Unsupervised Star Galaxy Classification with Cascade Variational Auto-Encoder

Hao Sun[1,2*], **Jiadong Guo**[2], **Edward J. Kim**[3], **Robert J. Brunner**[3]
[1]Peking University, [2]Peng Cheng Laboratory, [3]The University of Illinois at Urbana Champaign

Abstract

The increasing amount of data in astronomy provides great challenges for machine learning research. Previously, supervised learning methods achieved satisfactory recognition accuracy for the star-galaxy classification task, based on manually labeled data set. In this work, we propose a novel unsupervised approach for the star-galaxy recognition task, namely Cascade Variational Auto-Encoder (CasVAE). Our empirical results show our method outperforms the baseline model in both accuracy and stability.

1 Introduction

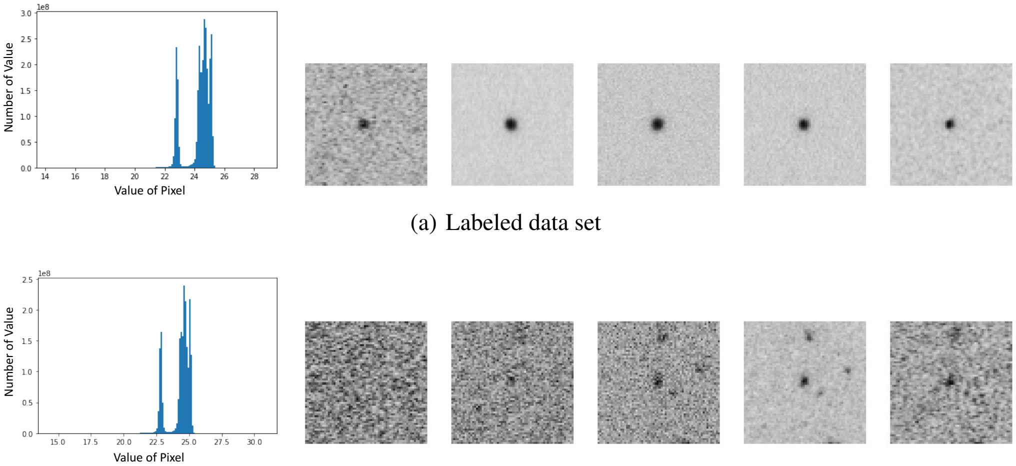
A main challenge in astronomical photometric surveys, e.g. the Sloan Digital Sky Survey (SDSS) [1], the Dark Energy Survey (DES) [2] and the Large Synoptic Survey Telescope (LSST) [3], is the need for object recognition. Their growing scale of collected data in current and future research makes it impossible for human experts to do classification manually. In previous work, machine learning and deep learning techniques were introduced to tackle the challenge [4–9], while all of those methods rely on data sets with labels assigned by experts.

The main drawbacks of previous learning approaches are twofold. On the one hand, labeling astronomical data is a time-consuming, interminable and error-prone job. On the other hand, the label process introduces prior knowledge into the data set, and cognition bias inevitably results. Learning through unsupervised algorithms can be a substitute but we need to extract useful features for the task; and, in most cases, specific network architecture design or domain knowledge is needed to achieve better task specific performance [10–13]. Applying previous unsupervised learning methods into in the domain of astronomy is not straightforward. Specifically, focusing on the star-galaxy classification task, prevailing unsupervised methods: the Auto-Encoders (AEs) and Variational Auto-Encoders (VAEs) [14, 15], are not able to provide satisfactory results for they are designed as generative models, which will learn the most useful hidden representation for identity mapping from the input space to the output space. Other traditional approaches like t-SNE and ISOMAP are not suitable to cope with large scale image inputs [16, 17], thus combination of AEs and manifold learning methods are used [18, 19].

