DEEPSMOTE: DEEP LEARNING FOR IMBALANCED DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite over two decades of progress, imbalanced data is still considered a significant challenge for contemporary machine learning models. With modern advances and rapid developments in deep learning, countering the problem of imbalanced data has become extremely important. The two main approaches to address this issue are based on loss function modifications and instance resampling, typically based on Generative Adversarial Networks (GANs) that may suffer from mode collapse. Therefore, there is a need for an oversampling method that is specifically tailored to deep learning models, can work on raw images while preserving their properties, and is capable of generating high quality, artificial images that can enhance minority classes and balance the training set. We propose DeepSMOTE - a novel oversampling algorithm for deep learning models. It is simple, yet effective in its design. It consists of only three major components: (i) an encoder/decoder framework; (ii) SMOTE-based oversampling; and (iii) a dedicated loss function enhanced with a penalty term. An important advantage of DeepSMOTE over GAN-based oversampling is that DeepSMOTE does not require a discriminator, and it generates high-quality artificial images that are both information-rich and suitable for visual inspection. DeepSMOTE code is publicly available: https://github.com/dd1github/DeepSMOTE.

1 INTRODUCTION

Learning from imbalanced data is among the crucial problems faced by the machine learning community (Krawczyk, 2016). Skewed distributions affect the training process of any classifier, leading to unfavourable bias towards the majority class(es). This may result in high error, or even complete omission, of the minority class(es). Such a situation cannot be accepted in most real-world applications (e.g., medicine or intrusion detection) and thus algorithms for countering the class imbalance problem have been a focus of intense research for over two decades (Fernández et al., 2018).

Motivation. While the imbalanced data problem adversely affects deep learning models, there has been limited research on how to counter this challenge. Two main directions have been loss function modifications (Cao et al., 2019) and resampling approaches (Bellinger et al., 2020). The deep learning resampling solutions are either pixel-based or use GANs for artificial instance generation (Mullick et al., 2019). Both of these approaches suffer from limitations. Pixel-based solutions often cannot capture complex data properties of images and are not capable of generating meaningful artificial images. GAN-based solutions require significant amounts of data, are difficult to tune, and may suffer from mode collapse. Therefore, there is a need for a novel oversampling method that is specifically tailored to deep learning models, can work on raw images while preserving their properties, and is capable of generating high-quality artificial images.

Summary of contributions. We propose DeepSMOTE - a novel oversampling algorithm for deep learning models based on the highly popular SMOTE method for shallow learning (Chawla et al., 2002). Our method bridges the advantages of metric-based resampling approaches, with a deep architecture capable of working with complex and high-dimensional data. DeepSMOTE consists of three major components: (i) an encoder/decoder framework; (ii) SMOTE-based oversampling; and (iii) a dedicated loss function enhanced with a penalty term. This approach allows us to embed effective SMOTE-based artificial instance generation within a deep encoder / decoder model for a streamlined and end-to-end process, including low dimensional embeddings, artificial image generation, and multi-class classification.

2 Oversampling imbalanced data for deep learning

Oversampling is a proven technique for combating class imbalance for traditional (i.e., shallow) learning models (Fernández et al., 2018). Several attempts have been made to extend oversampling methods, such as SMOTE, to deep learning models, with mixed results (Ando & Huang, 2017; Johnson & Khoshgoftaar, 2019).

Researchers have looked to other avenues within the deep learning ecosystem to replicate the benefits of oversampling with respect to the class imbalance problem. For example, generative models can achieve similar results as oversampling. GANs (Goodfellow et al., 2014), Variational Autoencoders (VAE) (Kingma & Welling, 2013), and Wasserstein Autoencoders (WAE) (Tolstikhin et al., 2017), have been successfully used within computer vision (Karras et al., 2020) and robotic control (Bonatti et al., 2019) to learn the latent distribution of data. Once the underlying distribution is learned, these models then sample from the distribution to produce images or actions.

VAEs operate by maximizing a variational lower bound of the data log-likelihood Hu et al. (2017); Doersch (2016). In deep learning models, the loss function in a VAE is typically implemented by combining a reconstruction loss with the Kullback-Leibler (KL) divergence. The KL divergence can be interpreted as an implicit penalty on the reconstruction loss. By penalizing the reconstruction loss, the model can learn to *vary* its reconstruction of the input data distribution and thus *generate* output (e.g., images) based on a latent distribution of the input. A key advantage of VAEs is that they can be implemented with a single deep learning model and they learn an embedding of the raw input (i.e., they reduce the raw input to a lower dimensional *feature space* so that the mean and variance can be calculated for purposes of the KL divergence). Although VAEs have sound theoretical foundations, they sometimes produce blurry images that may not always resemble the input Mescheder et al. (2017).

