UniOvi Team at LeQua 2024: Quantification via Gaussian Latent Space Representations

Olaya Pérez-Mon[0000−0002−4527−6698] and Pablo González[0000−0002−9250−0920]

Artificial Intelligence Center, University of Oviedo, Gijón, Asturias, Spain
UO257319@uniovi.es
gonzalezgpablo@uniovi.es

Abstract. Most traditional quantification methods often depend on prior probability shift assumptions to develop models that use classifiers for optimal prevalence estimation. In contrast, the idea of this team was to introduce an end-to-end neural network that utilizes Gaussian distributions to achieve invariant sample representations. This paper describes this deep learning architecture and its application to the LeQua 2024 quantification competition, specifically, addressing tasks T2, T3, and T4.

1 Motivation

Our main goal in this competition was to evaluate the performance of a novel deep-learning representation layer for quantification. This layer is based on modelling latent spaces using Gaussian distributions, to obtain an invariant representation of bags of examples. Using this layer in a deep neural network, with an architecture suited for quantification [1], we can undertake prevalence estimation tasks and compare the performance of our method against some of the most popular quantification algorithms, which served as baselines in the competition, as well as against the other participants.

We focused on tackling tasks T2 and T3, both complex multiclass problems, as well as T4, which, despite being a binary problem, presents the added challenge of covariate shift. We opted not to participate in task T1, a binary quantification problem where traditional quantification methods are highly effective and much simpler (using our approach for this task might have been seen as overkill). Our solution delivered consistently strong results in the multiclass tasks (T2 and T3), achieving a gold medal in both tasks. In the binary task with covariate shift, T4, we achieved fifth place. The setup and methodology employed to tackle the competition tasks will be detailed in Section 2.

2 Method

For this competition, we used a deep neural network designed to address quantification problems. Our approach builds on the same architecture as Deep

Quantification Network (DQN) [1], but instead of using pooling layers for representing the bags, we use Gaussian distributions. The architecture consists of three main components: i) a feature extraction module, which is adapted to the specific task, it projects features into latent spaces. ii) a bag representation module that summarizes the feature vectors into a numerical representation of each bag. In our case, Gaussian distributions are used to model the latent space. And, iii) a quantification module, a combination of dense linear layers, that learns the relationship between bag representations and bag prevalences, outputting prevalence estimates due to a softmax activation function.

This architecture is flexible, allowing the optimization of different loss functions, such as the Relative Absolute Error (RAE) for tasks T2 and T4, and the Normalized Match Distance (NMD) for T3. In addition, it can be trained using bags labelled by prevalence, with or without individual example labels.

2.1 Layer details

Our method introduces a novel approach to generating invariant bag features using learnable multivariate Gaussian distributions. This representation layer models the latent space, to which the bag examples are projected, by covering it with Gaussian distributions, each defined by a mean vector and a covariance matrix, which are learnable parameters of the network. Each example projected in the latent space is evaluated against the Gaussian distributions by calculating their likelihoods. The mean likelihood across all examples in a bag then shows how well each Gaussian represents the entire bag, resulting in a compact and invariant representation in the latent space.

In order to capture more useful information about the bags, we propose an extension of the network incorporating multiple latent space representations, each modelled with its own Gaussian distributions. To favor learning different latent spaces with different information, we use the Centered Kernel Alignment (CKA) score (for more information, please, see [2]). Since this metric measures the similarity between latent spaces, we incorporated it into the loss function we are minimizing to make them as distinct and informative as possible.

Another important aspect of our approach is that the network can be trained using only bags labelled by prevalence, but can also benefit when utilizing individual labelled examples, adding a classification layer to the feature extraction module that allows the network to learn representations, combining quantification and classification losses during training. It is also possible to train the net by combining these two techniques.

When labelled bags are limited, we use data augmentation techniques to generate new bags. There are two options: i) when only individual labels are available, we generate new bags using the APP protocol [3] that allows us to generate bags with a desired prevalence and, ii) when only bag prevalences are available, we can mix two real bags producing a new augmented bag whose prevalence is the average of the original bags. When possible, we can combine both strategies.

3 Experiments

The datasets used in this study correspond to the T2, T3 and T4 tasks from the LeQua 2024 competition. T2 is a multiclass problem with prior probability shift, T3 is a multiclass ordinal quantification problem, and T4 is a binary problem with covariate shift.

