Developing an image-based document search system

Final Presentation

Authors: Anika Apel, Piotr Chodyko, Kyle Hiroyasu, Festina Ismali, Hyein Koo
Project Leader: Dr. Ricardo Acevedo Cabra
Scientific Lead: Dr. Robert Pesch, Julia Kronburger (M.Sc.), Dr. Andre Ebert
TUM Co-Mentor: PhD candidate Olga Graf

Team

Anika Apel

Piotr Chodyko

Kyle Hiroyasu

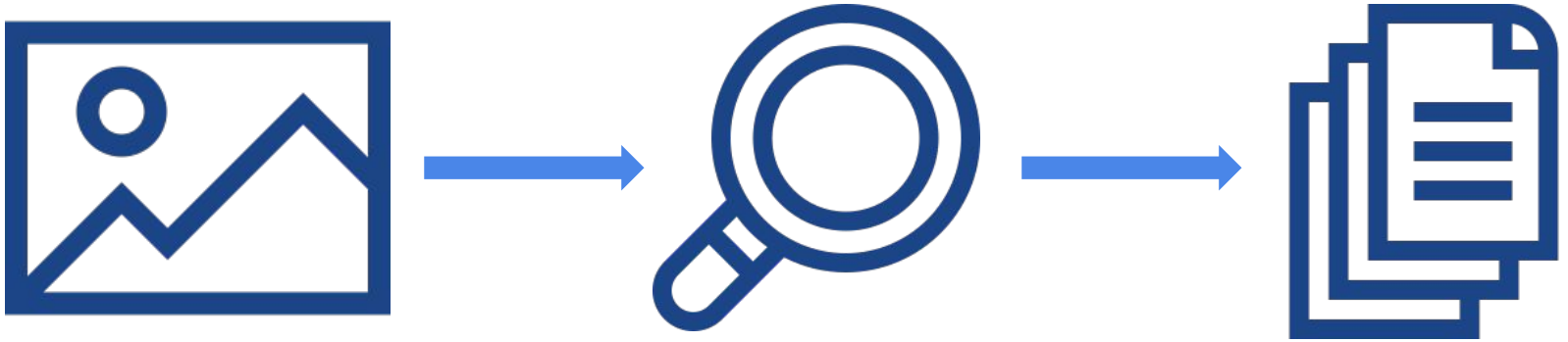
Festina Ismali

Hyein Koo

Contents

Introduction

Image-based Document Search



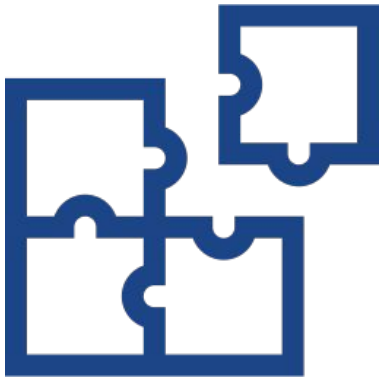
Query image

Search for related documents

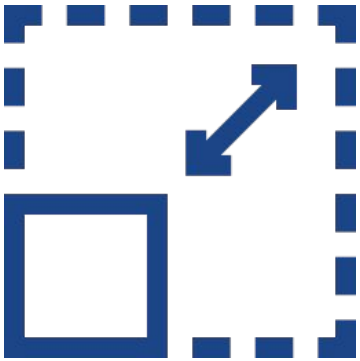
Return the most relevant documents

Project Goals

End-to-end

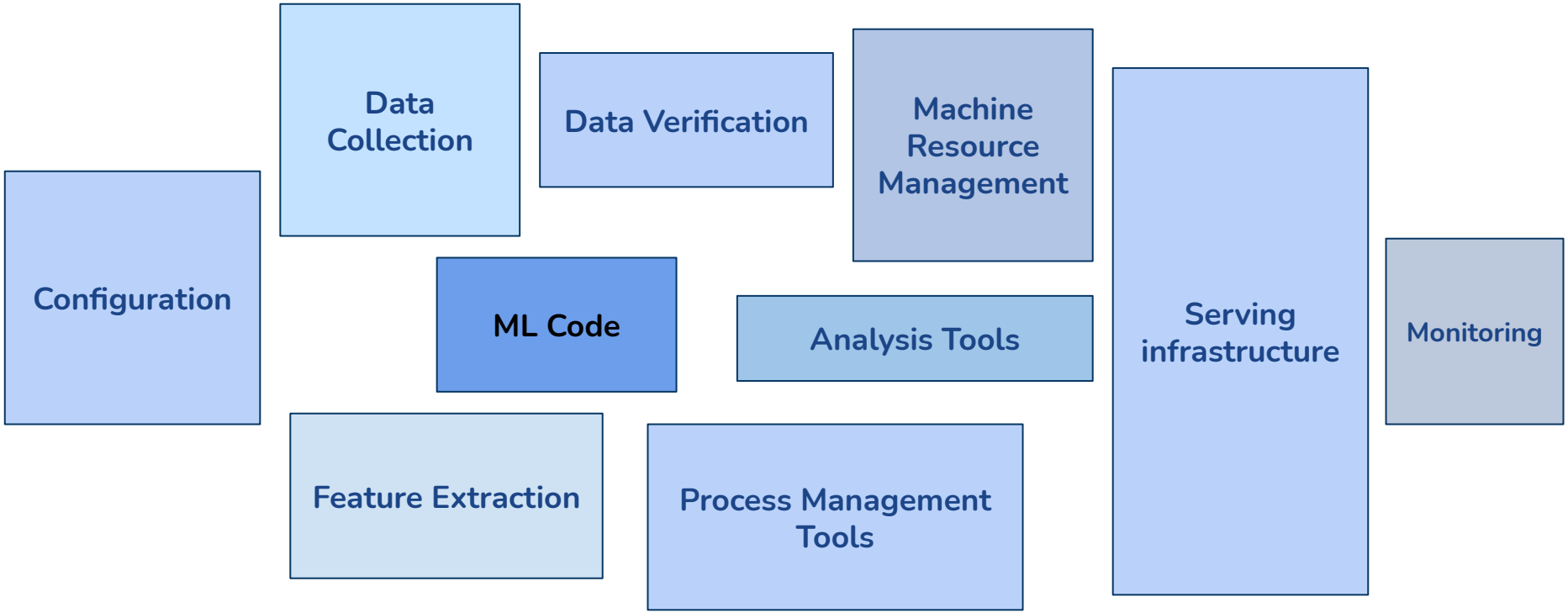


Modular



Scalable

End-to-End ML System Development



Inspired by [1]

