Deep Learning for Classical Japanese Literature

Tarin Clanuwat[*]
Center for Open Data in the Humanities

Mikel Bober-Irizar
Royal Grammar School, Guildford

Asanobu Kitamoto
Center for Open Data in the Humanities

Alex Lamb
MILA, Université de Montréal

Kazuaki Yamamoto
National Institute of Japanese Literature

David Ha
Google Brain

Abstract

Much of machine learning research focuses on producing models which perform well on benchmark tasks, in turn improving our understanding of the challenges associated with those tasks. From the perspective of ML researchers, the content of the task itself is largely irrelevant, and thus there have increasingly been calls for benchmark tasks to more heavily focus on problems which are of social or cultural relevance. In this work, we introduce Kuzushiji-MNIST, a dataset which focuses on *Kuzushiji* (cursive Japanese), as well as two larger, more challenging datasets, Kuzushiji-49 and Kuzushiji-Kanji. Through these datasets, we wish to engage the machine learning community into the world of classical Japanese literature.

1 Introduction

Recorded historical documents give us a peek into the past. We are able to glimpse the world before our time; and see its culture, norms, and values to reflect on our own. Japan has very unique historical pathway. Historically, Japan and its culture was relatively isolated from the West, until the Meiji restoration in 1868 where Japanese leaders reformed its education system to modernize its culture. This caused drastic changes in the Japanese language, writing and printing systems. Due to the modernization of Japanese language in this era, cursive Kuzushiji (くずし字) script is no longer taught in the official school curriculum. Even though Kuzushiji had been used for over 1000 years, most Japanese natives today cannot read books written or published over 150 years ago. [10, 20]

Figure 1: Most Japanese cannot read books over 150 years old, written in cursive *Kuzushiji* style. The 10 classes of Kuzushiji-MNIST, first column showing the modern *Hiragana* counterpart (left). Example of a Kuzushiji literature scroll, *Genjimonogatari Uta Awase*『源氏歌合絵巻』[21] (right).

[*]Corresponding author: tarin@nii.ac.jp, Center for Open Data in the Humanities, Tokyo, Japan.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

Figure 2: The difference between a text printed in 1772 and one printed in 1900. *Onna Daigaku*『女大学』 [13] is a book for women in the Edo period (left). *Shinpen Shūshinkyouten Vol.3*『新編修身教典三巻』 [6] is a textbook right after the standardization of Japanese in 1900 (right).

According to the General Catalog of National Books [19] there have been over 1.7 million books written or published in Japan prior to 1867. In addition to the number of registered books in the national catalog, we estimate that in total there are over 3 million unregistered books and a billion historical documents preserved nationwide. Despite ongoing efforts to create digital copies of these documents—a safeguard against fires, earthquakes, and tsunamis—most of the knowledge, history, and culture contained within these texts remains inaccessible to the general public. While we have many digitized copies of manuscripts and books, only a small number of people with Kuzushiji education are able to read them and work on them, leading to a huge dataset of Japanese cultural works which cannot be read by non-experts.

Figure 3: Our domain transfer experiment, generating Modern Kanji from the Kuzushiji Kanji for unseen characters. (Section 3.2).

In this paper we introduce a dataset specifically made for machine learning research to engage the community to the field of Japanese literature. In this work, we release three easy-to-use preprocessed datasets: Kuzushiji-MNIST, a dataset which focuses on *Kuzushiji* (cursive Japanese), as well as two larger, more challenging datasets, Kuzushiji-49 and Kuzushiji-Kanji. Kuzushiji-MNIST is designed as a drop-in replacement for the MNIST [16] dataset. In addition, we present baseline classification results on Kuzushiji-MNIST and Kuzushiji-49 using recent models, and also apply generative modelling to a domain transfer task between unseen Kuzushiji Kanji and Modern Kanji (See Figure 3). Through these datasets and experiments, we wish to intoducethe machine learning community into the world of classical Japanese literature.[2]

2 Kuzushiji Dataset

The Kuzushiji dataset is created by the National Institute of Japanese Literature (NIJL), and is curated by the Center for Open Data in the Humanities (CODH). In 2014, NIJL and other institutes begun a national project to digitize about 300,000 old Japanese books, transcribing some of them, and sharing them as open data for promoting international collaboration. During the transcription process, a bounding box was created for each character, but literature scholars did not think they were worth sharing. From a machine learning perspective, CODH suggested to make a separate dataset for bounding boxes on a page, because that can be used as the basis for many machine learning challenges and working towards automated transcription. As a result, the *full* Kuzushiji dataset was released in November 2016, and now the dataset contains 3,999 character types and 403,242 characters [5].

