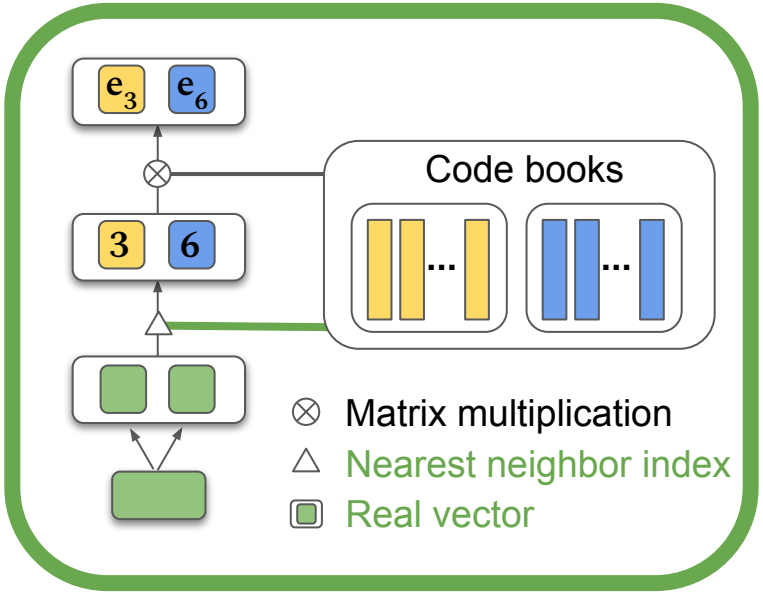


Discrete VAE	Posterior	- ELBO = L [,] + KL [||] Objective
Vector Quantized VAE (VQ-VAE) 	Point mass	- ELBO (constant KL) + VQ loss
Categorical VAE (CatVAE) 	Categorical	- ELBO
Hard Expectation Maximization (Hard EM) 	Point mass	- ELBO (constant KL)

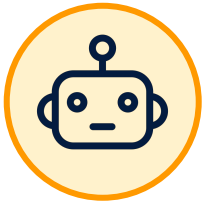

discrete latent
sentence vector
[3,6]
Encoder
E1
E2
E3
Decoder
w1
w2
w3
w4
E0
E1
E2
E3