In this work, we propose a method to produce accurate star-galaxy classification by learning directly from the pixel values of photometric images. Our proposed method, Cascade Variational Auto-Encoder, improves the vanilla VAE [15] to be more capable in classification tasks. We demonstrate our method on the SDSS data set that results in remarkable accuracy improvements over baseline method.

[*] Author is now affiliated with CUHK. This work was done in the visiting research intern program in the LCDM group at UIUC

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.



(a) Labeled data set

(b) Unlabeled data set

Figure 1: histograms of pixel values and some image samples

2 Method

In this section we will introduce our proposed solution for the unsupervised star-galaxy classification task. The dataset we used will be introduced in Sec.2.1. Sec.2.2 provides a detailed mathematical formalization of the problem, and proposes an improvement over the prior of vanilla VAEs. We discuss some baseline methods in Sec.2.3, and finally present our solution in Sec.2.4.

2.1 Dataset

The dataset we choose to use in this work is the SDSS dataset. Specifically, we have a labeled subset X_l and an unlabeled subset X_u. The labeled subset X_l contains 140,000 images, each of which has the shape of $64 \times 64 \times 5$, the 5 channels correspond to the u, g, r, i, z bands separately. The ground truth of their class, i.e. stars or galaxies, are also recorded in X_l. Further information of the objects like red-shift z and average magnitude are also recorded but not used in our training or evaluation process. In the unlabeled subset X_u, there are 100,000 images that share the same size with X_l. But we don't have ground truth labels for those images. We aim to train our model with X_u and evaluate our method on X_l. Fig. 1 shows the histograms of pixel values and some image samples of X_l and X_u before normalization.

2.2 Mathematical Formalization

The mathematical formulation of our task is related to the work of Model Agnostic Meta-Learning (MAML) [20] but we concentrate more on unsupervised classification tasks instead of meta-learning tasks. Unsupervised classification tasks can be interpreted as functional optimization problems when we aim to train the model with a labeled dataset and validate on the un-labeled dataset. In such problems, we have an labeled data set X_l, their corresponding labels Y_l and the unlabeled data set X_u. The optimization objective is

$$\min_{f} \parallel Y_l - f(X_l) \parallel^2 \tag{1}$$

where f is a classifier function that maps the original $64 \times 64 \times 5$ input images into a one dimensional scalar, i.e. the predicted label. As we can only learn from unlabeled data in unsupervised cases, we can only utilize the following objective in our optimization:

$$\min_{\mathcal{L}, f} \mathcal{L}(X_u, f(X_u)) \tag{2}$$

where f is the classifier that we will use to test our classification model on X_l, and $\mathcal{L}$ is a surrogate loss function to be determined, which varies between different methods. Ideally, optimization to-

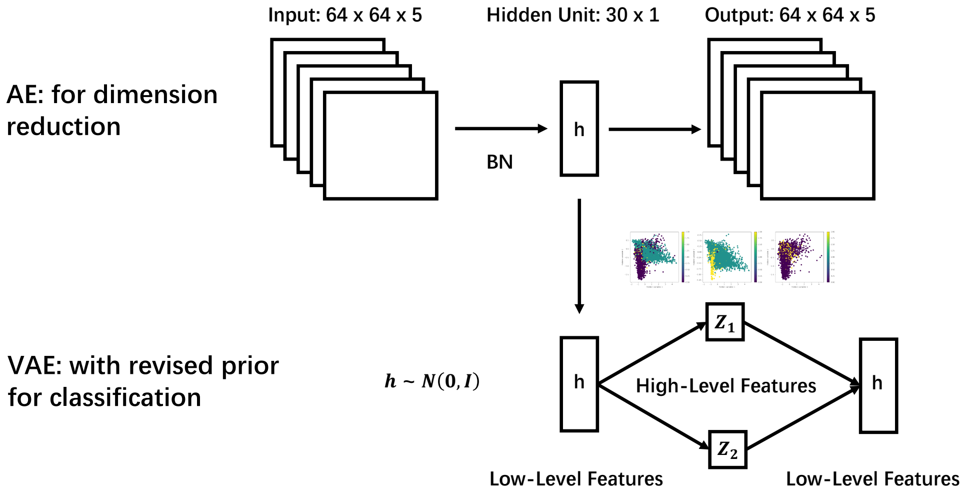


Figure 2: the learning paradigm of our proposed method. Our method can be interpreted as a VAE with its reproduction loss defined by an AE

wards the direction of minimizing the surrogate loss function $\mathcal{L}$ should drive f to be a good classifier automatically.