WAEs also exhibit generative qualities. Similar to VAEs, the loss function of a WAE is often implemented by combining a reconstruction loss with a penalty term. In the case of a WAE, the penalty term is expressed as the output of a discriminator network. Unlike VAEs, WAEs require a discriminator network and the introduction of a hyper-parameter in the loss function.

GANs have achieved impressive results in the computer vision arena (Wu et al., 2019; Chen et al., 2016). GANs formulate image generation as a min-max game between a generator and a discriminator network (Pfau & Vinyals, 2016). Despite their impressive results, GANs require the use of two networks, are sometimes difficult to train and are subject to mode collapse (i.e., the repetitive generation of similar examples) (Miyato et al., 2018; Salimans et al., 2016; Gulrajani et al., 2017; Arjovsky et al., 2017).

3 DeepSMOTE

We propose DeepSMOTE - a novel and breakthrough oversampling algorithm dedicated to enhancing deep learning models and countering the learning bias caused by imbalanced classes. In order for an oversampling method to be successfully applied to deep learning models, we believe that it should meet three essential criteria:

1. It should operate in an end-to-end manner by accepting raw input, such as images (i.e., similar to VAEs, WAEs and GANs).
2. It should learn a representation of the raw data and embed the data into a lower dimensional *feature space*, which can be used for oversampling.
3. It should readily generate output (e.g., images) that can be visually inspected, without extensive manipulation.

In this paper, we show through our design steps and experimental evaluation that Deep SMOTE meets these criteria. In addition, it produces sharp images without the need for a discriminator network. DeepSMOTE is straight-forward in its implementation. It consists of an encoder / decoder framework, a SMOTE-based oversampling method, and a loss function with a reconstruction loss and a penalty term. Each of these features is discussed below, while the overview of DeepSMOTE is presented in Algorithm 1.

Algorithm 1: DEEPSMOTE

Data: B: batches of imbalanced training data $B = \{b_1, b_2, \ldots, b_n\}$
Input: Model parameters: $\Theta = \{\Theta_0, \Theta_1, \ldots, \Theta_j\}$; Learning Rate: α
Output: Balanced training set.
Train the Encoder / Decoder:
for $e \leftarrow epochs$ **do**
  for $m \leftarrow B$ **do**
    $E_B \leftarrow encode(B)$
    $D_B \leftarrow decode(E_B)$
    $R_L = \frac{1}{n}\sum_{i=1}^{n}\left(D_B i - \hat{D_B i}\right)^2$
    $C_D \leftarrow sample(class\ data)$
    $E_S \leftarrow encode(C_D)$
    $P_E \leftarrow permute - order - of(E_S)$
    $D_P \leftarrow decode(P_E)$
    $P_L = \frac{1}{n}\sum_{i=1}^{n}\left(D_P i - \hat{D_P i}\right)^2$
    $\mathrm{T}_L = R_L + P_L$
    $\Theta_j := \Theta_j - \alpha\frac{\partial T_L}{\partial \Theta_j}\jmath(\Theta_0, \Theta_1)$
Generate Samples:
for $i \leftarrow no.\ of\ minority\ classes$ **do**
  $C \leftarrow select(class\ data)$
  $E \leftarrow encode(C)$
  $G \leftarrow SMOTE(E)$
  $S \leftarrow decode(G)$

Endoder/decoder framework. It is based on the DCGAN architecture established by Radford et al (Radford et al., 2015). Radford et al. employ a discriminator / generator in a GAN, which is fundamentally similar to an encoder / decoder because the discriminator effectively encodes input (absent the final, fully connected layer) and the generator (decoder) produces output. The encoder and decoder are trained in an end-to-end fashion. During DeepSMOTE training, an imbalanced dataset is fed to the encoder / decoder in batches. A reconstruction loss is computed on the batched data. All classes are used during training so that the encoder / decoder can learn to reconstruct both majority and minority class images from the imbalanced data. Because there are few minority class examples, majority class examples are used to train the model to learn the basic reconstruction patterns inherent in the data. This approach is based on the assumption that classes share some similar characteristics (e.g., all classes represent digits or faces).

Enhanced loss function. In addition to a reconstruction loss, the Deep SMOTE loss function contains a penalty term. The penalty term is based on a reconstruction of embedded images. DeepSMOTE's penalty loss is produced in the following fashion. During training, a batch of images is sampled from the training set. The number of sampled images is the same as the number of images used for reconstruction loss purposes; however, unlike the images used during the reconstruction loss phase of training, the sampled images are all from the same class. The sampled images are then reduced to a lower dimensional feature space by the encoder. During the decoding phase, the encoded images are *not* reconstructed by the decoder in the same *order* as the encoded images. By changing the *order* of the reconstructed images, which are all from the same class, we effectively introduce *variance* into the encoding / decoding process. This variance facilitates the generation of images during inference. In addition, by permuting the order of the data (instead of directly using SMOTE in the training phase), we overcome a common issue with metric based learning - the need to access the full training set, which can be computationally challenging in deep learning frameworks. For both reconstruction loss terms, we use mean squared error. DeepSMOTE is trained using gradient descent and the Adam optimizer (Kingma & Ba, 2014). Batch normalization is used to stabilize training (Ioffe & Szegedy, 2015).