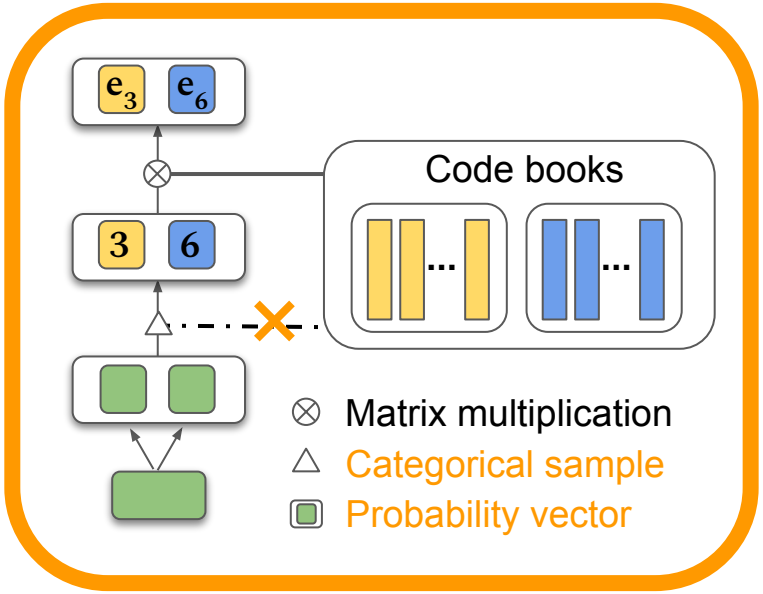
VQ-VAE



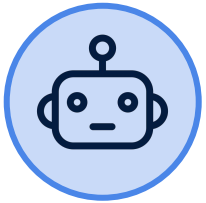

e3
e6
3
6
Code books
Matrix multiplication
Nearest neighbor index
Real vector

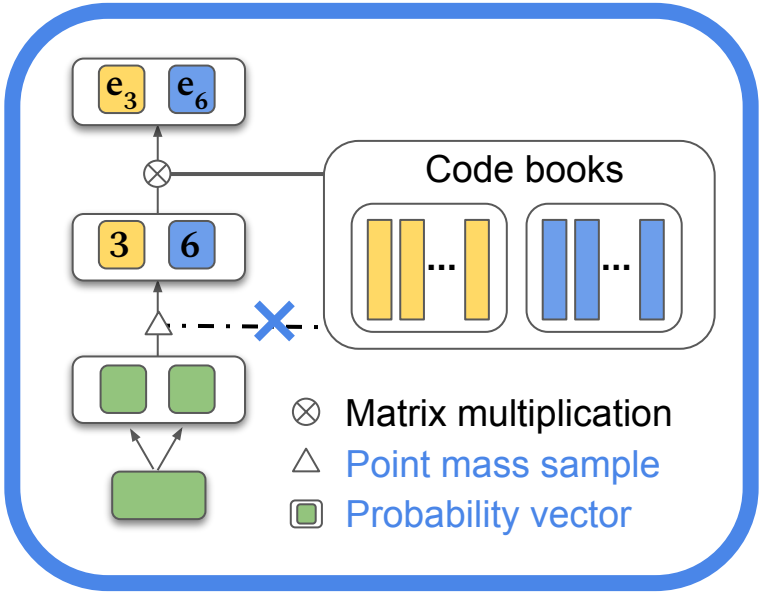
CatVAE



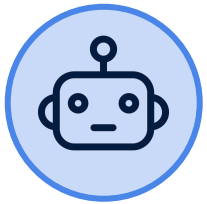

e3
e6
3
6
Code books
Matrix multiplication
Categorical sample
Probability vector

Hard EM

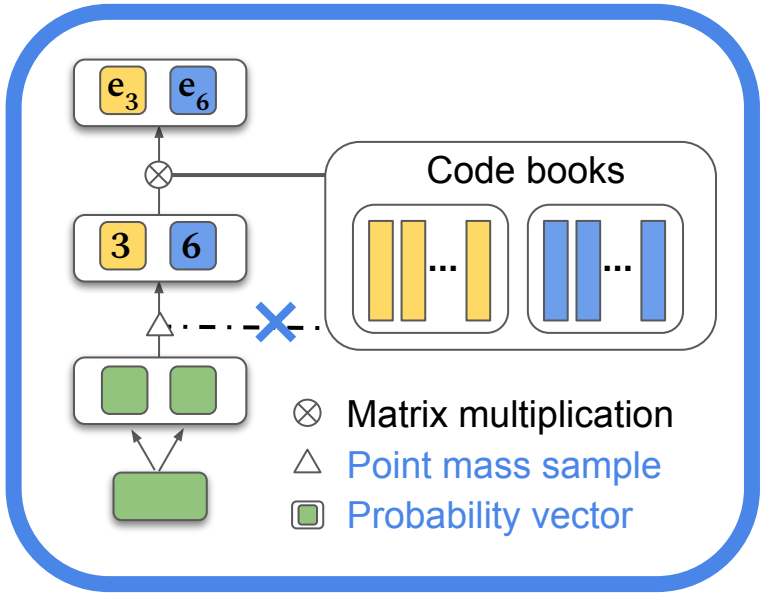


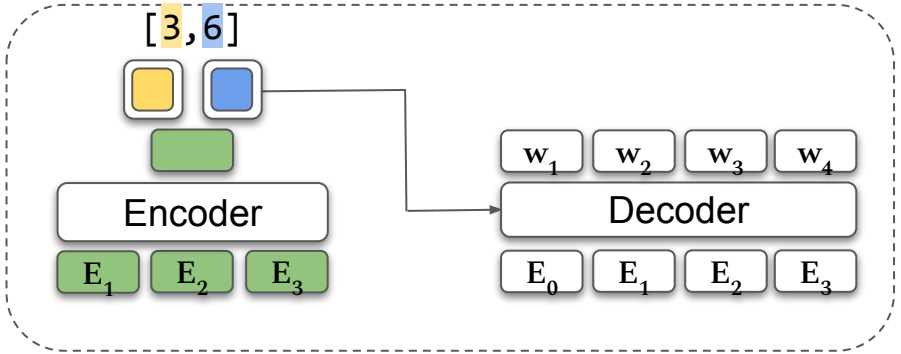

e_3
e_6
3
6
Code books
Matrix multiplication
Point mass sample
Probability vector

