Seminar for Master Students in
Software Engineering & Internet Computing

Master Thesis Proposal

Enhancing Image Retrieval with the Information Bottleneck Method

Student:
Marvin Seidl, 11777747

June 27, 2024

Advisor:
Univ.Prof. Mag.rer.soc.oec. Dr.rer.soc.oec. Schahram Dustdar

Co-Advisor:
Univ.Ass. Dipl.-Ing. Alireza Furutanpey, BSc

1 Problem Statement and Motivation

The task of *Content-Based Image Retrieval* is to find relevant images out of a database given the visual content of a *query* image. State-of-the-art methods use deep neural networks to extract image feature representations. The representations should be discriminative, compact, and robust. A common image retrieval setup to fulfill these requirements uses two types of feature representations: global and local features [3]. *Global features*, or *global descriptors* summarize the whole image into a single representation, an *embedding*. This representation of the visual content is compact, they capture high-level similarities well but lose the spatial information about individual elements in an image. *Local Features* are associated with specific regions of the image and describe *keypoints* for the depicted instance. They are especially suited for rigid objects [2].

Calculating the similarity between all local features of all images in the database is inefficient. *Re-ranking* is one approach to cut down on the search space: First, images are ordered by the similarity of their global descriptor, and then only the top-N results are ordered again by the similarity of their local descriptors. Modern approaches like *DELG* [2] extract both feature types simultaneously: *Global features* are extracted from the final layer of a convolutional neural network (CNN) and *local features* from the penultimate layer. During *re-ranking*, only the similarity based on local features is considered and the ordering by global similarity is discarded. We observe the following:

- By calculating the global similarity, we have gained information about the final order of the retrieved images. The re-ranking process of *DELG* [2] ignores this information.
- Both global and local features contain sufficient information to decide the final order, so local features should contain redundant information with the global feature. Since the capacity of features is limited, without this redundant component, the local features could better use their capacity to encode auxiliary information not already considered by the global feature.

Recent work [11] replaced the re-ranking process of *DELG* with one that also incorporates the global descriptor while still using the same CNN as a feature extractor, thus already alleviating the first issue (section 4.1 for details). The second issue remains, and we hypothesize that reducing the redundant information between global and local features increases retrieval performance.

2 Aim of the Thesis and Expected Results

We adopt the framework of the information bottleneck (IB) method to reason about the trade-off between preservation and compression in features. The use of the IB method in the context of this thesis is inspired by its application to multi-view representation learning, the goal of which is to map different data modalities into a single shared embedding space. The data can be heterogeneous, like image-text, video-text, or audio-text. Alternatively, it can be homogeneous, like two images depicting the same object

but captured from different sides. The IB method can be used to discard information that is not shared between two views [4]. Discarding non-shared information relies on the assumption that each view is sufficient for potential downstream tasks and that this task-relevant information is the same across each view. We consider our image retrieval system's global and local descriptors to be two views on the same underlying image data in the framework of Multi-View representation learning. The information bottleneck can be defined based on mutual information, a measure of statistical dependence between two random variables. Compared to correlation coefficients, mutual information can also capture non-linear dependencies at the cost of being harder to compute.

$$L = I(Z;Y) - \beta I(X;Z) \tag{1}$$

Equation 1 shows the maximization objective of the IB method. The first term encourages a representation Z to be predictive of the classification labels Y, i.e., we want the mutual information between representation and prediction to be high. The second term encourages the representation Z to "forget" about the input images X, i.e. we want the mutual information between representation and input to be low. The hyperparameter β controls the trade-off between the importance of both terms.

The main goal of this thesis is to develop a novel image retrieval system that improves upon previous work by explicitly modeling the information-theoretic relationship between global and local feature representations. By integrating the IB method into the training objective, we aim to improve the quality of the feature representations.

The thesis will answer the following research questions:

1. **Is there a relationship between mutual information and retrieval performance?**

   The feature extractor uses a proxy task (e.g., classification or pairwise distance losses) during training, making it hard to predict how well the learned representations will perform for retrieval.

2. **Which method of measuring and minimizing mutual information is the most suitable for our application?**

   Directly measuring mutual information in a very high-dimensional setting is not tractable. Several methods have been proposed to approximate or bound mutual information in recent years. However, there is no clear consensus yet on the practical limitations of these methods or which method performs best for our scenario (see section 4.2).

3. **Does applying the information bottleneck on global or local features alone improve retrieval performance?** So far, we outlined the mutual information minimization objective between global and local features. The IB method itself is primarily applied as a regularization technique. Accordingly, the thesis will also explore the effect of the IB method on retrieval performance when applying it to local and global features independently.