Artificial image generation. Once DeepSMOTE is trained, images can be generated with the encoder / decoder structure. The encoder reduces the raw input to a lower dimensional feature space, which is oversampled by SMOTE. The decoder then decodes the SMOTED features into images, which can augment the training set of a deep learning classifier. The main difference between the Deep SMOTE training and inference phases is that during inference, SMOTE is substituted for the order permutation step. SMOTE is used during inference to introduce variance; whereas, during training, variance is introduced by permuting the order of the training examples and also through the penalty loss.

4 EXPERIMENTAL STUDY

Benchmark datasets. Five popular datasets were selected as benchmarks for evaluating imbalanced data oversampling: MNIST (LeCun et al., 1998), Fashion-MNIST (Xiao et al., 2017), CIFAR-10 (Krizhevsky et al., 2009), the Street View House Numbers (SVHN) (Netzer et al., 2011), and Large-scale CelebFaces Attributes (CelebA) (Liu et al., 2015). Imbalance was introduced by random per-class instance selection. For the MNIST and Fashion-MNIST datasets, the class distributions are [4000, 2000, 1000, 750, 500, 350, 200, 100, 60, 40]; for the CIFAR-10 and SVHN datasets they are

[4500, 2000, 1000, 800, 600, 500, 400, 250, 150, 80]; and for CelebA they are [9000, 4500, 1000, 500, 160].

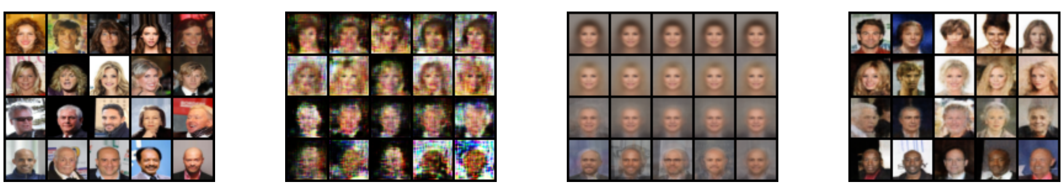
Reference methods. We compared DeepSMOTE to four pixel-based oversampling algorithms: SMOTE, Adaptive Mahalanobis Distance-based Over-sampling (AMDO) (Yang et al., 2018), Combined Cleaning and Resampling (MC-CCR) (Koziarski et al., 2020), and Radial-Based Oversampling (MC-RBO) (Krawczyk et al., 2020); as well as to two GAN-based methods: Balanced GAN (BAGAN) (Mariani et al., 2018) and Generative Adversarial Minority Oversampling (GAMO)(Mullick et al., 2019). All resampling methods were used to create a balance training set for Resnet-18 classifier (He et al., 2016).

Evaluation procedure. A 5-fold cross-validation was used for training and testing the selected algorithms. We evaluated their performance using three skew-insensitive metrics for multi-class imbalanced data: Average Class Specific Accuracy (ACSA), macro-averaged Geometric Mean (GM) and macro-averaged F1 measure (FM) (Fernández et al., 2018).

Table 1: Comparison of DeepSMOTE versus reference resampling method

	MNIST			*FMNIST*			*CIFAR*			*SVHN*			*CELEBA*		
	ACSA	GM	F1	ACSA	GM	F1	ACSA	GM	F1	ACSA	GM	F1	ACSA	GM	F1
SMOTE	81.48	83.99	82.44	67.94	74.84	67.12	28.02	50.08	29.58	70.18	76.33	71.80	60.29	70.48	60.03
AMDO	84.29	88.73	84.88	74.90	80.89	75.39	31.19	53.99	32.44	71.94	78.52	73.06	63.54	72.86	62.94
MC-CCR	86.19	92.04	86.46	78.58	86.17	79.03	32.83	56.68	33.91	72.01	80.94	74.26	65.23	77.14	64.88
MC-RBO	87.25	94.46	88.69	80.06	88.02	80.14	33.01	59.15	35.83	74.20	82.97	74.91	67.11	80.52	65.37
BAGAN	92.56	96.11	93.85	82.50	90.51	82.96	42.41	64.12	43.01	75.81	86.44	77.02	68.62	80.84	**68.33**
GAMO	95.45	97.61	95.11	83.05	90.76	83.00	44.72	65.72	**45.93**	75.07	86.00	76.68	66.06	79.11	64.85
DeepSMOTE	**96.16**	**98.11**	**96.44**	**84.88**	**91.63**	**83.79**	**45.26**	**66.13**	44.86	**79.59**	**88.67**	**80.71**	**72.40**	**82.91**	66.99