Hard EM

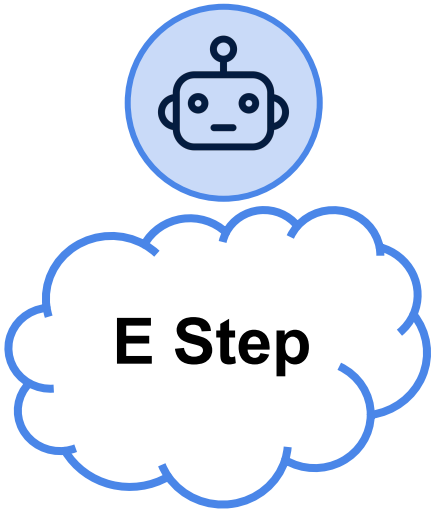


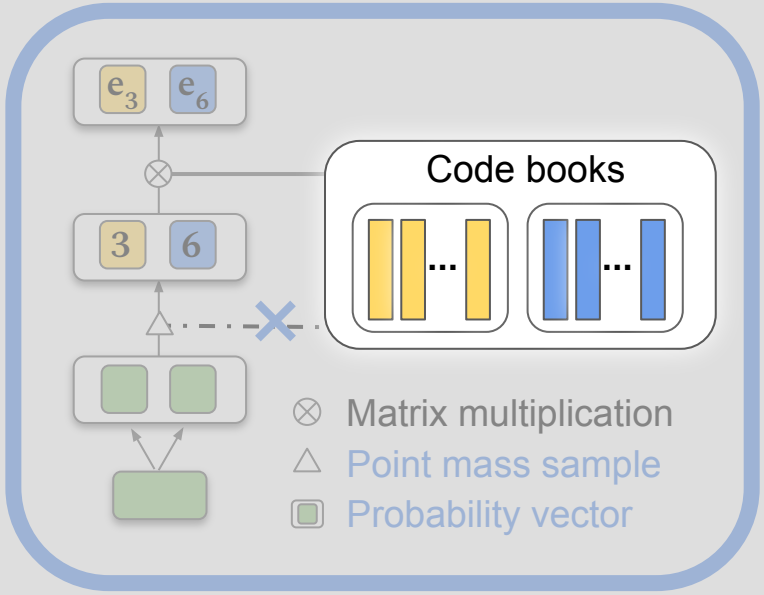
Alternating optimization

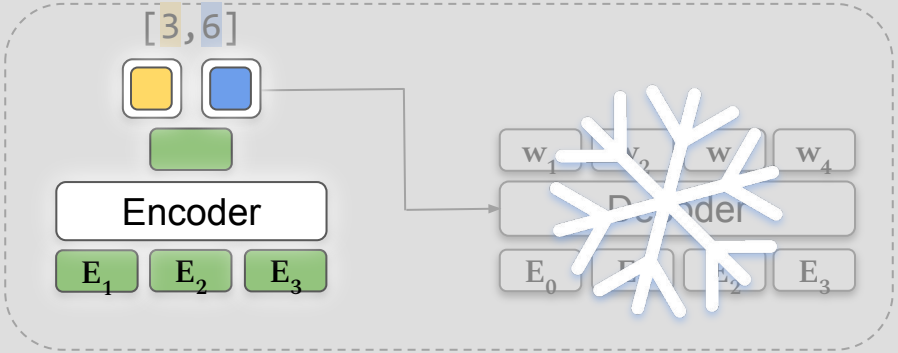




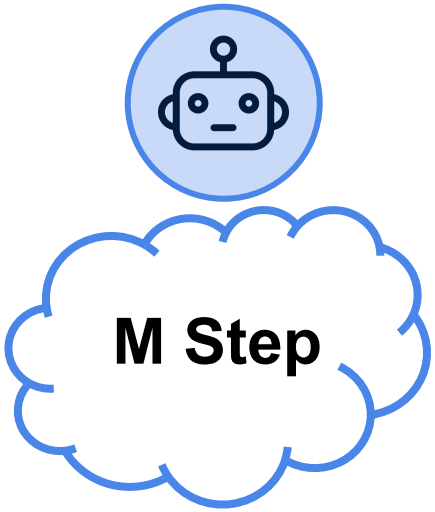
Hard EM

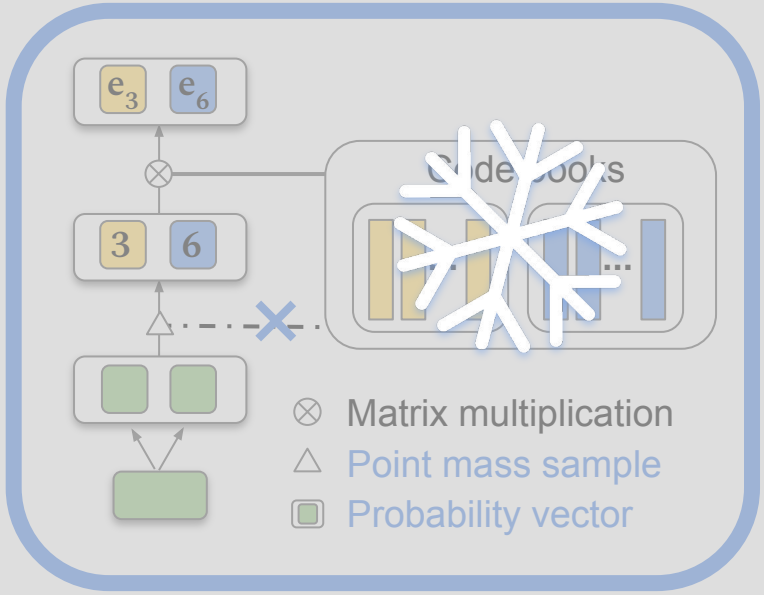

E Step

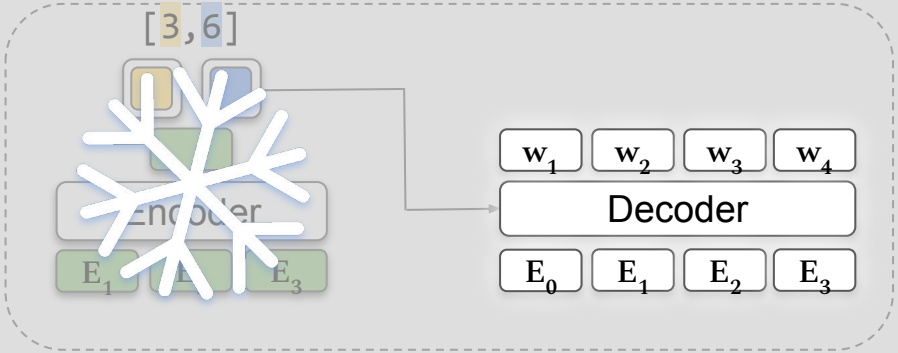

e3
e6
3
6