Image-Text Retrieval in Production

Process Management Tools
- Version control
- Model management

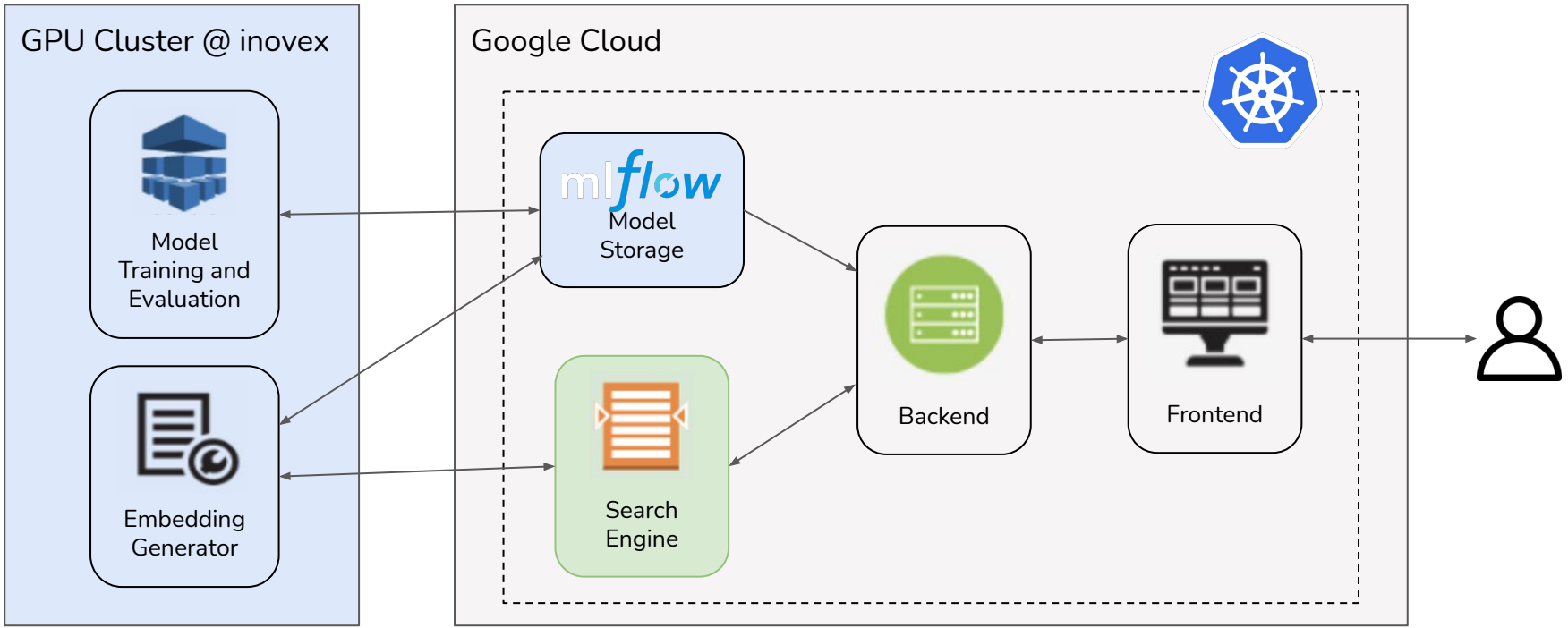
Feature extraction
- Backend

Serving Infrastructure
- Search system
- Web application

Machine resource management & Configuration tools
- Cloud infrastructure
- Iteratively add and modify deployments

End-to-End Application

Service Architecture



Service Architecture

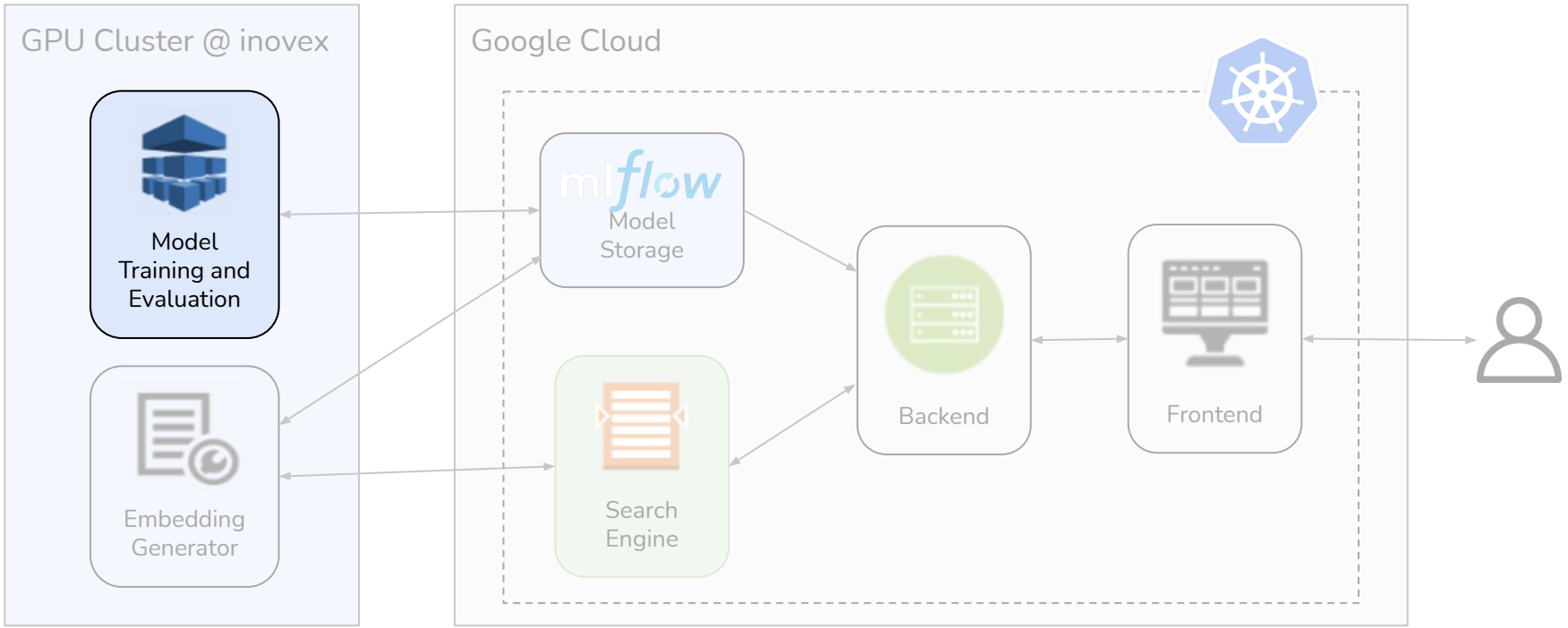


Image-to-Text Retrieval Models

Model	Code Available	Performance	Reproducible	Fast Inference
VSE++ [2]	✓	✓	✓	✓
VSRN [3]	✓	✓ ✓	○	○
OSCAR [4]	✓	✓ ✓ ✓	✗	✗

VSE++ [2]

A Boston terrier is running in the grass .

ResNet-152

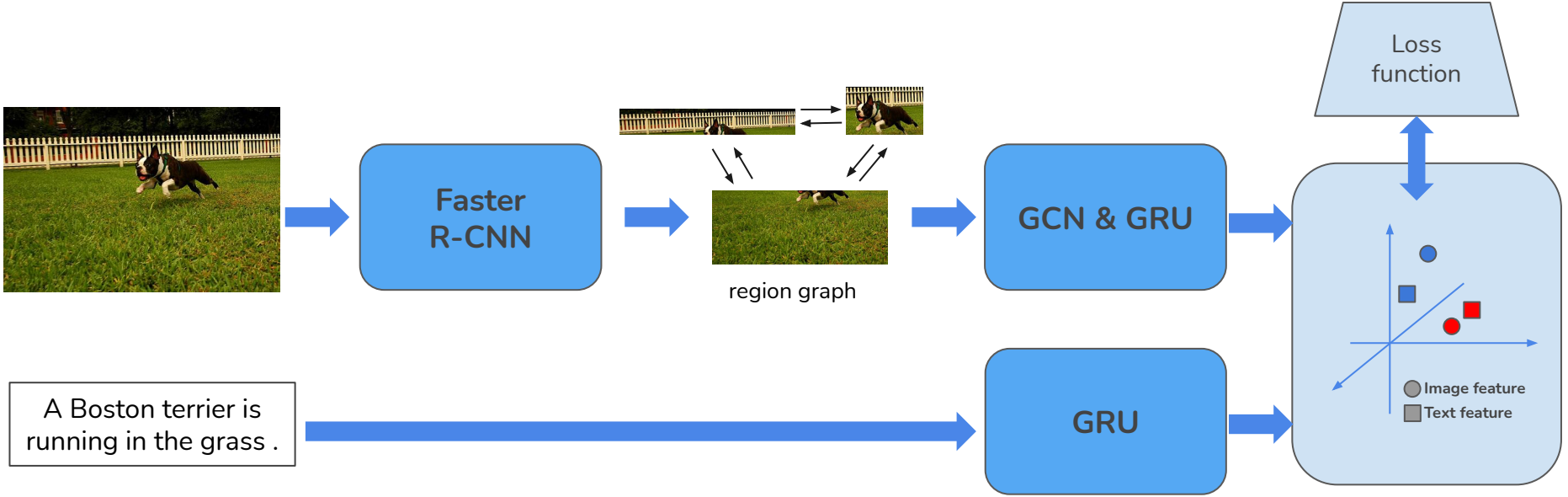
GRU

Loss function

Image feature

Text feature

VSRN [3]


Loss function
Faster R-CNN
region graph
GCN & GRU
A Boston terrier is running in the grass .
GRU
Image feature
Text feature

Model Evaluation

Dataset

Datasets	Size (images)	Train	Validation	Test
Flickr30k [5]	31,783	29,783	1,000	1,000
COCO [6]	123,287	113,287	5,000	5,000
Total	155,070	143,070	6,000	6,000

1	Boys kicking soccer ball in the grass under a tree.
2	Two boys are kicking a ball to each other in the park.
3	Two boys have made a goal out of two jackets in order to play soccer.
4	Two boys kick around a ball in a meadow.
5	Two kids play soccer in a field.

Recall@K

Percentage of queries for which at least one relevant document is among the top K results

1,

1	Two kids play soccer in a field.
2	A boy holds a red bucket up to a pony.
3	Two boys are kicking a ball to each other in the park.

