Training Multimodal Systems for Classification with Multiple Objectives

Jason Armitage[1], Shramana Thakur[1], Rishi Tripathi[1], Jens Lehmann[1,2], and Maria Maleshkova[1]

[1] University of Bonn, Germany
[2] Fraunhofer IAIS, Dresden, Germany

Abstract. We learn about the world from a diverse range of sensory information. Automated systems lack this ability as investigation has centred on processing information presented in a single form. Adapting architectures to learn from multiple modalities creates the potential to learn rich representations of the world - but current multimodal systems only deliver marginal improvements on unimodal approaches. Neural networks learn sampling noise during training with the result that performance on unseen data is degraded. This research introduces a second objective over the multimodal fusion process learned with variational inference. Regularisation methods are implemented in the inner training loop to control variance and the modular structure stabilises performance as additional neurons are added to layers. This framework is evaluated on a multilabel classification task with textual and visual inputs to demonstrate the potential for multiple objectives and probabilistic methods to lower variance and improve generalisation.

Keywords: Machine Learning · Multimodal Data · Probabilistic Methods.

1 Introduction

Human experience of the world is rooted in our ability to process and integrate information present in diverse perceptual modalities [1]. Multimodal approaches to machine learning are motivated by this ability and aim to develop rich representations that combine information from multiple sources [2]. Consider a machine learning system that models events in the world by processing inputs from online media sources. Representations of these events take the form of text, images, video, and audio. Systems that are able to process signals from a range of these inputs learn models that are more complete descriptions of the represented events with resulting benefits for inference and predictions [3]. Researchers have proposed related classifiers for performing event detection [4], source prediction [5], and activity recognition [6].

Multimodal machine learning presents a suite of methods for leveraging diverse data - but the development of systems that generalise to unseen samples leads to challenges arising both in practice and from the underlying theory of machine learning. Limited data resources are the most pressing concern in the first category. Data acquisition for multimodal systems is complicated by the requirement for combinations of samples in each input modality [7]. In the absence of large-scale data, neural networks learn sampling noise in the training data and report low scores on unseen samples [8]. Additional modalities also inflate parameter counts with the outcome that multimodal systems report high accuracy during training and low accuracy at test time [9]. Additional hidden layers can boost performance during training but introduce the requirement to prevent the interaction of parameters across the model from slowing convergence [10].

Multimodal fusion combines representations from constituent modalities into a single embedding. In recent years, the deployment of neural networks to generate fused embeddings has resulted in state-of-the-art performance on classification tasks of textual and visual samples. In theory, multimodal fusion methods capture information present in the input representations and produce outputs with complimentary information. Comparison with unimodal classifiers demonstrates that the introduction of additional modalities yields only modest performance gains [9]. In addition to overfitting, limitations on available data are acute when tasks require images or video.

Main contributions. We propose and build a novel approach to multimodal classification that introduces a second objective to learn fused embeddings trained with variational inference. To our knowledge, this use of multiple objectives - where one function is learned with a method from inverse probability - is unique in the research on multimodal representation learning. We go on to show that the range of methods for calibrating parameter updates developed within the latter approach offsets the overfitting associated with multimodal fusion. The benefits of these proposals are demonstrated empirically by adapting an existing end-to-end architecture to perform multilabel classification on a dataset of 25k samples of paired images and text. F-scores on classifying unseen samples provide measurement of the contribution from introducing a second probabilistic objective and related regularisation methods to multimodal classification tasks.

Structure of the analysis. We start with an outline of the use case identified for multimodal representation learning followed by a detailed specification of our proposed framework and related methods. The evaluation section presents topline results for the system and an ablation on regularisation methods in variational inference. Section 4 highlights the existing research informing our work and we conclude by summarising the main findings.

2 Use Case

Learning on multiple modalities presents opportunities to generate rich representations for enhancing performance on existing tasks and enables new appli-

cations [3]. This section introduces a form of classification task where samples are presented both in natural language and images. Systems that learn on these modalities are applied to a range of use cases related to archived and online media [5,4,3].

2.1 Multilabel Classification on Multiple Modalities

Label prediction underlies approaches to information retrieval [11] and classification [12] - and enables the downstream tasks of document retrieval [13], textual and visual entailment [14,15], and fact validation [16]. Assigning samples to a potential subset of multiple labels also characterises challenges in unimodal [17] and multimodal [18] real-world applications.

Multilabel classification on image and text inputs forms a benchmark task in the research on bimodal learning [18,19] and is also used here to assess our proposed approach to multimodal fusion. In creating the MM-IMDb dataset for movie genre classification, the authors were addressing a shortfall in training data to conduct multimodal classification [20]. The task in MM-IMDb is an instance of multilabel prediction over multiple modalities where titles have an average of 2.48 classes and the system undertakes a series of independent classifications. Metrics are computed by comparing outputs Y with target labels from the set D. The authors propose an architecture (referenced below as the *GMU baseline*) with gates to control information learned from modalities to perform classification. We build a version of this system in PyTorch and include results as a benchmark (see Figure 2 and Table 1).