Code books
Matrix multiplication
Point mass sample
Probability vector


[3,6]
Encoder
E1
E2
E3
w1
w4
E0
E3

Hard EM


M Step


e3 e6
3 6
Matrix multiplication
Point mass sample
Probability vector


[3,6]
w1 w2 w3 w4
Decoder
E0 E1 E2 E3

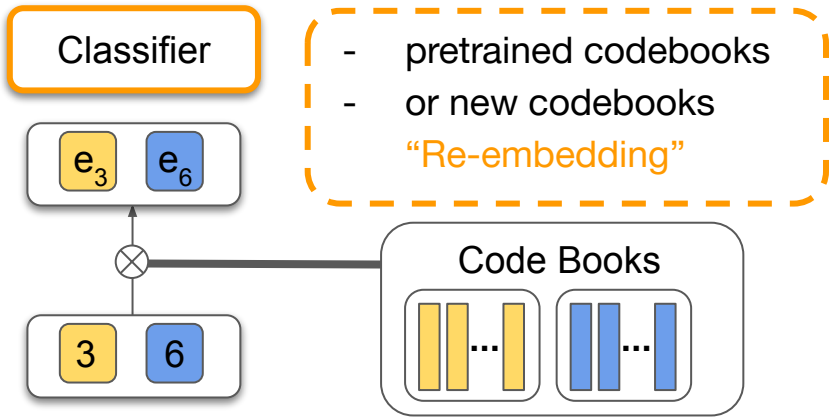
Goal

How to get discrete representation?