0,

1	A dog is running in a field.
2	A boy holds a red bucket up to a pony.
3	A young girl on a swing.

Precision@K

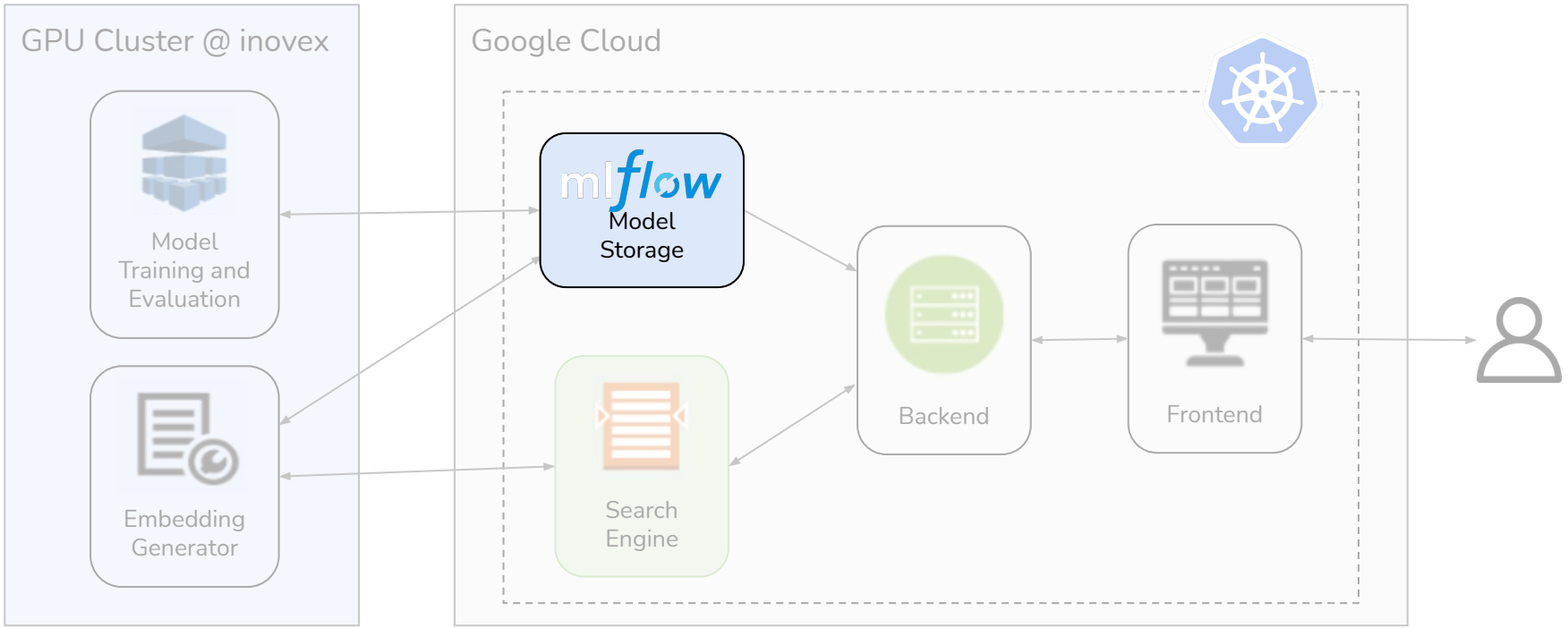
Average percentage of relevant documents among top K results

1	Boys kicking soccer ball in the grass under a tree.
2	Two boys are kicking a ball to each other in the park.
3	A young girl on a swing.

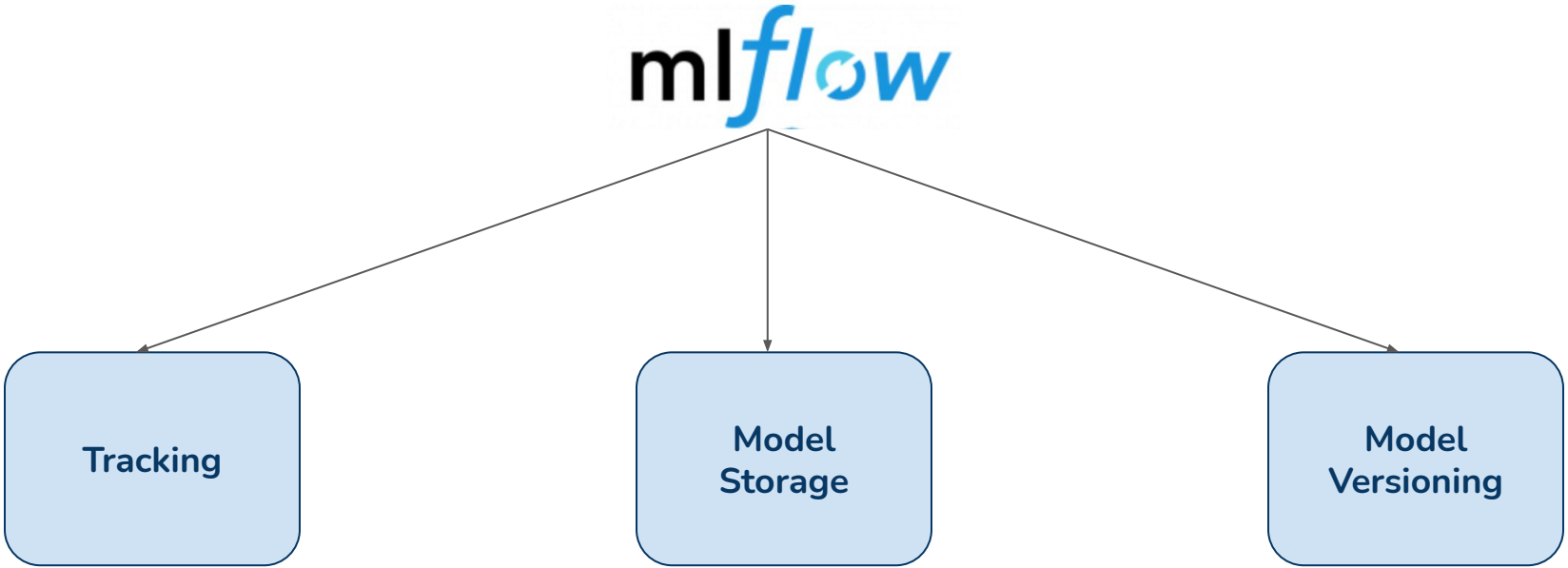
Model Results

Model	R@1	R@3	R@5	P@1	P@3	P@5
VSE++	0.343	0.542	0.634	0.343	0.388	0.244
VSRN w/ pre-computed features [7]	**0.461**	**0.667**	**0.759**	**0.461**	**0.392**	**0.339**
VSRN w/ Detectron2	0.315	0.516	0.607	0.315	0.275	0.241