[2]Location of dataset with instructions: `https://github.com/rois-codh/kmnist`

Figure 4: In addition to archives, classical books are circulated in manuscript bookstores, online auctions and the annual manuscript auction event held in Jimbocho, Tokyo. Ohya Shobo bookstore in Jimbocho (left). Edo books sold at the Sunday flea market at the Hanazono Shrine, Tokyo (right).

Our hope is that through releasing datasets in familiar formats, we can encourage dialog between the ML and Japanese literature communities. We pre-processed characters scanned from 35 classical books printed in the 18th century and organized the dataset into 3 parts: **(1)** Kuzushiji-MNIST, a drop-in replacement for the MNIST [16] dataset, **(2)** Kuzushiji-49, a much larger, but imbalanced dataset containing 48 Hiragana characters and one Hiragana iteration mark, and **(3)** Kuzushiji-Kanji, an imbalanced dataset of 3832 *Kanji* characters, including rare characters with very few samples.

Hiragana	Unicode	Samples	Sample Images
お (o)	U+304A	7000	
き (ki)	U+304D	7000	
す (su)	U+3059	7000	
つ (tsu)	U+3064	7000	
な (na)	U+306A	7000	

Hiragana	Unicode	Samples	Sample Images
は (ha)	U+306F	7000	
ま (ma)	U+307E	7000	
や (ya)	U+3084	7000	
れ (re)	U+308C	7000	
を (wo)	U+3092	7000	

Figure 5: The 10 classes of Kuzushiji-MNIST. Train and test set sizes are 6,000 and 1,000 per class.

Figure 6: The *hentaigana* for Ka (か) can be written using 12 different root characters (*jibo*, in red) [15], with some of these root characters themselves having multiple ways of being written. Many of the characters in our datasets have multiple ways of being written, so successful models need to be able to model the multi-modal distribution of each class, making the problem more challenging.

Since MNIST restricts us to 10 classes, much fewer than the 49 needed to fully represent Kuzushiji Hiragana, we chose one character to represent each of the 10 rows of Hiragana when creating Kuzushiji-MNIST. One characteristic of classical Japanese which is very different from modern one is that Classical Japanese contains *Hentaigana* (変修仮名). *Hentaigana* or *variant kana*, are Hiragana characters that have more than one form of writing, as they were derived from different Kanji. Therefore, one Hiragana class of Kuzushiji-MNIST or Kuzushiji-49 may have many characters

mapped to it. For instance, as seen in Figure 5, there are 3 different ways to write「つ」because this character was derived from different Kanji (川 and 津).

Another example of this many-to-one mapping is shown in Figure 6. Even though Kuzushiji-MNIST was created as drop-in replacement for the MNIST dataset, the characteristics of Hentaigana and Arabic numbers are completely different, and is one reason why we believe the Kuzushiji-MNIST dataset is more challenging than MNIST.

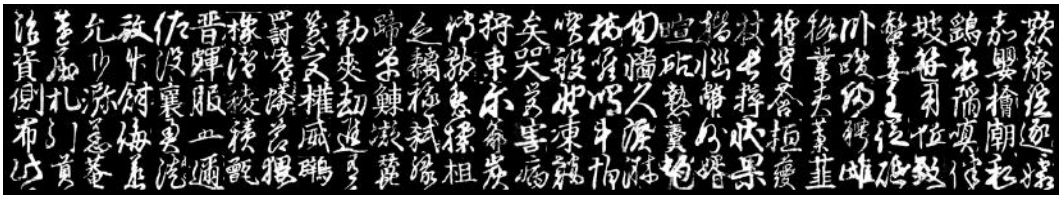

Figure 7: Examples of some of the 3832 classes in Kuzushiji-Kanji.