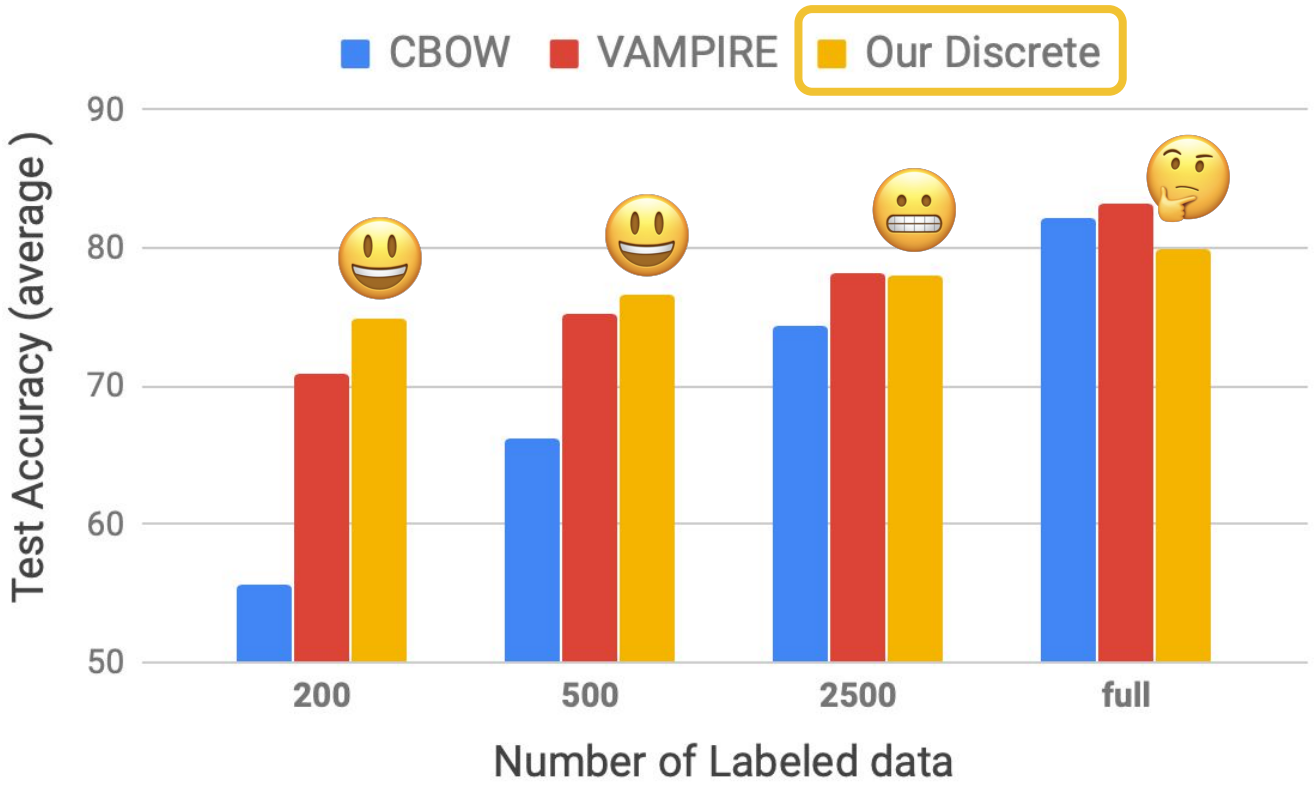
→ How to use discrete representation?