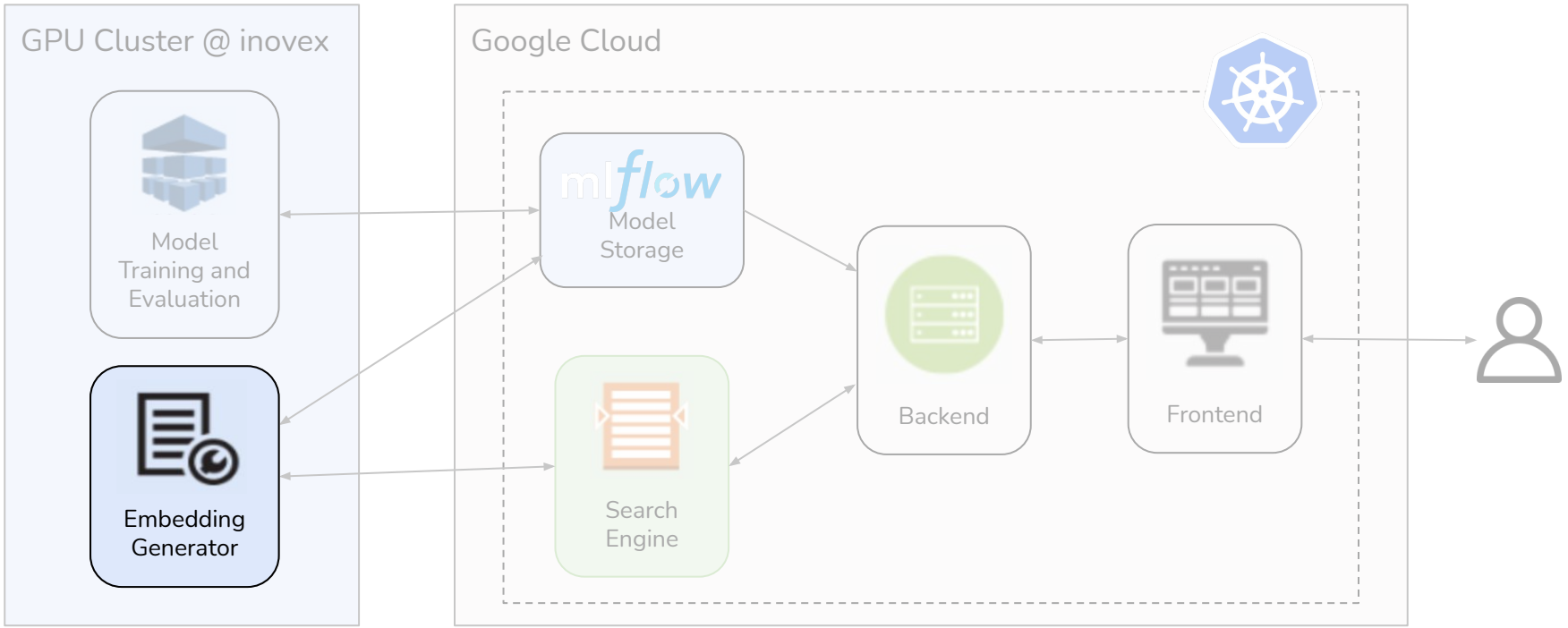
Service Architecture



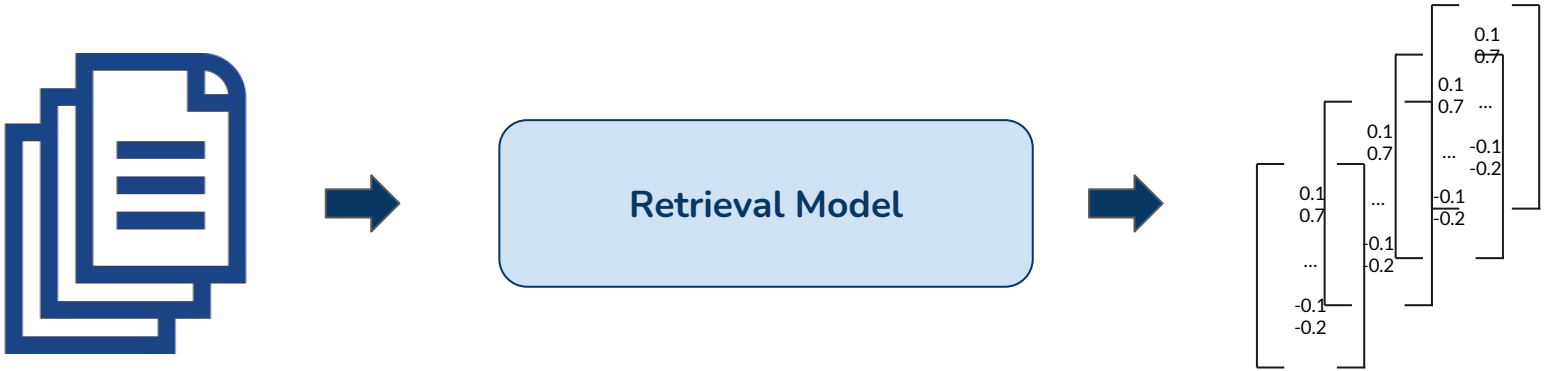
Model Management



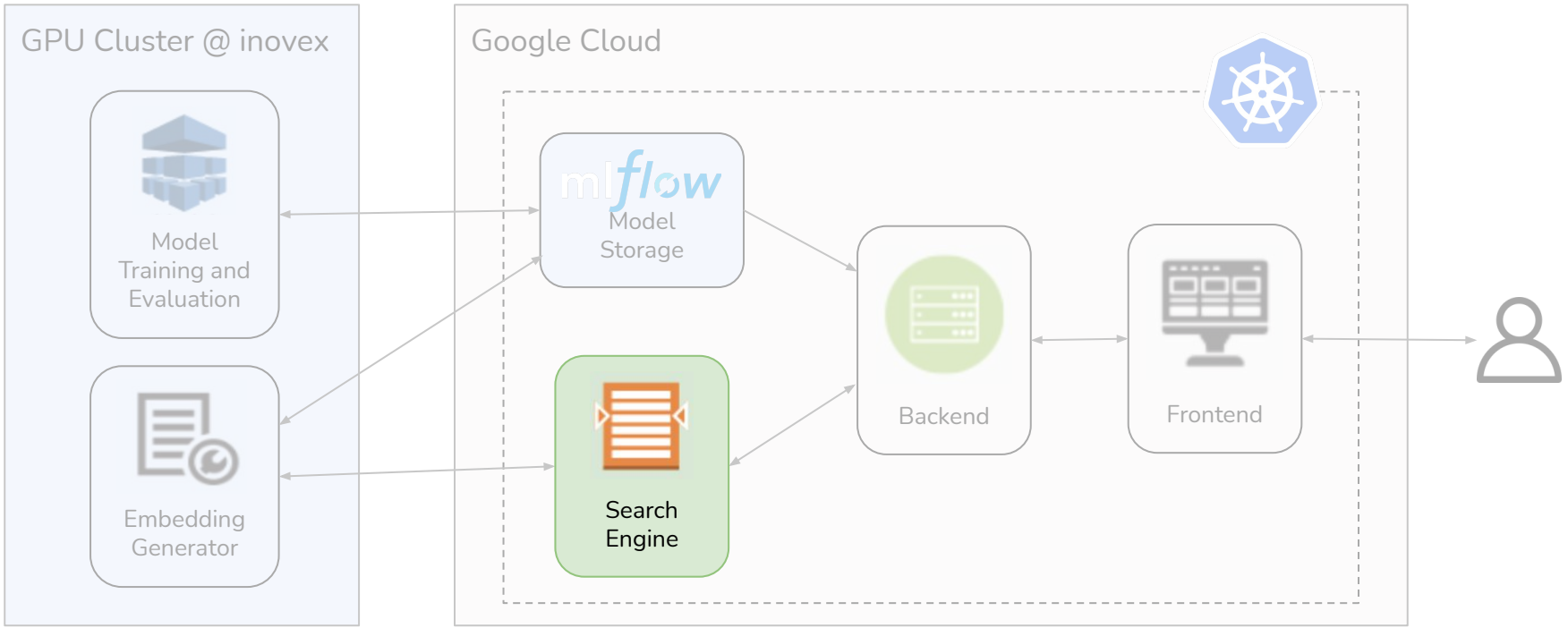
Service Architecture



Embedding Generator



Service Architecture



Search Engine

Image Encoder

0.8
0.2
...
0.6
-0.5

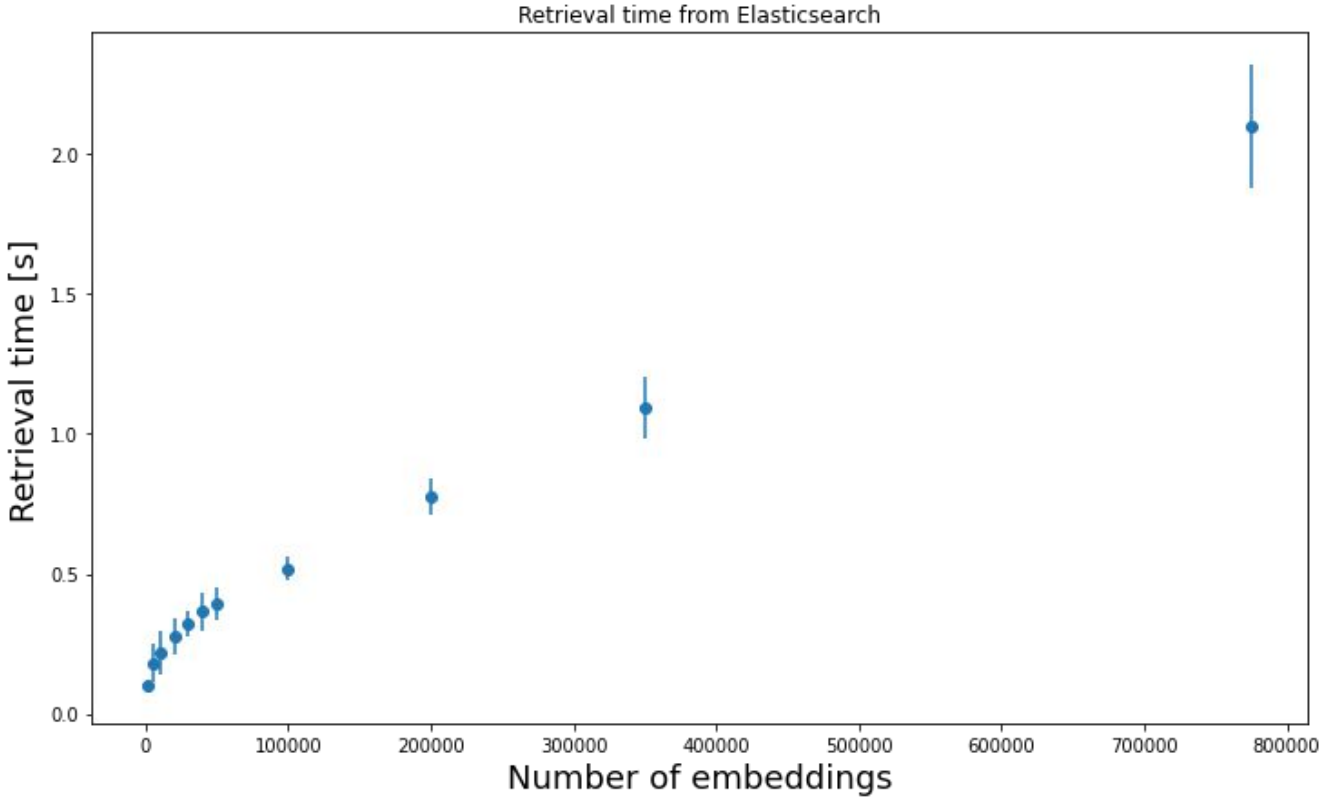
A Boston terrier is running in the grass .