2.2 Dataset

The MM-IMDb dataset constitutes samples for 25,959 movies assigned to one or more of 23 genre classes. Inputs processed in predicting class labels for each sample are a text summary averaging 92.5 words and an image poster. An additional 50 metadata fields of structured text are excluded from the multilabel prediction task to enable assessment of systems on natural language and images.

Systems are evaluated on a processed version of the dataset available from the authors' institution [*]. We extracted four columns from MM-IMDb: FileID, Genres (in one-hot encoded format), VGG16 image embeddings, word2vec text embeddings. Text embeddings are $300-dimension$ word2vec representations and image embeddings are $4096-dimension$ representations of features extracted by Arevalo *et al.* [20]. Image embeddings, text embeddings, and one-hot-encoded true labels are stored as separate tensors ahead of training. Training and cross-validation are performed with 70% of the data and systems are evaluated on the remaining 30% of samples.

[*] `http://lisi1.unal.edu.co/mmimdb/multimodal_imdb.hdf5`

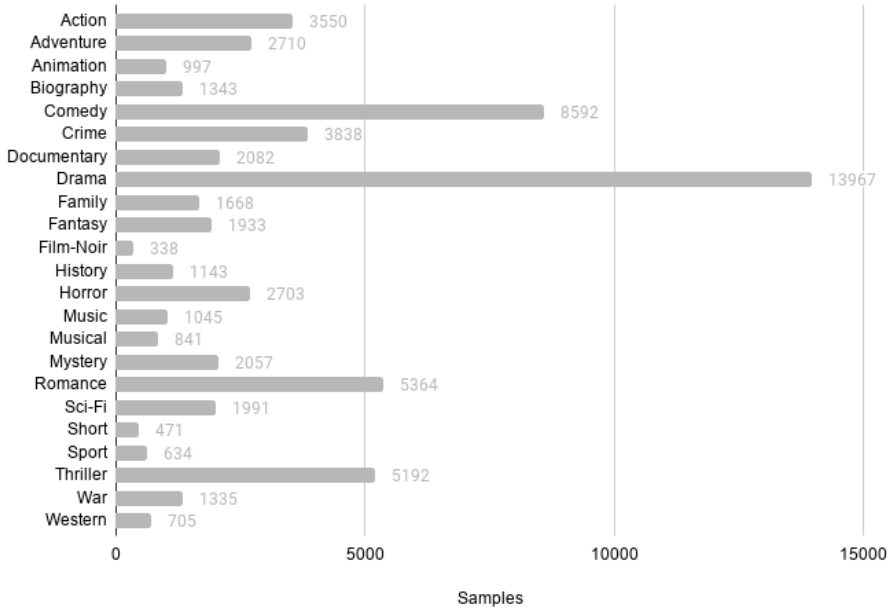

Fig. 1. Number of samples attributed to each genre in the MM-IMDb dataset.

2.3 Variance in Multimodal Classification Systems

High variance is a core challenge to system performance at test time in multimodal classification tasks [21]. Arevalo *et al.* [20] note the improvements that regularisation methods contribute to the architecture proposed for conducting classification on the MM-IMDb dataset. As a starting point, we examine the effects of excluding batchnorm [10] and constraining weight updates to an upper bound (max-norm) [8] on system performance. Validation accuracy curves in Figure 2 run to one epoch after the maximum $weighted - F1$ score (see Section 4) observed for different versions of the baseline system. The negative impact of variance is visible after only a few epochs when these methods are excluded from the system.

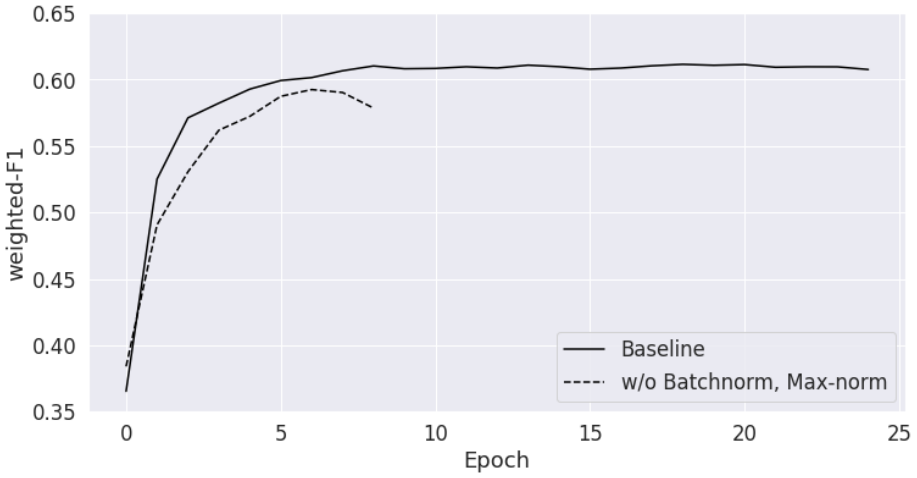

Fig. 2. Performance on the validation set of MM-IMDb by $weighted-F1$ for the GMU baseline system and a version excluding regularisation methods. Curves are plotted to maximum score +1 epoch.