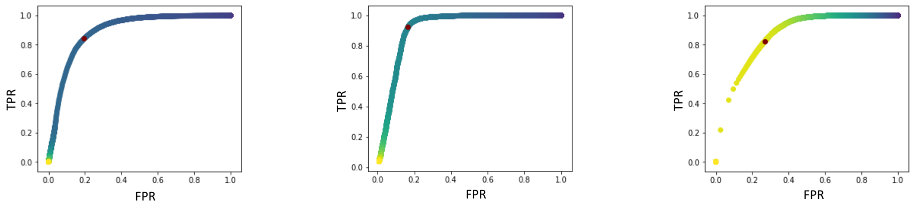
With such formalization, the crux is to find an appropriate surrogate loss function $\mathcal{L}$, with which the optimization goal of Eq.(2) can approximate the goal of Eq.(1). Previous work regard AEs and VAEs as dimension reduction models and utilize reproduction error as the surrogate loss $\mathcal{L}$, during which the bottleneck structure of AEs or VAEs can benefit classification due to its low dimension [21–23]. However, vanilla VAEs take a Gaussian prior in its hidden space to facilitate computation, and enables generating a consistent image by adding perturbations into the hidden variables before putting it to the decoder networks. Noticing the $\mathcal{N}(0, 1)$ prior is not conducive to unsupervised classification, Dilokthanaku et al. put forward GMVAE [24], but their work can not scale up to large scale input images. In our work, as we are interested in separating stars from galaxies, we need to do two-class object recognition, and a better and more convenient choice is to leverage a double-peak Gaussian prior to replace the vanilla Gaussian prior in VAEs. Denoting the prior of hidden space as $P(z)$, we use

$$P(z) \sim \alpha \mathcal{N}(-\mu, \sigma^2) + (1 - \alpha)\mathcal{N}(\mu, \sigma^2) \tag{3}$$

where α is a weighting factor. Although we can choose to use the proportion of stars and galaxies as a prior, we choose to use $\alpha = \frac{1}{2}$ in our experiments to avoid a dependence on prior knowledge.

2.3 Baseline Methods

In unsupervised clustering, AEs are always utilized to reduce the dimension of input data and Manifold Learning (ML) methods (e.g. ISOMAP, t-SNE) are applied to perform further dimension reduction and classification [25, 26, 19, 18]. In detail, those methods have three steps. First, such methods utilize VAEs or AEs to extract features from original images and reduce the dimensions into dim_{hid}, where $dim_{hid} \ll dim_{input}$. Next, they run an ML algorithm to map the d dimensional dim_{hid} into a one dimensional scalar. Finally, the last step is to use the scalar and a threshold, which can be determined from prior knowledge like star-galaxy proportion, to perform classification. In our experiments we will use the method of combining VAEs and ML as baseline.