Text Encoder

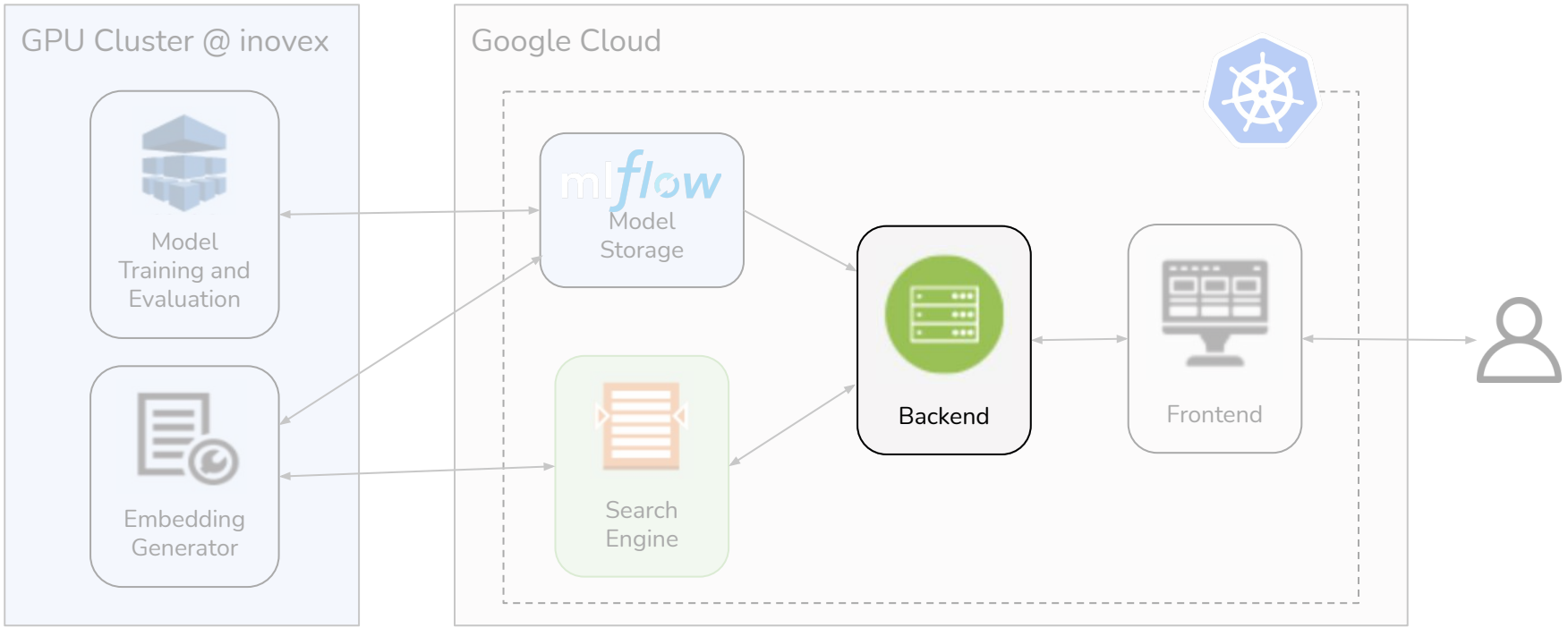
0.1
0.7
...
-0.1
-0.2

elastic

Elasticsearch Results



Service Architecture



Backend

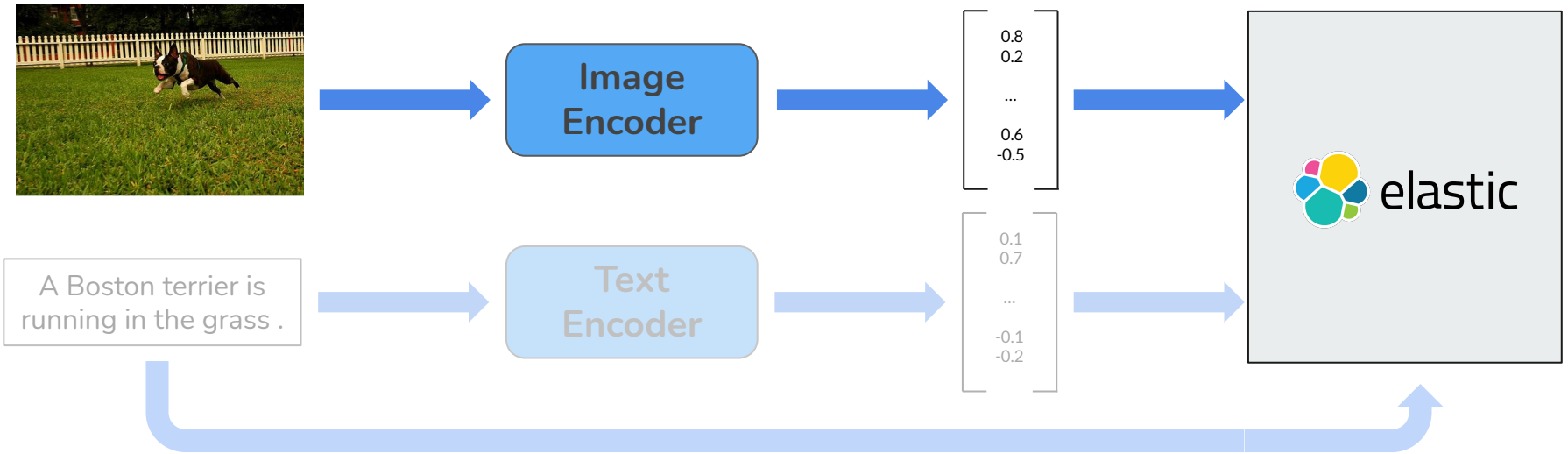


Image Encoder
0.8
0.2
...
0.6
-0.5
A Boston terrier is running in the grass .
Text Encoder
0.1
0.7
...
-0.1
-0.2
elastic

Backend

A Boston terrier is running in the grass .