3 Approach

We have examined the importance of regularisation in the use case above and continue with our proposal to provide additional controls for calibrating updates to a set of parameters Θ. In this section, we introduce a classification framework with multimodal fusion comprised of two modules trained with separate loss functions and a single computation of gradients w.r.t. inputs. This framework is the basis for our investigation into mitigating variance with the aim of improving classification performance over multiple modalities and is referred to as *PM+MO* below.

3.1 Classification Framework with Multimodal Fusion

A multilabel classification framework with bimodal inputs is a function $h(x)$ that takes pairs of samples $(x_i^t, x_i^i) = ((x_1^t, x_1^i), (x_2^t, x_2^i), ...(x_n^t, x_n^i))$ where t and i are text and image representations respectively. The resulting multimodal representations are mapped to a subset of labels $S \subseteq D$ or d_i and each output is classified $y = (y_1, y_2, ...y_n) \in \{0, 1\}$. In the proposed framework, the two objectives in $h(f_1(x), f_2(z))$ are computed sequentially with a single step of gradient computations. Module A learns the function $f_1(x)$ for multimodal fusion with a variational inference framework and ELBO as the objective. Module C conducts multilabel classification $f_2(z)$ on the outputs of A by optimising a standard loss described directly below.

Three wide hidden layers are the basis of the fused embedding module A. As with Arevalo *et al.* [20], input embeddings v^t and v^i are assigned to linear

functions and hyperbolic tangent $tanh(u_i^l)$ activations where u is the hyperbolic angle. In our proposal, weights and biases of hidden layers $[w_{i,j}^{I,l}, b_i^l]$ are random variables drawn from a Laplace L distribution $\theta_j^l \sim L(\mu_j, \sigma_j)$ with mean μ_j and variance σ_j optimised during training. Outputs from the unimodal embedding layers are fused using concatenation (1) and mixing (2) operations:

$$v_j^{cat} = [v_j^{t,o}, v_j^{i,o}] \tag{1}$$

$$z_j = v_j^{cat} * v_j^{i,o} + (1 - v_j^{cat}) * v_j^{t,o} \tag{2}$$

Loss on A is computed with stochastic variational inference using the variants of ELBO detailed in Section 3.2. Multimodal embeddings v^m form the inputs for the classifier module C. We align with Arevalo *et al.* [20] in implementing a multilayer perceptron with maxout activation $max_j(w_{i,j}^{I,l}x, b_i^l)$ as proposed by Goodfellow *et al.* [22]. A wide hidden layer receives v^m and the max of parameters for $v_{\lfloor m,o}$] are taken during activation. Binary cross-entropy combines sigmoid activation with cross-entropy loss to assign a probability for each class to outputs $(y_1, y_2, ...y_n) \in \{0, 1\}$ and summing the results $\sum_{c=1}^{M} y_{j,c}, log(p_{j,c})$.

Regularisation methods recommended by Arevalo *et al.* [20] - and retained in our framework - consist of batchnorm to learn $\gamma z + \beta$ for $\mu[z]$ and $[z]$ on batches β, max-norm to constrain weight updates $||w_j^l||$, and dropout with maxout activation in C. Both modules are optimised with variants of the Adam algorithm incorporating regularisation. Adam with gradient clipping in A - as implemented in Pyro PPL [23] - is run on each parameter during steps of variational inference. In the case of C, AdamW was used in place of Adam after initial testing. This algorithm acts on Loschilov and Hutter's proposal to replace L2 regularisation with decoupled weight decay when Adam is the optimiser [24].

3.2 Multimodal Fusion with Variational Inference

Multimodal embeddings are learned by inferring $[w_{i,j}^{I,l}, b_i^l]$ for the layers of A using variational inference. In each case, we assume that the posterior $p(\theta_j^l|X, Y)$ is drawn from a family of Laplace distributions Q. Computing the integrals for p is intractable and so we infer an approximate candidate from Q by selecting q_i with the lowest KL divergence from p [25].The posterior expression is reduced to $p(\theta_j^l|x)$ and defined as:

$$p(\theta_j^l|x) = \frac{p(\theta_j^l, x)}{p(x)}. \tag{3}$$

Minimising the KL divergence

$$KL(q(\theta_j^l)||p(\theta_j^l|x)) \tag{4}$$

is equivalent to maximising the evidence lower bound (ELBO) [26]

$$KL(q(\theta_j^l)||p(\theta_j^l|x)) = -(E_q[logp(\theta_j^l, x)] - E_q[logq(\theta_j^l)] \tag{5}$$

where E_q is the expected value under q. In practice the parameters are stochastic gradients sampled from the optimal variational distribution.