Use Case - Classification

- Codebook as integer's embedding
- Train a lightweight classifier with a few labeled data, for low resource

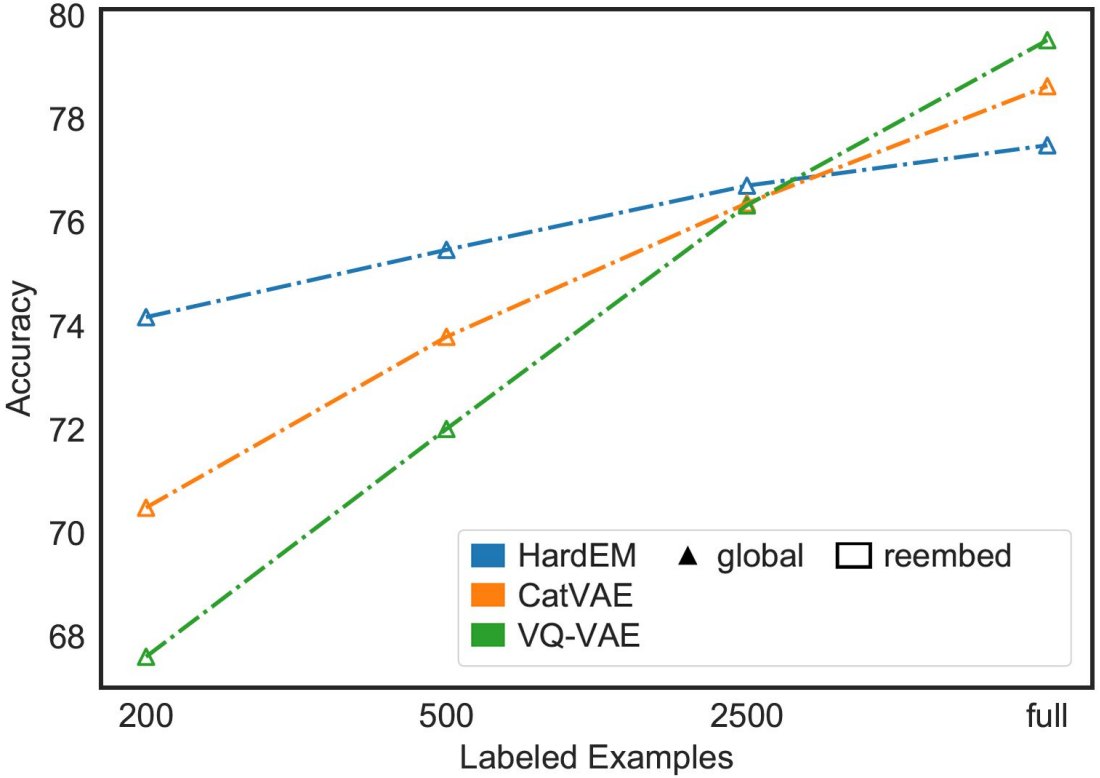


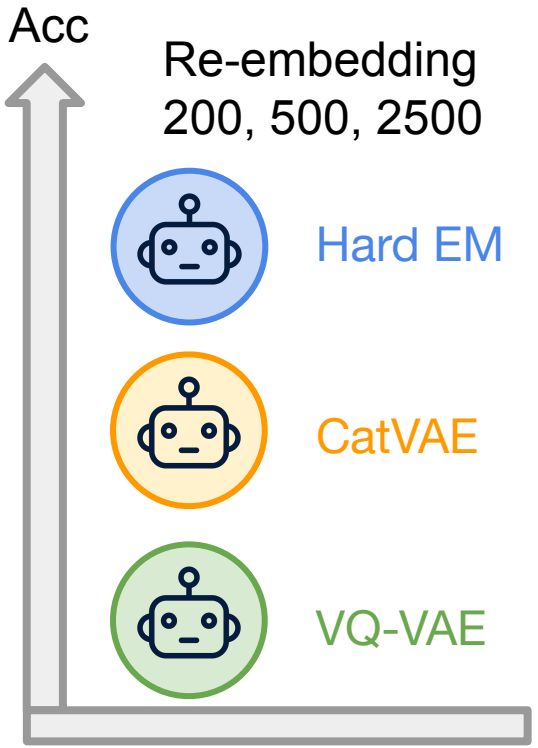
Classification: discrete matches continuous



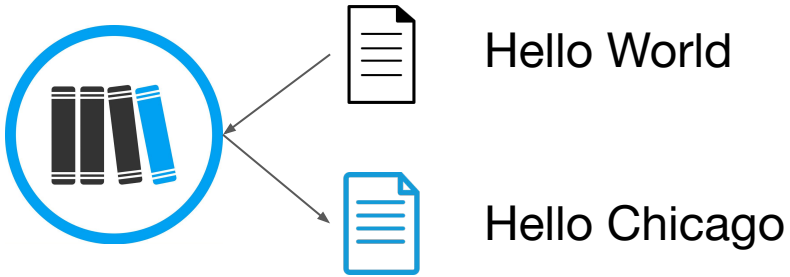
Gururangan et al., 2019. Variational Pretraining for Semi-supervised Text Classification (VAMPIRE)