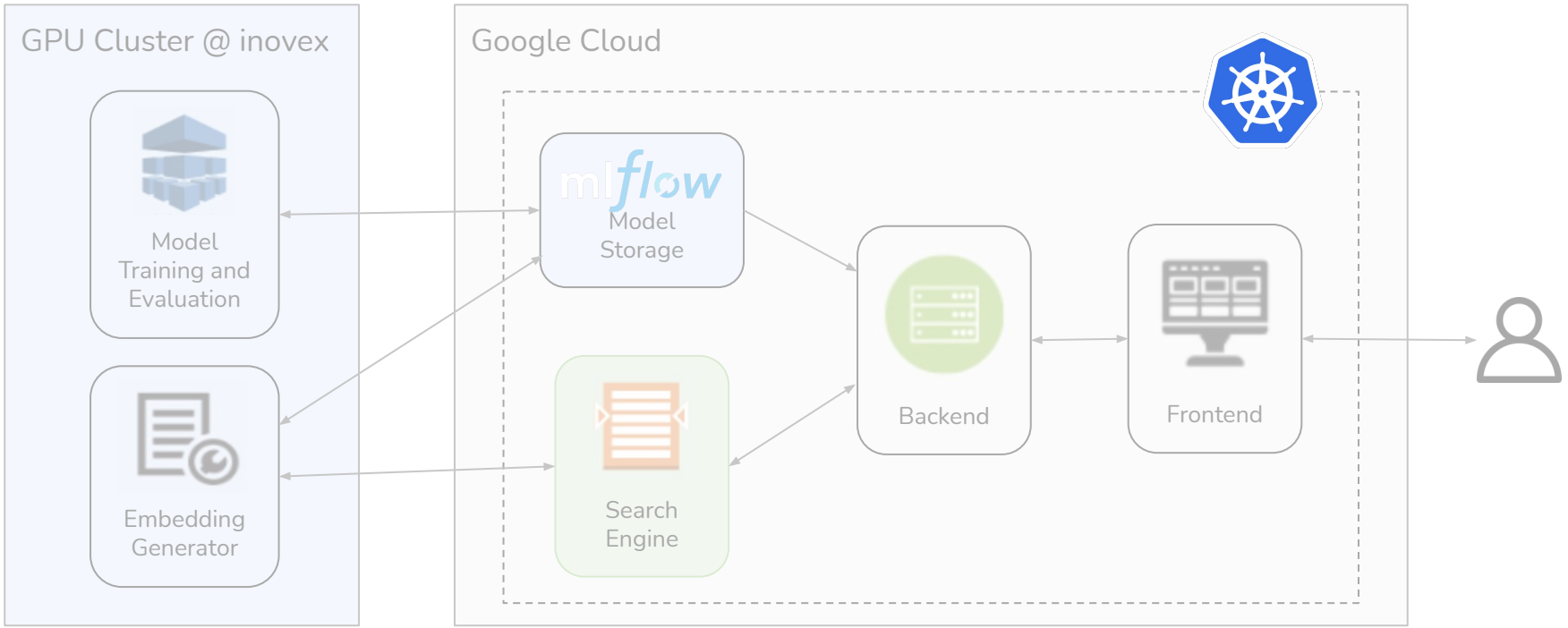
A Boston terrier is running in the grass .

Text Encoder

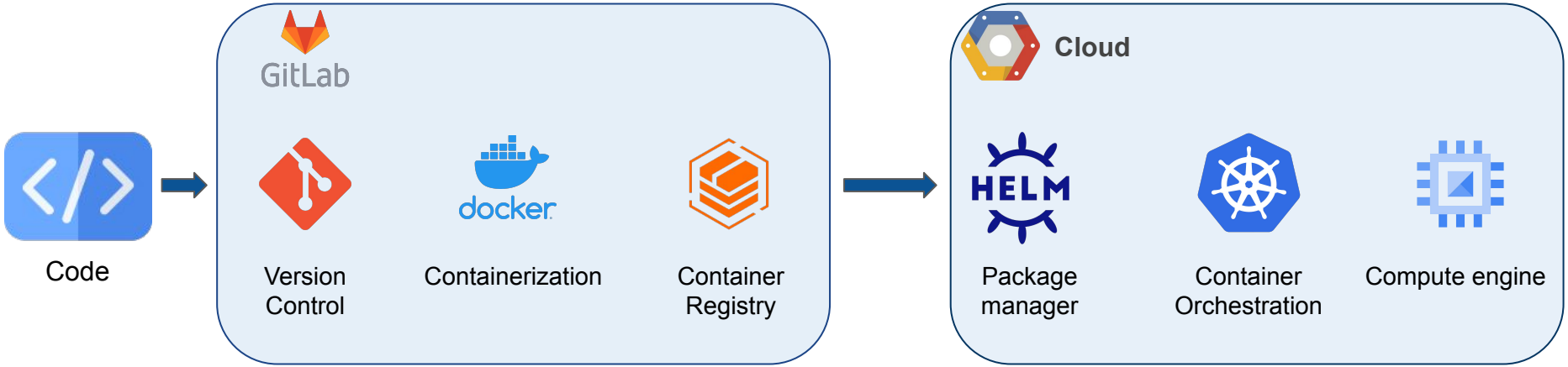
0.1
0.7
...
-0.1
-0.2

elastic

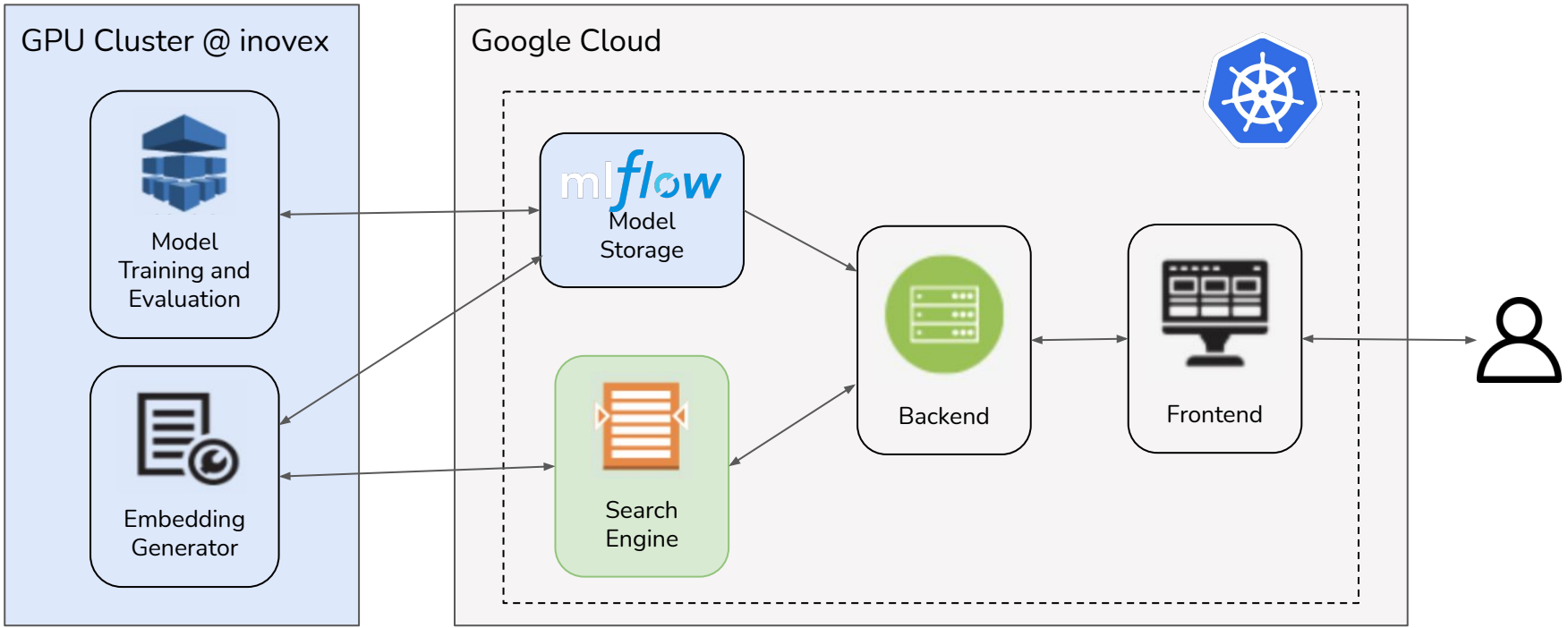
Service Architecture



Deployment


GitLab
Code
Version Control
docker
Containerization
Container Registry
Cloud
HELM
Package manager
Container Orchestration
Compute engine