Three variants of ELBO are evaluated in trained versions of the PM+MO framework. The first implementation (ELBOv1) samples s from q_i and computes the expected value in a basic form:

$$E_{q_{s(x)}}[logp(\theta_j^l, x) - logq_s(x)]. \tag{6}$$

ELBOv2 [23] uses the Rao-Blackwellization strategy proposed by Ranganath *et al.* to reduce variance when estimating gradients by replacing random variables with conditional expectations of the variables [27]. The third variant (λKL) also addresses variance in gradient estimation by including a term to limit the regularising influence of the KL term at initiation - and then scaling up the level as training progresses [28]. L1 and L2 norms in the training steps of variational inference present additional controls to regulate parameter updates.

4 Evaluation

An analysis of the PM+MO framework comprises training and testing system variants on the task of multilabel classification on text and images from the MM-IMDb dataset. System performance is measured with $micro-F1$, $macro-F1$, $weighted{-}F1$, and $samples{-}F1$. $F1$ is a standard metric for measuring accuracy in multiclass classification tasks [29] and is computed as the mean of precision and recall

$$f_1^{sample} = \frac{1}{N}\sum_{N}^{i=1}\frac{2|\hat{y_i} \cap y_i|}{|\hat{y_i}| + |y_i|} \tag{7}$$

where N is the number of samples and y_i is the tuple of predictions. Each of the four methods is an average of F-scores computed in the following ways: per sample (samples), across all system outputs (micro), by genre (macro), or by genre and with a weighted average on positive samples for each label. Performance at system level is reported for all of these metrics and $weighted-F1$ is referenced in comparisons between systems in the text.

4.1 System Configuration

Systems in the evaluation are all trained on a single Tesla K80 GPU and with a batch size of 512. Priors for the weights and biases of each layer in builds with variational inference are modeled using Laplace distributions initialised at $\mu = 0.1$ and $\sigma = 0.01$. Parameters for these distributions are learned during gradient steps with the three variants of ELBO detailed above. In the version with KL scaling (λKL), the scaling term is set at $\lambda = 0.2$ following tests in the range $(0.1, 1.0)$. L1 and L2 updates to parameters are set by a different $\lambda = 0.1$ in all tests.

4.2 Results

Experiments aim to measure the impact of training with multiple objectives, variational inference, and probabilistic regularisation methods when conducting multimodal fusion for classification tasks. Assessment starts with a comparison of the best performing version of the PM+MO framework - PM+MO (λKL+$L2$) - with the GMU baseline. An ablation analysis on several PM+MO variants provides a granular analysis of regularisation methods associated with variational inference. Reported numbers are the means of scores calculated over five complete cycles of training and testing.

Table 1. F-scores for PM+MO and GMU Baseline (mean over 5 cycles)

System	F-score			
	Micro	Macro	Weighted	Samples
PM+MO ($\lambda KL + L2$)	0.620	0.549	0.617	0.620
PM+MO (λKL+$L2$+1024)	0.602	0.524	0.599	0.607
GMU Baseline	0.618	0.528	0.608	0.617

Topline results for our proposed framework and the GMU baseline are presented in Table 1. Hyperparameter settings were optimised for each system with differences in learning rate (PM+MO=0.005, GMU=0.001) and dropout (PM+MO=0.9, GMU=0.7). A version of the PM+MO framework with a fused embedding module including ELBO+KL scaling and L2 regularisation in the fused embedding model - and wide layers with 3000 neurons - scored higher on all F-scores than the GMU baseline ($weighted - F1 = +0.009$). The baseline combines a Gated Multimodal Unit and a simple classifier with maxout activation - and is trained end-to-end with a single binary cross-entropy objective. Regularisation and hyperparameter settings conform to specifications shared in the publication [20] and repository. Code conversion from Theano to PyTorch is a contributor to the difference in scores for this build of the GMU system against those stated in the original research ($weighted - F1 = 0.617$).

A version of our framework with 1024 neurons in each linear layer completes the topline analysis. Lower scores for this approach underline the benefits of training with wide layers when regularisation offsets variance. As a final check of the benefits of training with a combination of variational inference and wide layers, we tested a version of the GMU baseline with wide layers and observed lower accuracy to the reported system. Training time per epoch on a single GPU for the best performing PM+MO system is 5.25 secs compared to 1.10 secs for the GMU baseline. Total training time for the former is still low at 2m11s - but extended training times are significant considerations in large-scale data regimes [30].