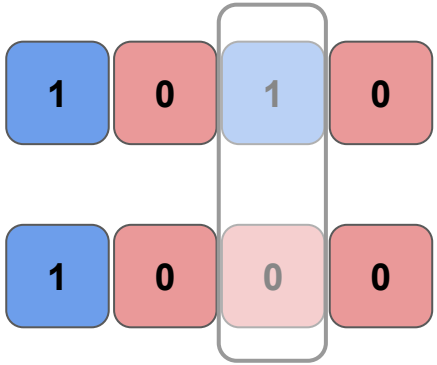
Classification: compare discrete





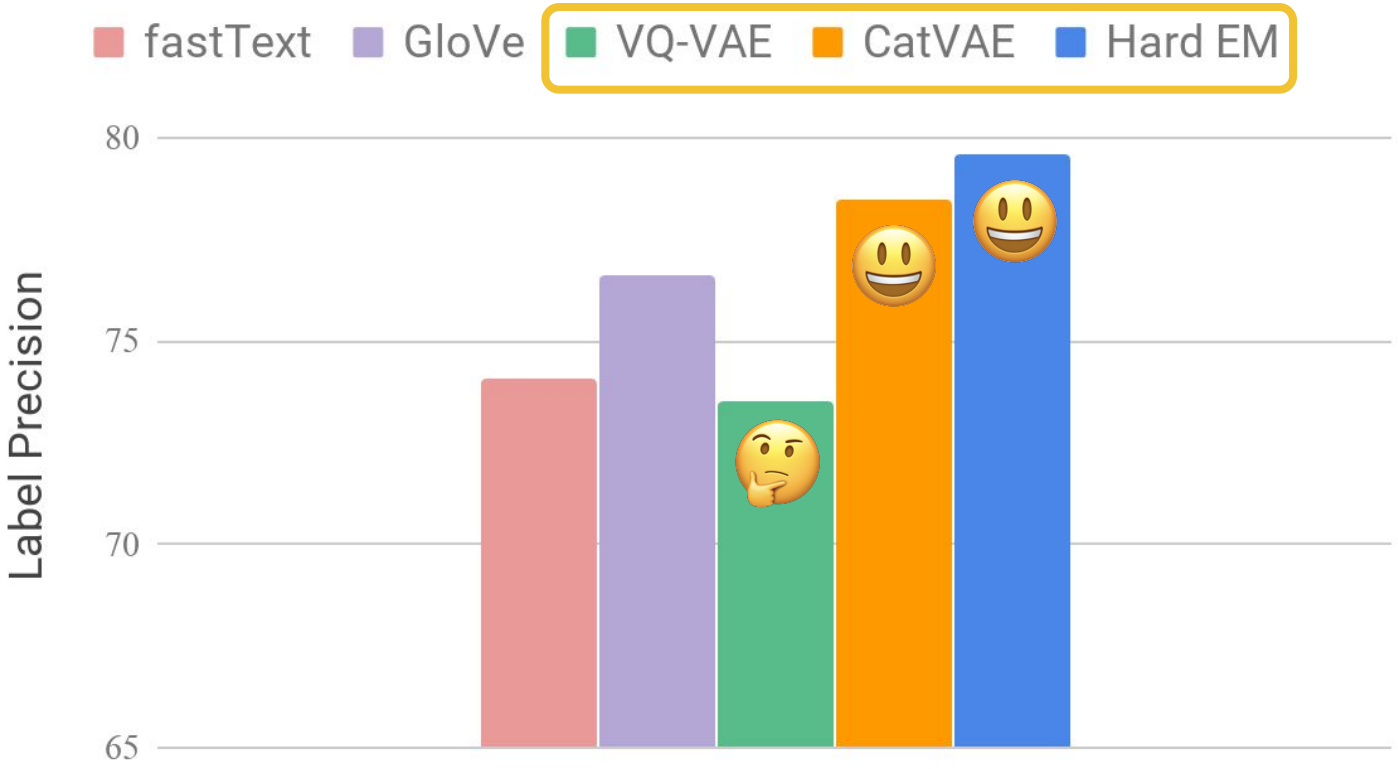
Use Case - Nearest Neighbor Based Retrieval