(a) VAE+ML (b) DKL-VAE + ML (c) CasVAE

Figure 3: ROC and AUC of each method, (a) VAE + ML as baseline; (b) VAE with double-peak Gaussian prior + ML; (c) Cascade VAE. Points with different colors shows different thresholds (dark/blue color shows smaller threshold and light color shows larger threshold). The red points in each figure shows the point closest to the $(0, 1)$ point

Table 1: Performance comparison of each method, VAE + ML is the baseline method and DKL-VAE +ML is based on our improvement of the Gaussian prior used in VAE. Results are averaged by 100 experiments

Method	**Mean AUC**	**Highest AUC**	**Lowest AUC**
VAE + ML	0.75	0.89	0.61
DKL-VAE + ML	0.84	0.92	0.74
CasVAE	**0.90**	**0.91**	**0.90**

2.4 Cascade VAE

Our proposed method, namely Cascade VAE (CasVAE), introduces a hierarchical structure based on VAEs to facilitate unsupervised classification. Based on previous work, most successful generative models often use only a single layer of latent variables [27, 28], and multiple layers only show modest performance increases in quantitative metrics such as log-likelihood [29, 30]. We choose to use shallow neural networks in our model and separating the dimension reduction process and the classification process.

The modelling ability of VAEs varies when using different number of hidden units. While more hidden units lead to sufficient representation ability, less hidden units prompts higher level features, which is more useful for classification [31]. To address this trade-off problem, we proposed a cascade structure. In the first phase of our model, we focus on the dimension reduction part of unsupervised classification, where a normal AE with 30 hidden units is utilized. We apply batch normalization (BN) [32] in AE to bridge the gap between different data distributions. The effect of AE in our model can be interpreted as an low-level feature extractor. In the second phase, we utilize a VAE with 2 hidden units in its hidden layer for classification. The first of the two hidden units follows the vanilla VAE to use a simple Gaussian prior and is used to store information useful for VAE-reproduction. The second hidden unit is equipped with a revised Double-peak Gaussian prior to encourage separation. Fig. 2 shows our architecture, which can be interpreted as a two-channel VAE with its loss function defined by an AE.

3 Experiments

We approximate the calculation of KL-divergence in VAEs in several ways after we introduce the double-peak Gaussian prior in Eq.(3). We evaluate our method based on 4 different kinds of substitutes of KL-divergence. Those loss terms are double-peak KL-Scaling (DKLSC), double-peak KL (DKL), Wasserstein loss (W) and pseudo Wasserstein loss (PW). More details on the calculation can be found in Supplementary materials

In our experiment, we use the Receiver Operating Characteristic curve (ROC) to visualize the classification results and use Area Under Curve (AUC) to evaluate the performance of our models. We use grid search for the selection of hyper-parameters and repeat each experiment with different initializations to test the stability of each method. Fig. 3 shows some of our results, We first tried VAE + Manifold Learning (VAE+ML) and show the ROC curve in Figure Fig. 3(a). Fig. 3(b) shows our results with a double-peak Gaussian prior. Finally, Fig. 3 (c) shows our result when we use the Cascade VAE method (CasVAE).

The first two methods are quite sensitive to the initialization of the neural networks in our experiment. The instability mainly comes from the manifold learning algorithm we use after VAEs. Specifically, Manifold learning algorithms are often sensitive to the outliers. While there are many multi-object images in our dataset, i.e., stars and galaxies may appear in the same image, the prediction we hope to make is on the class of the object in the center of the image. These mixture images may be classified according to the central objects in supervised learning by attention mechanism [33]. But in unsupervised learning and especially in manifold clustering, they are outliers. Our CasVAE method uses one hidden unit as reproducer and one hidden unit as classifier, so that we can use the classifier unit to do classification instead of using ML in other approaches. The performance of each method in terms of AUC is shown in Table 1, indicating the excellence performance of our method as well as its stability.

4 Conclusion

In this work, we proposed a new approach for unsupervised star-galaxy classification, namely Cascade Variational Auto-Encoder (CasVAE). We highlight two improvements over vanilla VAEs. We first introduce an AE in CasVAE to perform dimension reduction, simplifying the problem to a great extent. Moreover, our revision of the Gaussian prior in vanilla VAEs enables CasVAE to abstract high level features that can be used to perform classification in its inner phase. Compared with previous approaches, CasVAE is able to achieve the highest star-galaxy classification accuracy with high stability, which ensures the reproducibility of our proposed method in applications.