The high class imbalance in Kuzushiji-49 and Kuzushiji-Kanji is due to the appearance frequency in the real source books, and kept that way to represent the real data distribution. Kuzushiji-49, as the name suggests, has 49 classes (266,407 images) and Kuzushiji-Kanji has a total of 3832 classes (140,426 images), ranging from 1,766 examples to only a single example per class. Kuzushiji-MNIST and Kuzushiji-49 consist of grayscale images of 28x28 pixel resolution, consistent with the MNIST dataset, while the Kuzushiji-Kanji images are of a larger 64x64 pixel resolution.

Hiragana	Unicode	Samples	Sample Images
あ (a)	U+3042	7000	
い (i)	U+3044	7000	
う (u)	U+3046	7000	
え (e)	U+3048	903	
お (o)	U+304A	7000	
か (ka)	U+304B	7000	
き (ki)	U+304D	7000	
く (ku)	U+304F	7000	
け (ke)	U+3051	5481	
こ (ko)	U+3053	7000	
さ (sa)	U+3055	7000	
し (shi)	U+3057	7000	
す (su)	U+3059	7000	
せ (se)	U+305B	4843	
そ (so)	U+305D	4496	
た (ta)	U+305F	7000	
ち (chi)	U+3061	2983	
つ (tsu)	U+3064	7000	
て (te)	U+3066	7000	
と (to)	U+3068	7000	
な (na)	U+306A	7000	
に (ni)	U+306B	7000	
ぬ (nu)	U+306C	2399	
ね (ne)	U+306D	2850	
の (no)	U+306E	7000	

Hiragana	Unicode	Samples	Samples Images
は (ha)	U+306F	7000	
ひ (hi)	U+3072	5968	
ふ (fu)	U+3075	7000	
へ (he)	U+3078	7000	
ほ (ho)	U+307B	2317	
ま (ma)	U+307E	7000	
み (mi)	U+307F	3558	
む (mu)	U+3080	1998	
め (me)	U+3081	3946	
も (mo)	U+3082	7000	
や (ya)	U+3084	7000	
ゆ (yu)	U+3086	1858	
よ (yo)	U+3088	7000	
ら (ra)	U+3089	7000	
り (ri)	U+308A	7000	
る (ru)	U+308B	7000	
れ (re)	U+308C	7000	
ろ (ro)	U+308D	2487	
わ (wa)	U+308F	2787	
ゐ (i)	U+3090	485	
ゑ (e)	U+3091	456	
を (wo)	U+3092	7000	
ん (n)	U+3093	7000	
ゝ (iteration mark)	U+309D	4097	

Figure 8: Kuzushiji-49 description. Training/Test split is $\frac{6}{7}$ and $\frac{1}{7}$ of each class respectively.

In all three datasets, the characters in the train and test sets are sampled from the same 35 books, meaning the data distributions of each class are consistent between the two sets. While Kuzushiji-MNIST is balanced across classes, Kuzushiji-49 has several rare characters with a small number of samples (such as「ゑ」which has only ~ 400 samples).

On the other hand, Kuzushiji-Kanji is a highly imbalanced dataset due to the natural frequency of Kanji appearing in the Kuzushiji literature. In Kuzushiji-Kanji, the number of samples range from

over a thousand to only one sample. This dataset is created for more creative experimental tasks rather than merely for classification and character recognition benchmarks.

Our design of a drop-in replacement for MNIST was inspired by the popular Fashion-MNIST [25], a dataset of fashion items that is considerably more difficult than the original MNIST dataset, while maintaining ease of use. One aspect of Fashion-MNIST that we believe decreases model performance compared to MNIST is that many fashion items, such as shirts, T-shirts, or coats look very similar at 28x28 pixel resolution in grayscale, making many samples ambiguous even for humans (Human performance on Fashion-MNIST is only 83.5% [24]). A characteristic of Kuzushiji-MNIST that makes it more difficult compared to MNIST is that there are in fact multiple very different ways to write certain characters, while each way of writing is still unambiguous at 28x28 pixel resolution for human readers, meaning we believe there is less of a performance 'cap'. Another difference is that while fashion trends come and go, and what constitute a shirt may be different a hundred years from now, Kuzushiji will always remain Kuzushiji. We believe both Fashion-MNIST and Kuzushiji-MNIST will be useful companions to the original MNIST dataset for the research community.