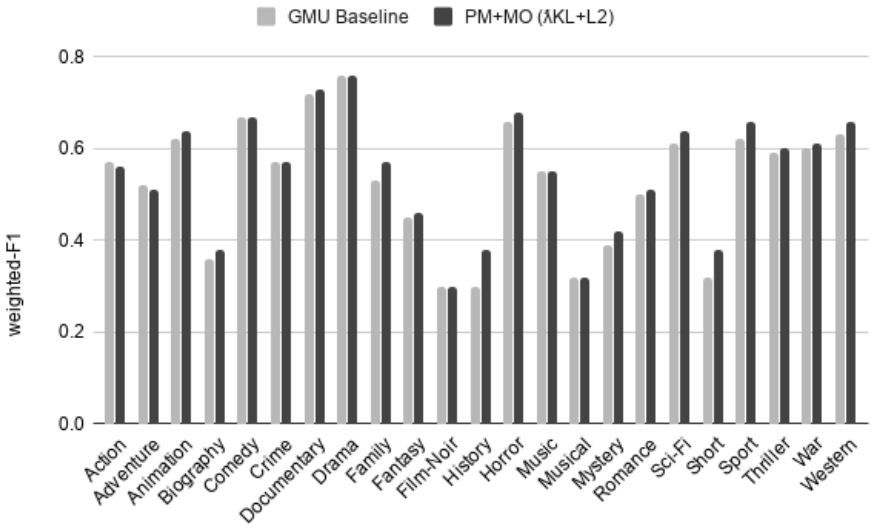

Fig. 3. PM+MO and GMU Baseline performance on genres by $weighted - F1$ (mean over 5 cycles).

Measurements of accuracy on individual classes are presented in Figure 3. The most performant PM+MO system reported higher $weighted - F1$ scores in relation to the GMU baseline for 15 of the 23 movie genres. Classification accuracy matched or exceeded the baseline on all genres where $weighted - F1$ for the latter system was less than 0.5.

Table 2. Ablation for PM+MO with F-scores (mean over 5 cycles)

Specification	F-score			
	Micro	Macro	Weighted	Samples
PM+MO ($\lambda KL + L2$)	0.620	0.549	0.617	0.620
PM+MO ($\lambda KL + L1$)	0.612	0.545	0.613	0.611
PM+MO (λKL)	0.612	0.544	0.611	0.612
PM+MO (ELBOv2+L2)	0.618	0.543	0.614	0.619
PM+MO (ELBOv2)	0.615	0.541	0.611	0.616
PM+MO (ELBOv1+L2)	0.617	0.544	0.614	0.618
PM+MO (ELBOv1)	0.612	0.543	0.609	0.613
M+MO (2 units minus VI)	0.611	0.529	0.602	0.611

Scores for several versions of the PM+MO framework are presented in Table 2 with the objective of comparing the impact of regularisation strategies. Hyper-parameter settings are uniform across all runs with the exception of the specific

methods noted in rows. ELBO versions incorporating methods for managing variance outperform basic implementations of ELBO (ie ELBOv1). KL scaling with L2 regularisation delivers a marginal improvement ($weighted - F1 = +0.003$) on the same configuration with Rao-Blackwellization (ELBOv2+L2). Supplementary L2 norm penalties on parameter updates boost accuracy on all configurations. A build with multiple objectives and excluding variational inference (M+MO) returns the lowest $weighted{-}F1$ (-0.015 w.r.t. PM+MO ($\lambda KL{+}L2$)).

5 Related Work

5.1 Multimodal Representation Learning

Researchers have investigated the role of neural networks in combining representations from multiple modalities to perform end tasks for several decades [31]. Multimodal fusion is deployed in classification tasks when all constituent modalities are present during training and inference [32]. Coordinated embedding methods are an alternative method for these tasks [3]. Separation between vectors is retained in these approaches by projecting the textual and visual representations into a common $d - dimensional$ space and introducing a constraint [33]. Fusion-based learning results in a single output vector: one advantage for sticking with this approach in our framework is to facilitate transfer between modules. Multimodal embeddings are also learned by Silberer and Lapata to perform word similarity and object classification [34]. The process proposed for this and related methods [34,35] differs from the method in our system by including Singular Value Decomposition (SVD) to integrate constituent embeddings.

5.2 Multiple Modules and Objectives

System architectures composed of multiple modules form a foundational area in the research on neural networks. A primary objective in this literature is the construction of classification frameworks that generalise to unseen samples [36]. Auda *et al.* [37] detailed several approaches to decomposing tasks and designed a classifier with multiple modules to solve components for separate sub-tasks. In this case, a voting layer acted as a constraint on outputs from individual components [38]. Early implementations of modular architectures for task decomposition were trained with a single objective - or were separated into distinct models. Secondary losses are implemented during training in the related areas of representation learning and transfer learning. Our system approximates Zhang et al's proposal to introduce an auxiliary objective and module into a classification framework [36]. The method selection in this research differs to ours in implementing unsupervised learning for the auxiliary components. Du *et al.* [39] proposed a system for measuring cosine similarity between auxiliary and main losses when the former contributes to the latter [39]. In contrast to our work, positive transfer with multiple losses are applied in instances where source and target tasks share related objectives.