Acknowledgement: This work was supported and funded by School of Physics of Peking University (The Undergraduate Student Summer Research Intern Fund). We would like to acknowledge discussions with Binxu Wang, Feng Cheng and Dogezi Song.

References

[1] Jennifer K Adelman-McCarthy, Marcel A Agüeros, Sahar S Allam, Kurt SJ Anderson, Scott F Anderson, James Annis, Neta A Bahcall, Ivan K Baldry, JC Barentine, Andreas Berlind, et al. The fourth data release of the sloan digital sky survey. *The Astrophysical Journal Supplement Series*, 162(1):38, 2006.

[2] University of Chicago Lawrence Berkeley National Laboratory Cerro-Tololo Inter-American Observatory Dark Energy Survey Collaboration: Fermilab, University of Illinois at Urbana-Champaign and Brenna Flaugher. The dark energy survey. *International Journal of Modern Physics A*, 20(14):3121–3123, 2005.

[3] Nicholas Kaiser, Herve Aussel, Barry E Burke, Hans Boesgaard, Ken Chambers, Mark Richard Chun, James N Heasley, Klaus-Werner Hodapp, Bobby Hunt, Robert Jedicke, et al. Pan-starrs: a large synoptic survey telescope array. In *Survey and Other Telescope Technologies and Discoveries*, volume 4836, pages 154–164. International Society for Optics and Photonics, 2002.

[4] Eduardo Vasconcellos, Raíssa Carvalho, R. Gal, F. LaBarbera, Hugo Capelato, Haroldo Campos Velho, Marina Trevisan, and and Ruiz. Decision tree classifiers for star/galaxy separation. *The Astronomical Journal*, 141:189, 05 2011.

[5] Edward J. Kim and Robert J. Brunner. Star–galaxy classification using deep convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*, 464(4):4463–4475, 10 2016.

[6] Ignacio Sevilla-Noarbe and Penélope Etayo-Sotos. Effect of training characteristics on object classification: An application using boosted decision trees. *Astronomy and Computing*, 11:64–72, 2015.

[7] Ross Fadely, David W Hogg, and Beth Willman. Star-galaxy classification in multi-band optical imaging. *The Astrophysical Journal*, 760(1):15, 2012.

[8] Edward J Kim, Robert J Brunner, and Matias Carrasco Kind. A hybrid ensemble learning approach to star–galaxy classification. *Monthly Notices of the Royal Astronomical Society*, 453(1):507–521, 2015.

[9] Harshil M Kamdar, Matthew J Turk, and Robert J Brunner. Machine learning and cosmological simulations–i. semi-analytical models. *Monthly Notices of the Royal Astronomical Society*, 455(1):642–658, 2015.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[11] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[12] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[14] Hervé Bourlard and Yves Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4-5):291–294, 1988.

[15] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[16] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[17] Mukund Balasubramanian and Eric L Schwartz. The isomap algorithm and topological stability. *Science*, 295(5552):7–7, 2002.

[18] Sawon Pratiher and Subhankar Chattoraj. Manifold learning & stacked sparse autoencoder for robust breast cancer classification from histopathological images. *arXiv preprint arXiv:1806.06876*, 2018.

[19] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008.

[20] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

[21] Brain Leke Betechuoh, Tshilidzi Marwala, and Thando Tettey. Autoencoder networks for hiv classification. *Current Science (00113891)*, 91(11), 2006.

[22] Jie Geng, Jianchao Fan, Hongyu Wang, Xiaorui Ma, Baoming Li, and Fuliang Chen. High-resolution sar image classification via deep convolutional autoencoders. *IEEE Geoscience and Remote Sensing Letters*, 12(11):2351–2355, 2015.

[23] Wenjun Sun, Siyu Shao, Rui Zhao, Ruqiang Yan, Xingwu Zhang, and Xuefeng Chen. A sparse auto-encoder-based deep neural network approach for induction motor faults classification. *Measurement*, 89:171–178, 2016.