3 Experiments

3.1 Classification Baselines for Kuzushiji-MNIST and Kuzushiji-49

Model	MNIST [16]	Kuzushiji-MNIST	Kuzushiji-49
4-Nearest Neighbour Baseline	97.14%	91.56%	86.01%
Keras Simple CNN Benchmark [4]	99.06%	95.12%	89.25%
PreActResNet-18 [11]	**99.56%**	97.82%	96.64%
PreActResNet-18 + Input Mixup [26]	99.54%	98.41%	97.04%
PreActResNet-18 + Manifold Mixup [22]	99.54%	**98.83%**	**97.33%**

Table 1: Test set accuracy, computed as mean of per-class accuracies to address class imbalance.

We present baseline classification results on Kuzushiji-MNIST and Kuzushiji-49 in Table 1. We consider 4 different baselines: A simple 4-nearest neighbours algorithm, a small 2-layer convolutional network, an 18-layer ResNet [11], and a ResNet that incorporates a manifold mixup regularizer [22]. For the training setup details, please refer to the GitHub repository that contains the dataset. By comparing the performance numbers to the original MNIST dataset using various different approaches, we hope these results will provide a sense of the relative difficulty of our dataset.

3.2 Domain Transfer from Kuzushiji-Kanji to Modern Kanji

In addition to classification, we are interested in more creative uses of our dataset. While existing work [3, 12, 17, 23] on domain transfer focuses on pixel images, we explore instead the transfer from pixel images to *vector* images, across two different domains. Our proposed model aims to generate Modern Kanji versions of a given Kuzushiji-Kanji input, in both pixel and stroke-based formats.

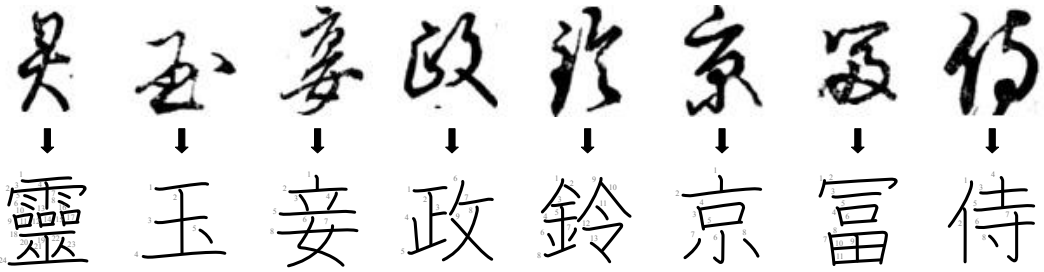

Figure 9: Kuzushiji-Kanji 64x64px samples (Top) and stroke-based Modern Kanji versions (Bottom).

We employ KanjiVG [1], a font for Modern Kanji in a stroke-ordered format. Variational Autoencoders [14, 18] provide a latent space for both Kuzushiji-Kanji and a pixel version of KanjiVG. A Sketch-RNN [8] model is then trained to generate Modern Kanji strokes, conditioned on the VAE's latent space. Predicting pixel versions of Modern Kanji using a VAE also aids human transcribers as the blurry regions of the output can be interpreted as uncertain regions to focus on. In addition to the earlier Figure 3, see Figure 10 below for a demonstration of our model on test set examples.

Figure 10: More domain transfer examples including the VAE pixel reconstructions for both domains.

In Figure 11, we present an overall diagram of our approach. We first train two separate Convolutional Variational Autoencoders, one on the Kuzushiji-Kanji dataset, and also a second on a pixel version of KanjiVG dataset rendered to 64x64 pixel resolution for consistency. The architecture for the VAE is identical to [9] and both datasets are compressed into their own respective 64-dimensional latent space, z_{old} and z_{new}. As in previous work [8], we do not optimize the KL loss term below a certain threshold, ensuring some information capacity while enforcing the Gaussian prior on z.