5.3 Probabilistic Deep Learning

Probabilistic methods in this research extend an approach to machine learning where the assessment of architectures is based on inverse probability [40]. Here the plausibility of the model - or in our case, the parameters in each layer - are computed w.r.t. to the data. Variational inference is a non-deterministic method that replaces elements in probabilistic inference with approximations when the computation of integrals is intractable. A family of distributions is placed over the model parameters and the candidate distribution with the lowest KL divergence from the true posterior is selected [27]. ELBO formulates this minimisation as optimisation and rewards candidate distributions that maximise both $p(z|x)$ and a spread of uncertainty. The ELBO term in our system extends optimisation with simple operations to regularise parameter updates. Ma *et al.* [41] modify ELBO with a supplementary regularisation term to improve representation learning - although the objective of this technique is to reward diversity in the selection of candidate distributions [41]. In contrast to our proposals, enhancements or substitutes for ELBO in this and other research are centred on variational autoencoder (VAE) architectures [42,43]. The selection of Laplace distributions in our system is informed by the ability of these distributions to model data with a high level of heterogeneity [44]. Stochastic variational inference and related methods are implemented in our system using the Pyro PPL [23].

5.4 Regularisation Methods

Our framework retains means for reducing variance when conducting multimodal fusion proposed by Arevalo *et al.* [20]. Batch normalisation was introduced to minimise the impact of changes in parameters on the distributions for activations [10]. Goodfellow *et al.* describe the maxout activation function as an averaging technique in neural network-based architectures that compliments dropout [22]. In initial testing, we verified the performance gain from selecting maxout activation in the classifier module and retained it in the PM+MO architecture. Contributions that describe interactions between parameters and optimiser algorithms [24,45] supported our decision to test different forms of managing weight updates. The context differs as we extend these techniques to training representations with variational inference.

6 Conclusion

In this research, we have demonstrated that a framework of sub-modules trained with variational inference as one of multiple objectives leads to improvements in performance on multimodal classification. The proposed framework supports wide layers and higher learning rates when compared with systems trained with a single objective. Improvements in generalisation over multiple objective systems that exclude variational inference are also demonstrated. An evaluation of

regularisation methods associated with variational inference underlines the advantages of probabilistic approaches in extending options to calibrate parameter updates during training and offset overfitting in multimodal systems. We plan to train on additional modalities and extend probabilistic methods in representation learning as further contributions to the research on improving generalisation in multimodal systems.

Acknowledgements

The project leading to this publication has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 812997.

References

1. Sepulcre, J., Sabuncu, M.R., Yeo, T.B., Liu, H., Johnson, K.A.: Stepwise Connectivity of the Modal Cortex Reveals the Multimodal Organization of the Human Brain. Journal of Neuroscience 32(31), 10649–10661 (Aug 2012), http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.0759-12.2012
2. Bruni, E., Tran, N.K., Baroni, M.: Multimodal Distributional Semantics. Journal of Artificial Intelligence Research p. 47 (2013)
3. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal Machine Learning: A Survey and Taxonomy. arXiv:1705.09406 [cs] (Aug 2017), http://arxiv.org/abs/1705.09406, arXiv: 1705.09406
4. Petkos, G., Papadopoulos, S., Kompatsiaris, I.: Social event detection using multimodal clustering and integrating supervisory signals | Proceedings of the 2nd ACM International Conference on Multimedia Retrieval. In: ICMR '12: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval (Jun 2012), https://dl.acm.org/doi/abs/10.1145/2324796.2324825, archive Location: world Library Catalog: dl.acm.org
5. Ramisa, A.: Multimodal News Article Analysis. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. pp. 5136–5140. International Joint Conferences on Artificial Intelligence Organization, Melbourne, Australia (Aug 2017), https://www.ijcai.org/proceedings/2017/737
6. Yang, X., Ramesh, P., Chitta, R., Madhvanath, S., Bernal, E.A., Luo, J.: Deep Multimodal Representation Learning from Temporal Data. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5066–5074. IEEE, Honolulu, HI (Jul 2017), http://ieeexplore.ieee.org/document/8100021/
7. Amato, G., Behrmann, M., Bimbot, F., Caramiaux, B., Falchi, F., Garcia, A., Geurts, J., Gibert, J., Gravier, G., Holken, H., Koenitz, H., Lefebvre, S., Liutkus, A., Lotte, F., Perkis, A., Redondo, R., Turrin, E., Vieville, T., Vincent, E.: AI in the media and creative industries. arXiv:1905.04175 [cs] (May 2019), http://arxiv.org/abs/1905.04175, arXiv: 1905.04175
8. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting p. 30
9. Wang, W., Tran, D., Feiszli, M.: What Makes Training Multi-Modal Classification Networks Hard? arXiv:1905.12681 [cs] (Dec 2019), http://arxiv.org/abs/1905.12681, arXiv: 1905.12681