[24] Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

[25] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. *arXiv preprint arXiv:1611.05148*, 2016.

[26] Leonardo Badino, Claudia Canevari, Luciano Fadiga, and Giorgio Metta. An auto-encoder based approach to unsupervised learning of subword units. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7634–7638. IEEE, 2014.

[27] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[28] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems*, pages 4790–4798, 2016.

[29] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *Advances in neural information processing systems*, pages 3738–3746, 2016.

[30] Philip Bachman. An architecture for deep, hierarchical generative models. In *Advances in Neural Information Processing Systems*, pages 4826–4834, 2016.

[31] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[32] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Appendix

A Detailed calculation of KL-divergence

A.1 An scaling technique

$$
\begin{aligned}
&D_{KL}\left(N(\mu,\sigma^2)\|\frac{1}{2}N(-m,s^2)+\frac{1}{2}N(m,s^2)\right)\\
=&D_{KL}\left(2*\frac{1}{2}N(\mu,\sigma^2)\|\frac{1}{2}N(-m,s^2)+\frac{1}{2}N(m,s^2)\right)\\
\leq&D_{KL}\left(\frac{1}{2}N(\mu,\sigma^2)\|\frac{1}{2}N(-m,s^2)\right)+D_{KL}\left(\frac{1}{2}N(\mu,\sigma^2)\|\frac{1}{2}N(m,s^2)\right)\\
=&-\frac{1}{2}\log 2\left(s^2+\log\sigma^2-\frac{1}{2}(\mu-m)^2-\frac{1}{2}(\mu+m)^2-\sigma^2\right)
\end{aligned}
$$

A.2 Exact solution

$$
\begin{aligned}
&D_{KL}\left(N(\mu,\sigma^2)\|\frac{1}{2}N(-m,s^2)+\frac{1}{2}N(m,s^2)\right)\\
=&\int_{-\infty}^{\infty}N(\mu,\sigma^2)\log\frac{N(\mu,\sigma^2)}{\frac{1}{2}N(-m,s^2)+\frac{1}{2}N(m,s^2)}dx\\
=&\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\log\frac{\frac{1}{\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\frac{1}{2s}[e^{-\frac{(x-m)^2}{2s^2}}+e^{-\frac{(x+m)^2}{2s^2}}]}dx\\
=&\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\log\frac{2s}{\sigma}dx\\
-&\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}[\frac{(x-\mu)^2}{2\sigma^2}-\frac{(x-m)^2}{2s^2}]dx\\
-&\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\log[1+e^{-\frac{2mx}{s^2}}]dx\\
=&\alpha-\beta-\gamma
\end{aligned}
$$

$$
\begin{aligned}
\alpha&=\log\frac{2s}{\sigma}\\
\beta&=-\frac{(m-\mu)^2+\sigma^2-s^2}{2s^2}\\
\gamma&\approx-\frac{2m[-\sigma e^{-\frac{\mu^2}{2\sigma^2}}+\sqrt{\frac{\pi}{2}}\mu Erfc(\frac{\mu}{\sqrt{2}\sigma})]}{s^2}
\end{aligned}
$$

with an approximation of $Erfc(x)\approx 1-tanh(1.19x)$

$$
\begin{aligned}
&D_{KL}\left(N(\mu,\sigma^2)\|\frac{1}{2}N(-m,s^2)+\frac{1}{2}N(m,s^2)\right)\\
\approx&\log\frac{2s}{\sigma}+\frac{(m-\mu)^2+\sigma^2-s^2}{2s^2}\\
+&\frac{2m\{-\sigma e^{-\frac{\mu^2}{2\sigma^2}}+\sqrt{\frac{\pi}{2}}\mu[1-tanh(1.19\frac{\mu}{\sqrt{2}\sigma})]\}}{s^2}
\end{aligned}
$$