Service Architecture



Demo

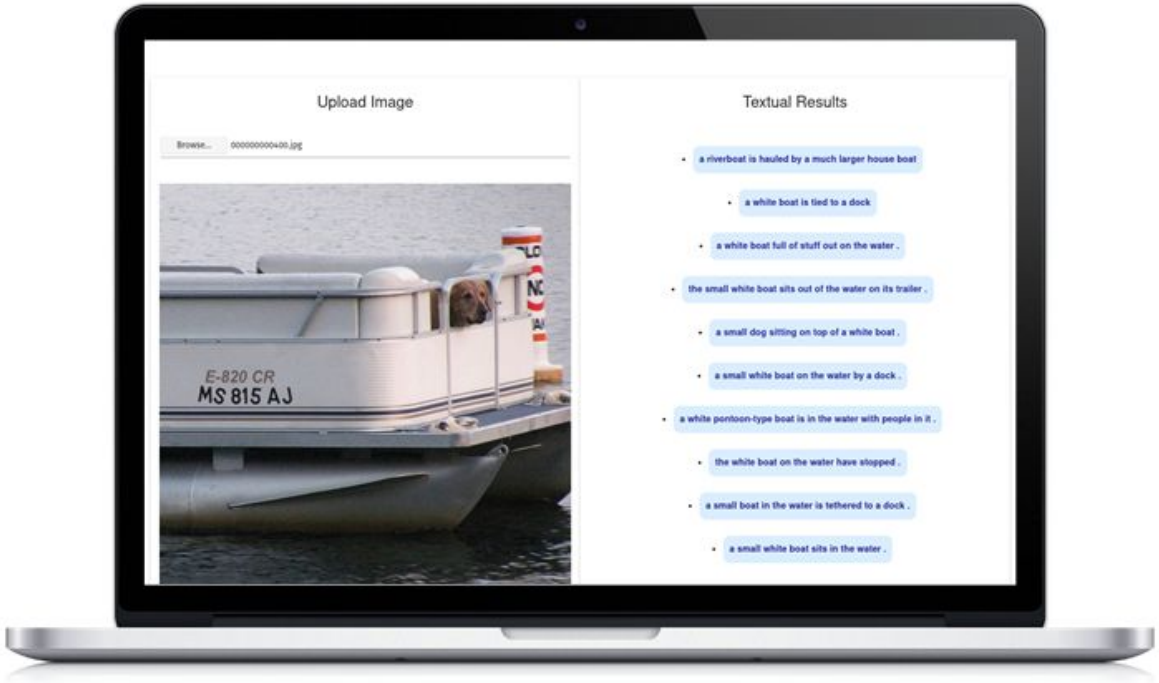
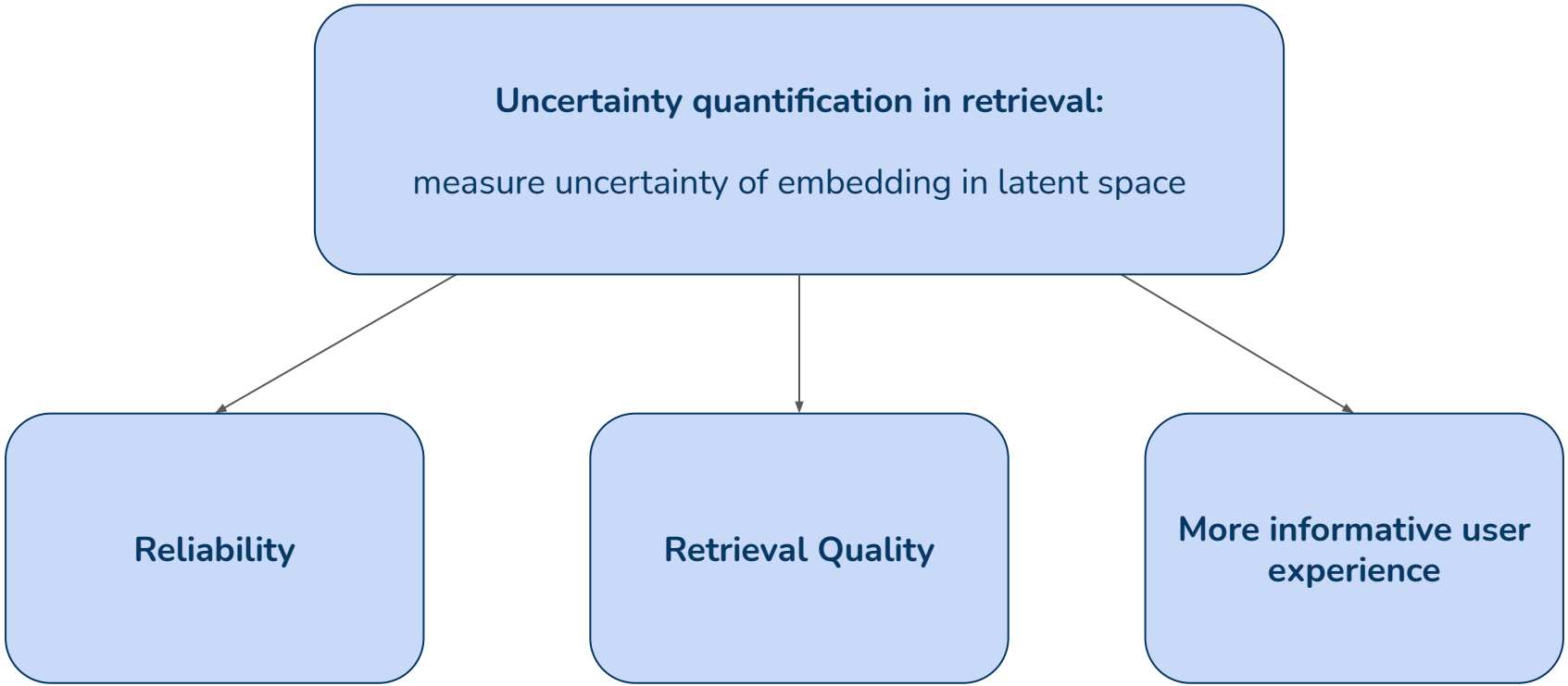


image-text-retrieval.inovex.de

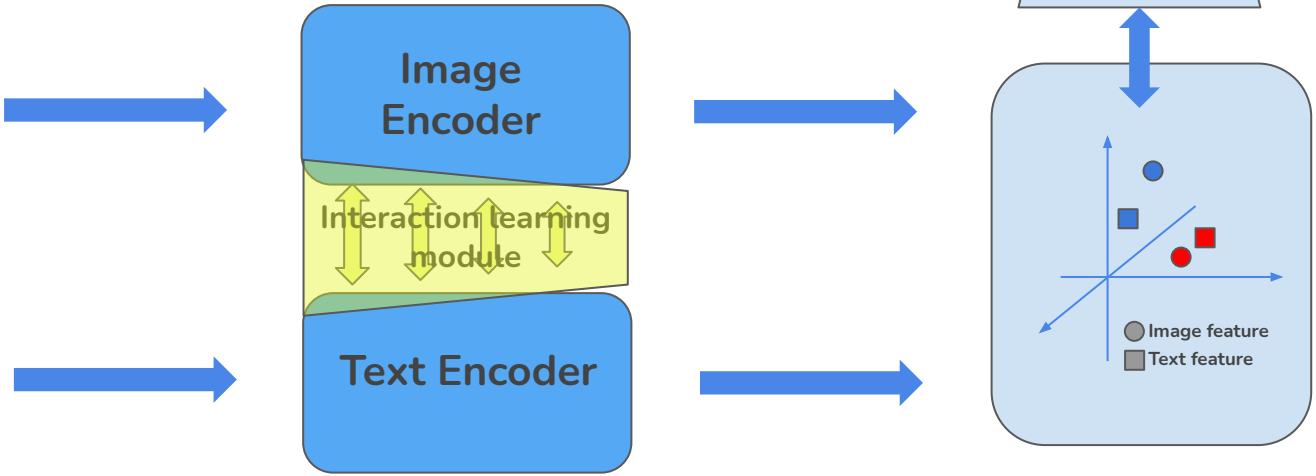
Discussion

Uncertainty Quantification



Interaction learning [11]

A Boston terrier is running in the grass .



Thanks for your attention!
Questions?

References

[1] Sculley, David, et al. "Hidden technical debt in machine learning systems." *Advances in neural information processing systems* 28 (2015): 2503-2511.

[2] Faghri, Fartash, et al. "Vse++: Improving visual-semantic embeddings with hard negatives." *arXiv preprint arXiv:1707.05612* (2017).

[3] Li, Kunpeng, et al. "Visual semantic reasoning for image-text matching." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

[4] Li, Xiujun, et al. "Oscar: Object-semantics aligned pre-training for vision-language tasks." *European Conference on Computer Vision*. Springer, Cham, 2020.

[5] Peter Young et al. “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions”. In: *Transactions of the Association for Computational Linguistics* 2 (2014), pp.67-78.

[6] Xinlei Chen et al. “Microsoft COCO Captions: Data Collection and Evaluation Server”. In: *CoRR* abs/1504.00325 (2015).