The official loss function for T2 and T4 is the Relative Absolute Error (RAE) and Normalized Mean Distance (NMD) for task T3 (see [4] for more information on quantification loss functions).

The experimental setup was tailored to optimize the network according to the official loss function for each task. Our quantification method's flexibility allows for various training data configurations, as explained in Section 2. Specifically, for T2 and T4, half of the training bags were generated using APP from example-labelled data, while the other half were labelled by prevalence (with 700 out of 1,000 bags used for training and 300 for validation). To prevent overfitting, the labelled data was augmented by randomly mixing real bags, as described in Section 2.

For T3, the network was trained exclusively on bags generated using APP from the example-labeled dataset, with the 1,000 training bags used for validation and early stopping. In this case, the label information was incorporated into the network helping its convergence.

The number of Gaussian distributions to cover the latent space was fixed to 100. For T2, the most challenging task, we used 18 different latent spaces, each in 5 dimensions. The number of latent spaces for T3 was set to 9, while the other parameters remained unchanged. The limitation in the number of latent spaces was given due to memory limitations in the GPU used for training (12Gb). For T4, we utilized 1 latent space with 17 dimensions and 156 Gaussians. For T4, these parameter values were found using Optuna [5]. For T2 and T3 parameters were chosen manually, without using any automated procedure.

3.1 Results

In this section, we present the performance of our proposed method on the three tasks: T2, T3, and T4.

For the T2 dataset, see Table 1, our method significantly outperforms all baseline methods as well as other participants' ones, in terms of the optimized loss function, Relative Absolute Error (RAE). The performance gap is significant, demonstrating the effectiveness of our approach in handling this complex multiclass problem with prior probability shift.

The T3 dataset presents a multiclass ordinal problem with five classes, where the objective loss function to optimize is the NMD, see Table 1. In this case, our method also demonstrates superior performance, although the margin of improvement is narrower compared to T2. Nevertheless, our approach proves to be more effective in dealing with the complexities of ordinal quantification.

Finally, for the T4 dataset, see Table 1, our method was outperformed by other teams' approaches. In this case, our thought was that the ability to train

Ranking	T2 (RAE)	T3 (NMD)	T4 (RAE)
1	**0.9217 (Ours)**	**0.0644 (Ours)**	**0.1093 (tobiaslotz)**
2	1.0302 (tobiaslotz)	0.0659 (tobiaslotz)	0.1150 (EMQ)
3	1.0786 (hustav)	0.0668 (PCC)	0.1156 (DistMatching-y)
4	1.1616 (EMQ)	0.0690 (KDEy)	0.1180 (KDEy)
5	1.1942 (PACC)	0.0721 (juanjodelcoz)	0.1298 (Ours)

Table 1: Top five competitors for each of the tasks in which we competed.

the network using bags with a particular type of shift might prove an advantage when this shift is reproduced in the test bags. In practice, quantification methods using an underlying classifier and designed for prior probability shift, performed substantially well in this task, with errors very similar to those obtained by teams in task T1.

4 Conclusions

We have reached several interesting conclusions. Firstly, our method has demonstrated strong adaptability across a variety of quantification scenarios, achieving robust results in the presence of prior probability shift, covariate shift and, even in ordinal quantification problems. It is particularly effective in handling multiclass problems, which are often more complex and demanding.

Additionally, our method can optimize different loss functions, such as RAE and NMD, delivering strong results in both cases. Furthermore, our approach is flexible regarding training data. It can handle situations where data is labelled by individual examples, where bags are labelled by prevalence, or when there is a combination of both cases.

References

1. Lei Qi, Mohammed Khaleel, Wallapak Tavanapong, Adisak Sukul, and David Peterson. A framework for deep quantification learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part I*, pages 232–248. Springer, 2021.
2. Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited, 2019.
3. Andrea Esuli, Alessandro Fabris, Alejandro Moreo, and Fabrizio Sebastiani. *Learning to Quantify*. Springer, Cham, CH, 2023.
4. Fabrizio Sebastiani. Evaluation measures for quantification: An axiomatic approach. *Information Retrieval Journal*, 23(3):255–288, 2020.
5. Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.