3 Methodology

The goals of the thesis will be achieved through the following steps:

- **Literature review** Suitable state-of-the-art methods and evaluation schemas for image retrieval have to be identified in the literature. This step was partially completed by selecting specific methods to improve upon as outlined in section 1. However, a crucial aspect of the implementation will be the measurement and minimization of mutual information between different components of the retrieval model. A particular focus of the literature search is the different methods to measure mutual information and the limitations of these methods.
- **Model Implementation and Verification** The methodology involves developing a novel image retrieval system incorporating the information bottleneck principle. We base our work on existing architectures of *DELG* [2] and *Reranking Transformer* [11] (for details, see section 4.1). The loss function of the existing method will be adapted to incorporate terms related to the mutual information between input images, feature representations, and class labels.
- **Model evaluation** To establish a baseline for comparison, the proposed image retrieval system will be compared against the state-of-the-art methods described in section 4.1. Comparison will use the datasets and evaluation protocol of these methods. Datasets used by these methods are: *GLDv2* [12], *Revisited Oxford and Paris* [8] and *Stanford Online Products (SOP)* [10]. Mean average precision (mAP) and recall @ rank (r@k) are commonly used metrics. The effect on retrieval performance of different methods for measuring mutual information will be compared to each other. If applicable, the individual contribution of model components will be assessed separately. Relating the mutual information between representations to the retrieval performance could be done by plotting the information plane [9] of representations. It shows the mutual information between model input and representation on one axis and between representation and output on the other.
- **Ablation Studies** Our experiments involve numerous hyperparameters, such as training time, loss weight scales, and stochasticity during train time. Accordingly, we will perform ablation experiments several times using the same configurations to determine whether performance gains are from our proposed method.

4 State of the Art

4.1 Image Retrieval

As introduced in section 1, *DELG* [2] extracts both global and local features simultaneously from a CNN backbone. An attention module selects which local features should be used. Similarity of global features is calculated based on their dot product. The re-ranking process based on local features uses *RANSAC* [5] to perform geometric verification: First, local features are paired by L2 distance and their spatial location in the

feature map is recorded. Second, an affine transformation is fitted on the spatial locations of all feature pairs. Third, based on a threshold value, the fitted *RANSAC* model produces several inlier pairs. This number is used as the metric for the re-ranking. Geometric verification has several downsides: It requires rigid objects, e.g., a human changing poses between images, would change the relative location of detected keypoints, and is sensitive to large viewpoint changes. It is also a computationally expensive iterative process that requires many local features.

Reranking Transformer [11] directly predicts image pair similarity based on local and global features. In the context of this thesis, this offers an important benefit over geometric verification: The re-ranking process of *DELG* ignores the relative similarity based on global features in the top-N results and only considers the output of geometric verification. It is unclear how our proposed modification on the information content of local features would influence the final retrieval performance and how the weighting of global and local feature similarity would need to be changed to accommodate. By having a model learn the final joint similarity score as in *Reranking Transformer*, this problem can be delegated to the model itself.

4.2 Mutual Information Estimation

A major challenge for applying the IB method to modern machine learning is the calculation of mutual information in high dimensional space which is intractable in general. Variational approximation and the optimization of a lower bound on the IB objective are commonly used instead [7]. These methods require that the distribution of one of the random variables is known. Since the true underlying distribution of input images is unknown, this restriction is usually placed upon the learned representation.

In recent years, a number of general purpose estimators of mutual information have been introduced: e.g, MINE [1] or NWJ [6]. These methods require an auxiliary neural network that optimizes a lower bound by estimating expectations over sample distributions. Another promising direction is the usage of kernel-based methods like Rényi's α-order entropy which has already been applied to the multi-view bottleneck [13].

5 Context within the Master Program

Machine Learning, introductory course with a broad range of general topics about machine learning. *Artificial Intelligence Seminar*, the seminar topic was in the field of Music Recommender systems. *Security, Privacy, and Explainability in Machine Learning*, includes topics like model interpretability and transparency, statistical notions of bias, and fairness. *Advanced Information Retrieval*, includes topics like search engines, evaluation of search results, and machine learning architectures to create embeddings for words and passages of text. *Medical Image Processing* (Data Science/Medical Informatics/Visual Computing Curriculum), contains general computer vision topics like image segmentation, pixel classification and image registration. Courses from the *Data Science* curriculum that are adjacently relevant to the thesis: *Knowledge Graphs, Data-oriented Programming Paradigms, Recommender Systems, Data-intensive Computing.*

References

[1] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, R. Devon Hjelm, and Aaron C. Courville. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, 2018.

[2] Bingyi Cao, André Araujo, and Jack Sim. Unifying Deep Local and Global Features for Image Search. In *Computer Vision – ECCV 2020*, pages 726–743, Cham, 2020.

[3] Wei Chen, Yu Liu, Weiping Wang, Erwin M. Bakker, Theodoros Georgiou, Paul W. Fieguth, Li Liu, and Michael S. Lew. Deep learning for instance retrieval: A survey. *EEE Trans. Pattern Anal. Mach. Intell.*, 45(6):7270–7292, 2023.

[4] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *8th International Conference on Learning Representations, ICLR 2020*, 2020.

[5] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.

[6] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. F-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 271–279, 2016.

[7] Ben Poole, Sherjil Ozair, Aäron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, pages 5171–5180, 2019.

[8] Filip Radenovic, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, pages 5706–5715, 2018.

[9] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *CoRR*, abs/1703.00810, 2017.

[10] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pages 4004–4012, 2016.

[11] Fuwen Tan, Jiangbo Yuan, and Vicente Ordonez. Instance-level image retrieval using reranking transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021*, pages 12085–12095, 2021.

[12] Tobias Weyand, André Araújo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2 - A large-scale benchmark for instance-level recognition and retrieval. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, pages 2572–2581, 2020.

[13] Qi Zhang, Shujian Yu, Jingmin Xin, and Badong Chen. Multi-view information bottleneck without variational approximation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022*, pages 4318–4322, 2022.