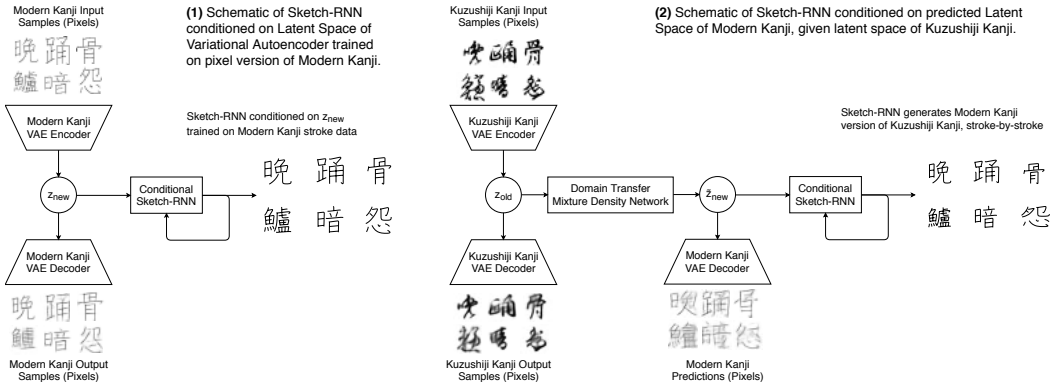

Figure 11: Overview of our approach. **(1)** We first train a VAE on pixel version of KanjiVG (Modern Kanji), and a Sketch-RNN model to generate stroke versions of KanjiVG conditioned on the latent space, z_{new}. **(2)** We train a VAE on Kuzushiji-Kanji, and train a Mixture Density Network [2] to predict $P(z_{\text{new}}|z_{\text{old}})$. We generate stroke versions of Modern Kanji based on the predicted $\widetilde{z}_{\text{new}}$.

Algorithm 1 Summary of training procedure in domain transfer experiment.

1. Train two separate Variational Autoencoders [14, 18] on pixel version of KanjiVG and Kuzushiji-Kanji.
2. Train Mixture Density Network [2] to model $P(z_{\text{new}}|z_{\text{old}})$ as mixture of Gaussians.
3. Train Sketch-RNN [8] to generate KanjiVG strokes conditioned on either z_{new} or $\widetilde{z}_{\text{new}} \sim P(z_{\text{new}}|z_{\text{old}})$.

We then train a Mixture Density Network (MDN) [2] with 2 hidden layers to model the density function of $P(z_{\text{new}}|z_{\text{old}})$ approximated as a mixture of Gaussians. We can then sample a latent vector $\widetilde{z}_{\text{new}}$ in the domain of Modern Kanji, given a latent vector z_{old} encoded from Kuzushiji-Kanji. We note that training two separate VAE models on each dataset is much more efficient and achieves better results compared to training a single model end-to-end, which in our experience does not work well, and might explain why previous works [3, 12, 17, 23] require the use of an adversarial loss.

Previous work [7, 27] utilized MDN-RNN to generate stroke-based Chinese characters. In our last step, we train a Sketch-RNN [8] decoder model to generate Modern Kanji conditioned on $\widetilde{z}_{\text{new}}$. There are around 3,600 overlapping Kanji characters between the two datasets. For characters that are not in Kuzushiji-Kanji, we condition the model on the z_{new} encoded from KanjiVG data to generate the stroke data also from KanjiVG, see (1) in Figure 11. For characters that are in the overlapping 3,600 set, we use the $\widetilde{z}_{\text{new}}$ sampled from the MDN conditioned on z_{old}, to generate the stroke data also from KanjiVG, as per (2) in Figure 11. By doing this, the Sketch-RNN training procedure can fine tune aspects of the VAE's latent space that may not capture well parts of the data distribution of Modern Kanji when trained only on pixels, by training it again on the stroke version of the dataset.