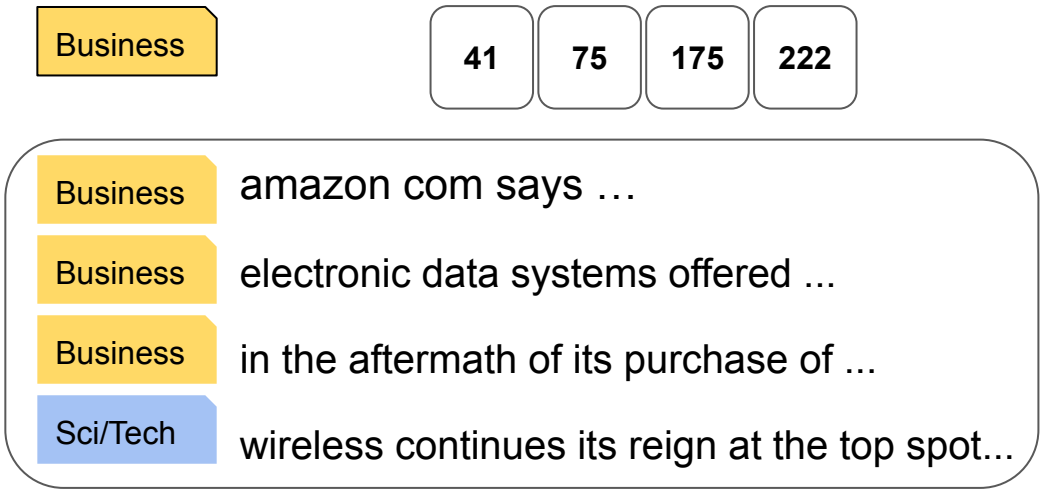


Hamming Distance = 1

Retrieval: discrete matches continuous



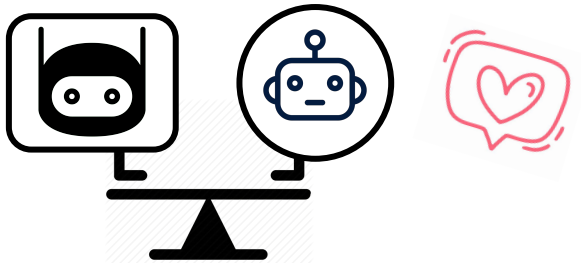
Analysis - Interpretable Topical Cluster



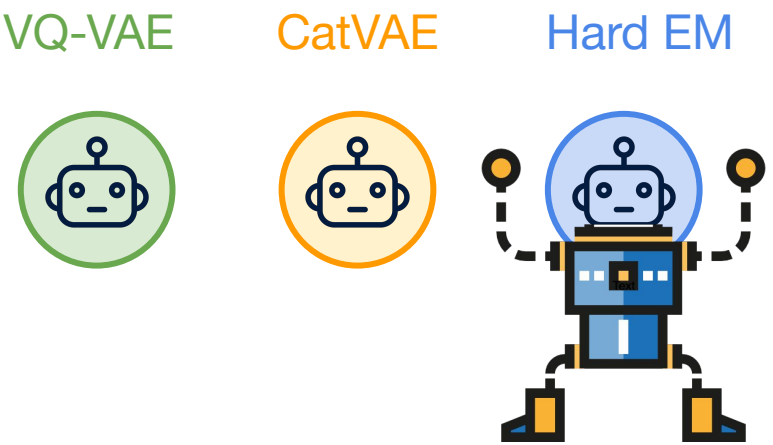
Sentence cluster

Takeaway

1. **Why you should use discrete ?**
   = matches continuous
   + compression + interpretability



2. **Which discrete to use?**
   Hard EM > CatVAE > VQ-VAE
   - classification
     low resource & re-embedding
   - document retrieval



Shuning Jin

Sam Wiseman

Karl Stratos

Karen Livescu