[7] Peter Anderson et al. “Bottom-up and top-down attention for image captioning and visual question answering”. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, pp. 6077–6086.

[8] Taha, Ahmed, Yi-Ting Chen, Xitong Yang, Teruhisa Misu, und Larry Davis. „Exploring Uncertainty in Conditional Multi-Modal Retrieval Systems“. *arXiv:1901.07702 [cs]*, (2019).

[9] Hama, Kenta, Takashi Matsubara, Kuniaki Uehara, und Jianfei Cai. „Exploring Uncertainty Measures for Image-Caption Embedding-and-Retrieval Task“. *arXiv:1904.08504 [cs, stat]*, (2019).

[10] Warburg, Frederik, Martin Jørgensen, Javier Civera, und Søren Hauberg. „Bayesian Triplet Loss: Uncertainty Quantification in Image Retrieval“. *arXiv:2011.12663 [cs]*, (2020).

[11] Jianan Chen et al. “Review of Recent Deep Learning Based Methods for Image-Text Retrieval”. In: *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE. 2020, pp. 167-172

Appendix

GCP Monthly Costs (January)

Service	SKU	Subtotal
Compute Engine	E2 Instance Core running in Frankfurt	€ 153.36
Compute Engine	E2 Instance Ram running in Frankfurt	€ 82.20
Compute Engine	Storage PD Capacity in Frankfurt	€ 26.47
Compute Engine	Network Load Balancing	€ 17.82
Compute Engine	Misc. Networking Services	€ 14.58
Cloud Storage	Download Worldwide Destinations	€ 18.88
Cloud Storage	Standard Storage Frankfurt	€ 0.71
Cloud Storage	Regional Standard Class B Operations	€ 0.09
	Total	€ 314.11

Uncertainty Quantification -Approaches for Retrieval

Monte-Carlo Dropout [8]

Deep Bayesian Neural Networks [9]

Stochastic embeddings [10]

Uncertainty Quantification - MC Dropout [7]

Key Idea: cast triplet loss as a regression loss and estimate epistemic uncertainty with **MC Dropout**

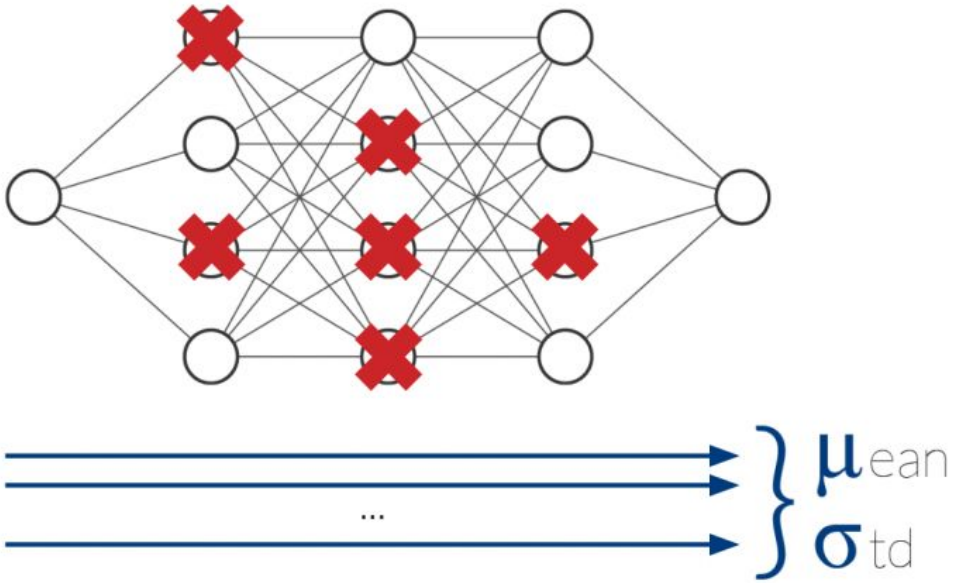


Figure taken from https://www.inovex.de/blog/uncertainty-quantification-deep-learning/

Uncertainty Quantification - BNN [8]

Key Idea: Consider embedding-and-retrieval task as regression/classification task and apply bayesian neural networks

- Use neural networks with stochastic components after each weight layer (enabled during inference and training)

- Stochastic components:
  - Stochastic batch normalization
  - Dropout

Uncertainty Quantification - Stochastic Embeddings [9]

Key Idea: Stochastic embeddings instead of deterministic ones

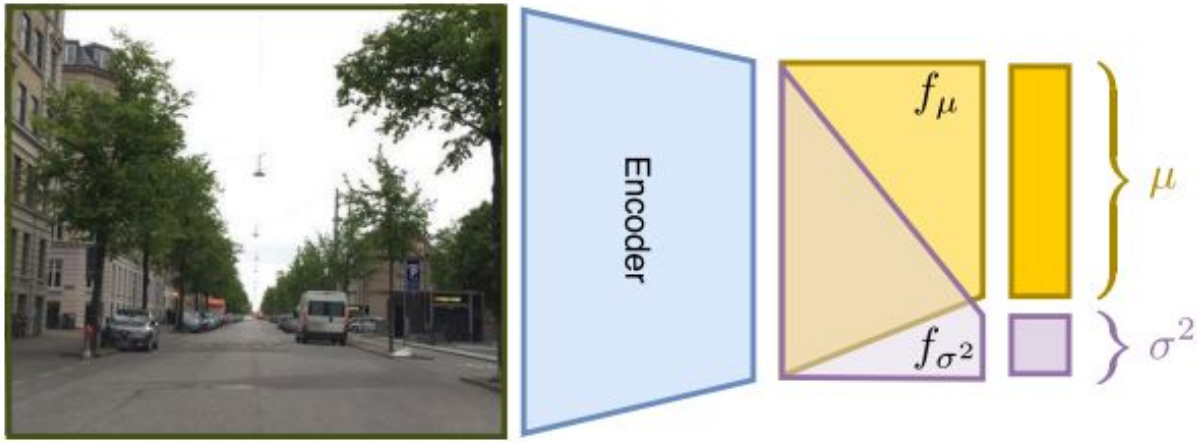


Figure: Architecture for Image Encoder with Stochastic Embeddings [9]

Pairwise learning [11]

A Boston terrier is running in the grass .

Image Encoder

Text Encoder

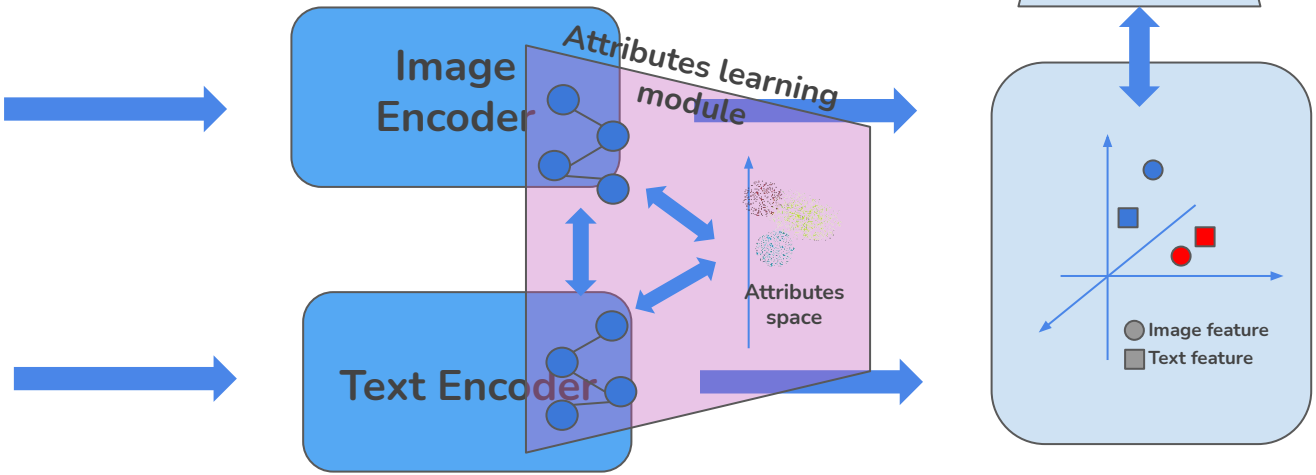
Loss function

Image feature

Text feature

Attributes learning [11]

A Boston terrier is running in the grass .



Interaction learning: OSCAR [4]

A Boston terrier is running in the grass .