4 Future Directions

We believe the Kuzushiji datasets will not only serve as a benchmark to advance classification algorithms, but also contribute to more creative areas such as generative modelling, adversarial examples, few-shot learning, transfer learning and domain adaptation. To foster community building, we plan to organize machine learning competitions using Kuzushiji datasets to encourage further development of these research areas. We are also working on expanding the size of the dataset, and by next year, the size of the full Kuzushiji dataset will expand to over a million character images. We hope these efforts will encourage further collaboration between different research fields and at the same time, help preserve the cultural knowledge and heritage of Japanese history.

References

[1] U. Apel et al. KanjiVG, 2009. `http://kanjivg.tagaini.net`.

[2] C. M. Bishop. Mixture Density Networks. *Technical Report*, 1994.

[3] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 7, 2017.

[4] F. Chollet et al. Keras, 2015. `https://keras.io`.

[5] C. for Open Data in the Humanities. Kuzushiji dataset, 2016. `http://codh.rois.ac.jp/char-shape/`.

[6] Fukyūsha. *Shinpen Shūshinkyouten Vol.3 (*『新編修身教典 巻三』*)*. Fukyūsha, 1900.

[7] D. Ha. Recurrent Net Dreams Up Fake Chinese Characters in Vector Format with TensorFlow, 2015. `http://otoro.net/kanji-rnn`.

[8] D. Ha and D. Eck. A Neural Representation of Sketch Drawings. In *International Conference on Learning Representations*, 2018. `https://openreview.net/forum?id=Hy6GHpkCW`.

[9] D. Ha and J. Schmidhuber. Recurrent World Models Facilitate Policy Evolution. *arXiv preprint arXiv:1809.01999*, 2018. `https://worldmodels.github.io/`.

[10] Y. Hashimoto, Y. Iikura, Y. Hisada, S. Kang, T. Arisawa, and D. Kobayashi-Better. The Kuzushiji Project: Developing a Mobile Learning Application for Reading Early Modern Japanese Texts. *DHQ: Digital Humanities Quarterly*, 11(1), 2017. `http://dh2016.adho.org/static/data/254.html`.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976. IEEE, 2017.

[13] E. Kaibara. *Onna Daigaku (*『女大学』*)*. Shinsaibashijunkeichō, 1772. `http://www.wul.waseda.ac.jp/kotenseki/html/bunko30/bunko30_g0371/index.html`.

[14] D. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*, 2014.

[15] K. Kodama. *Kuzushiji Yōrei Jiten*. Kondō Shuppansha, 1980.

[16] Y. LeCun. The MNIST database of handwritten digits, 1998. http://yann.lecun.com/exdb/mnist/.

[17] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017.

[18] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *International Conference on Machine Learning*, pages 1278–1286, 2014.

[19] I. Shoten. *General Catalog of National Books (*『国書総巻録』*)*. Iwanami Shoten, 2002.

[20] K. Takashiro. Notation of the Japanese Syllabary seen in the Textbook of the Meiji first Year. *The bulletin of Jissen Women's Junior College*, 34:109–119, mar 2013. https://ci.nii.ac.jp/els/contents110009587135.pdf?id=ART0010042265.

[21] Unknown. *Scroll Genjimonogatari Uta Awase (*『源氏歌合絵巻』*)*. National Institute of Japanese Literature, c. 1500. http://codh.rois.ac.jp/pmjt/book/200014735/.

[22] V. Verma, A. Lamb, C. Beckham, A. Najafi, A. Courville, I. Mitliagkas, and Y. Bengio. Manifold Mixup: Learning Better Representations by Interpolating Hidden States. *ArXiv e-prints*, June 2018.

[23] L. Wolf, Y. Taigman, and A. Polyak. Unsupervised Creation of Parameterized Avatars. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 1539–1547. IEEE, 2017.

[24] H. Xiao et al. Fashion-MNIST: A MNIST-like fashion product database, 2017. https://github.com/zalandoresearch/fashion-mnist.

[25] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[26] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*, 2018.

[27] X. Zhang, F. Yin, Y. Zhang, C. Liu, and Y. Bengio. Drawing and Recognizing Chinese Characters with Recurrent Neural Network. *CoRR*, abs/1606.06539, 2016.