Experiment 1: DeepSMOTE vs reference resampling. Table 1 presents the comparison of DeepSMOTE with reference resampling methods. We can see that pixel-based resampling approaches performed much worse than deep learning methods. Only the MC-RBO approach was able to return results not far from GAN-based methods. This shows that pixel-based resampling cannot be used efficiently to train robust deep learning classifiers. GAN-based reference methods performed better, although still out-performed by DeepSMOTE on all three metrics. This shows that DeepSMOTe is at the same time simple and effective, being able to create information-rich artificial instances following topologies of minority classes and capturing their local properties.



(a) Originals (b) BAGAN (c) GAMO (d) DeepSMOTE

Figure 1: CelebA minority class images. Row 1 (brown hair); row 2 (blond hair); row 3 (gray hair); row 4 (bald)

Experiment 2: Quality of artificially generated images. Figure 1 presents the artificially generated images for CelebA dataset by BAGAN, GAMO, and the proposed DeepSMOTE. We can clearly see the quality of DeepSMOTE-generated images. This can be attributed to DeepSMOTE using an efficient encoding/decoding architecture with an enhanced loss function, as well as preserving class topology via metric-based instance imputation. Outcomes of both experiments demonstrates that DeepSMOTE generates artificial images that are both information-rich (improve discrimination abilities of deep classifiers and counter the majority bias) and are of high visual quality.

5 Conclusion

We proposed DeepSMOTE, which marries the simplicity of metric learning with deep architectures that work on complex data. DeepSMOTE works in an end-to-end process, and generates high quality images that can be used for data augmentation and overcoming class imbalance.

REFERENCES

Shin Ando and Chun Yuan Huang. Deep over-sampling framework for classifying imbalanced data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 770–785. Springer, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Colin Bellinger, Roberto Corizzo, and Nathalie Japkowicz. Remix: Calibrated resampling for class imbalance in deep learning. *CoRR*, abs/2012.02312, 2020. URL `https://arxiv.org/abs/2012.02312`.

Rogerio Bonatti, Ratnesh Madaan, Vibhav Vineet, Sebastian Scherer, and Ashish Kapoor. Learning visuomotor policies for aerial navigation using cross-modal representations. *arXiv preprint arXiv:1909.06993*, 2019.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 1565–1576, 2019.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv preprint arXiv:1606.03657*, 2016.

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

Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk, and Francisco Herrera. *Learning from Imbalanced Data Sets*. Springer, 2018. ISBN 978-3-319-98073-7. doi: 10.1007/978-3-319-98074-4. URL `https://doi.org/10.1007/978-3-319-98074-4`.

Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61:863–905, 2018.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Zhiting Hu, Zichao Yang, Ruslan Salakhutdinov, and Eric P Xing. On unifying deep generative models. *arXiv preprint arXiv:1706.00550*, 2017.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Michal Koziarski, Michal Wozniak, and Bartosz Krawczyk. Combined cleaning and resampling algorithm for multi-class imbalanced data with label noise. *Knowl. Based Syst.*, 204:106223, 2020.

Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

Bartosz Krawczyk, Michal Koziarski, and Michal Wozniak. Radial-based oversampling for multiclass imbalanced data classification. *IEEE Trans. Neural Networks Learn. Syst.*, 31(8):2818–2831, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Giovanni Mariani, Florian Scheidegger, Roxana Istrate, Costas Bekas, and Cristiano Malossi. Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655*, 2018.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *International Conference on Machine Learning*, pp. 2391–2400. PMLR, 2017.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Sankha Subhra Mullick, Shounak Datta, and Swagatam Das. Generative adversarial minority oversampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1695–1704, 2019.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

David Pfau and Oriol Vinyals. Connecting generative adversarial networks and actor-critic methods. *arXiv preprint arXiv:1610.01945*, 2016.

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

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*, 2016.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

Yan Wu, Jeff Donahue, David Balduzzi, Karen Simonyan, and Timothy Lillicrap. Logan: Latent optimisation for generative adversarial networks. *arXiv preprint arXiv:1912.00953*, 2019.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Xuebing Yang, Qiuming Kuang, Wensheng Zhang, and Guoping Zhang. AMDO: an over-sampling technique for multi-class imbalanced problems. *IEEE Trans. Knowl. Data Eng.*, 30(9):1672–1685, 2018.