10. Ioffe, S., Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv:1502.03167 [cs] (Mar 2015), http://arxiv.org/abs/1502.03167, arXiv: 1502.03167
11. Chen, H.: Machine learning for information retrieval: Neural networks, symbolic learning, and genetic algorithms. Journal of the American Society for Information Science 46(3), 194–216 (1995), https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199504%2946%3A3%3C194%3A%3AAID-ASI4%3E3.0.CO%3B2-S
12. Aggarwal, C.C., Zhai, C.: A Survey of Text Classification Algorithms. In: Aggarwal, C.C., Zhai, C. (eds.) Mining Text Data, pp. 163–222. Springer US, Boston, MA (2012), https://doi.org/10.1007/978-1-4614-3223-4_6
13. Hinton, G.E., Salakhutdinov, R.R.: Reducing the Dimensionality of Data with Neural Networks. Science 313(5786), 504–507 (Jul 2006), https://science.sciencemag.org/content/313/5786/504, publisher: American Association for the Advancement of Science Section: Report
14. Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R.: A SICK cure for the evaluation of compositional distributional semantic models p. 8
15. Xie, N., Lai, F., Doran, D., Kadav, A.: Visual Entailment: A Novel Task for Fine-Grained Image Understanding. arXiv:1901.06706 [cs] (Jan 2019), http://arxiv.org/abs/1901.06706, arXiv: 1901.06706
16. Lehmann, J., Gerber, D., Morsey, M., Ngonga Ngomo, A.C.: DeFacto - Deep Fact Validation. In: Cudré-Mauroux, P., Heflin, J., Sirin, E., Tudorache, T., Euzenat, J., Hauswirth, M., Parreira, J.X., Hendler, J., Schreiber, G., Bernstein, A., Blomqvist, E. (eds.) The Semantic Web – ISWC 2012. pp. 312–327. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg (2012)
17. Nam, J., Kim, Y.B., Mencía, E.L., Park, S., Sarikaya, R., Fürnkranz, J.: Learning Context-DependentLabel Permutations for Multi-Label Classification p. 10
18. Kiela, D., Grave, E., Joulin, A., Mikolov, T.: Efficient Large-Scale Multi-Modal Classification. arXiv:1802.02892 [cs] (Feb 2018), http://arxiv.org/abs/1802.02892, arXiv: 1802.02892
19. Pang, Y., Ma, Z., Yuan, Y., Li, X., Wang, K.: Multimodal learning for multi-label image classification. In: 2011 18th IEEE International Conference on Image Processing. pp. 1797–1800. IEEE, Brussels, Belgium (Sep 2011), http://ieeexplore.ieee.org/document/6115811/
20. Arevalo, J., Solorio, T., Montes-y Gómez, M., González, F.A.: Gated Multimodal Units for Information Fusion. arXiv:1702.01992 [cs, stat] (Feb 2017), http://arxiv.org/abs/1702.01992, arXiv: 1702.01992
21. Radu, V., Tong, C., Bhattacharya, S., Lane, N.D., Mascolo, C., Marina, M.K., Kawsar, F.: Multimodal Deep Learning for Activity and Context Recognition. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1(4), 1–27 (Jan 2018), http://dl.acm.org/citation.cfm?doid=3178157.3161174
22. Goodfellow, I.J., Warde-Farley, D., Mirza, M., Courville, A., Bengio, Y.: Maxout Networks. arXiv:1302.4389 [cs, stat] (Sep 2013), http://arxiv.org/abs/1302.4389, arXiv: 1302.4389
23. Bingham, E., Chen, J.P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., Goodman, N.D.: Pyro: Deep Universal Probabilistic Programming. arXiv:1810.09538 [cs, stat] (Oct 2018), http://arxiv.org/abs/1810.09538, arXiv: 1810.09538
24. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. arXiv:1711.05101 [cs, math] (Jan 2019), http://arxiv.org/abs/1711.05101, arXiv: 1711.05101

25. Wingate, D., Weber, T.: Automated Variational Inference in Probabilistic Programming. arXiv:1301.1299 [cs, stat] (Jan 2013), http://arxiv.org/abs/1301.1299, arXiv: 1301.1299
26. Minka, T.: Divergence Measures and Message Passing (Jan 2005), https://www.microsoft.com/en-us/research/publication/divergence-measures-and-message-passing/
27. Ranganath, R., Gerrish, S., Blei, D.M.: Black Box Variational Inference. arXiv:1401.0118 [cs, stat] (Dec 2013), http://arxiv.org/abs/1401.0118, arXiv: 1401.0118
28. Sønderby, C.K., Raiko, T., Maaløe, L., Sønderby, S.K., Winther, O.: Ladder Variational Autoencoders. arXiv:1602.02282 [cs, stat] (May 2016), http://arxiv.org/abs/1602.02282, arXiv: 1602.02282
29. Madjarov, G., Kocev, D., Gjorgjevikj, D., Dzeroski, S.: An extensive experimental comparison of methods for multi-label learning. Pattern Recognit. 45(9), 3084–3104 (2012), http://dblp.uni-trier.de/db/journals/pr/pr45.html#MadjarovKGD12
30. Ma, M., Pouransari, H., Chao, D., Adya, S., Serrano, S.A., Qin, Y., Gimnicher, D., Walsh, D.: Democratizing Production-Scale Distributed Deep Learning. arXiv:1811.00143 [cs] (Nov 2018), http://arxiv.org/abs/1811.00143, arXiv: 1811.00143
31. Yuhas, B., Goldstein, M., Sejnowski, T.: Integration of acoustic and visual speech signals using neural networks \textbar IEEE Communications Magazine. IEEE Communications Magazine 27 (1989), https://dl.acm.org/doi/10.1109/35.41402
32. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.: Multimodal Deep Learning (2011), https://people.csail.mit.edu/khosla/papers/icml2011_ngiam.pdf
33. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M.A., Mikolov, T.: DeViSE: A Deep Visual-Semantic Embedding Model. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 26, pp. 2121–2129. Curran Associates, Inc. (2013), http://papers.nips.cc/paper/5204-devise-a-deep-visual-semantic-embedding-model.pdf
34. Silberer, C., Lapata, M.: Learning Grounded Meaning Representations with Autoencoders. In: ACL (1). pp. 721–732. The Association for Computer Linguistics (2014), http://dblp.uni-trier.de/db/conf/acl/acl2014-1.html#SilbererL14
35. Bruni, E., Boleda, G., Baroni, M., Tran, N.K.: Distributional Semantics in Technicolor. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 136–145. Association for Computational Linguistics, Jeju Island, Korea (Jul 2012), https://www.aclweb.org/anthology/P12-1015
36. Zhang, Y., Lee, K., Lee, H.: Augmenting Supervised Neural Networks with Unsupervised Objectives for Large-scale Image Classification. arXiv:1606.06582 [cs] (Jun 2016), http://arxiv.org/abs/1606.06582, arXiv: 1606.06582
37. Auda, G., Kamel, M.: Multimodal Deep Learning. Journal of Intelligent and Robotic Systems (1998), https://dl.acm.org/doi/10.1023/A%3A1007925203918
38. Auda, G., Kamel, M., Raafat, H.: A new neural network structure with cooperative modules. In: Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94). vol. 3, pp. 1301–1306 vol.3. Orlando, FL (Jun 1994)
39. Du, Y., Czarnecki, W.M., Jayakumar, S.M., Pascanu, R., Lakshminarayanan, B.: Adapting Auxiliary Losses Using Gradient Similarity. arXiv:1812.02224 [cs, stat] (Dec 2018), http://arxiv.org/abs/1812.02224, arXiv: 1812.02224

40. MacKay, D.J.C.: Bayesian methods for adaptive models. Ph.D. thesis, California Institute of Technology (1992)
41. Ma, X., Zhou, C., Hovy, E.: MAE: Mutual Posterior-Divergence Regularization for Variational AutoEncoders. arXiv:1901.01498 [cs, stat] (Jan 2019), http://arxiv.org/abs/1901.01498, arXiv: 1901.01498
42. Alemi, A.A., Poole, B., Fischer, I., Dillon, J.V., Saurous, R.A., Murphy, K.: Fixing a Broken ELBO. arXiv:1711.00464 [cs, stat] (Feb 2018), http://arxiv.org/abs/1711.00464, arXiv: 1711.00464
43. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. arXiv:1606.03657 [cs, stat] (Jun 2016), http://arxiv.org/abs/1606.03657, arXiv: 1606.03657
44. Geraci, M., Borja, M.C.: Notebook: The Laplace distribution. Significance 15(5), 10–11 (Oct 2018), https://rss.onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2018.01185.x
45. Hanson, S.J., Pratt, L.Y.: Comparing Biases for Minimal Network Construction with Back-Propagation. In: Touretzky, D.S. (ed.) Advances in Neural Information Processing Systems 1, pp. 177–185. Morgan-Kaufmann (1989), http://papers.nips.cc/paper/156-comparing-biases-for-minimal-network-construction-with-back-propagation.pdf