TextCaps : Handwritten Character Recognition with Very Small Datasets

Vinoj Jayasundara, Sandaru Jayasekara, Hirunima Jayasekara, Jathushan Rajasegaran,
Suranga Seneviratne and Ranga Rodrigo

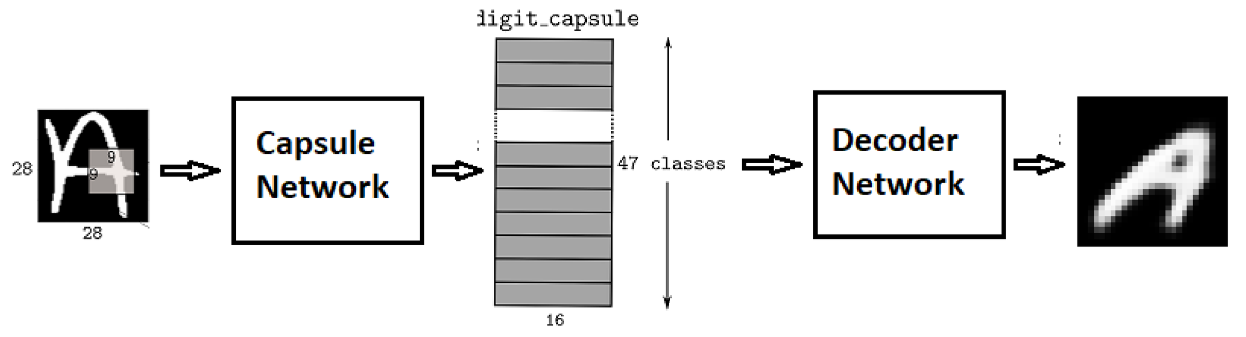
IEEE 2019 **W**inter **C**onference on **A**pplications of **C**omputer **V**ision

MOTIVATION

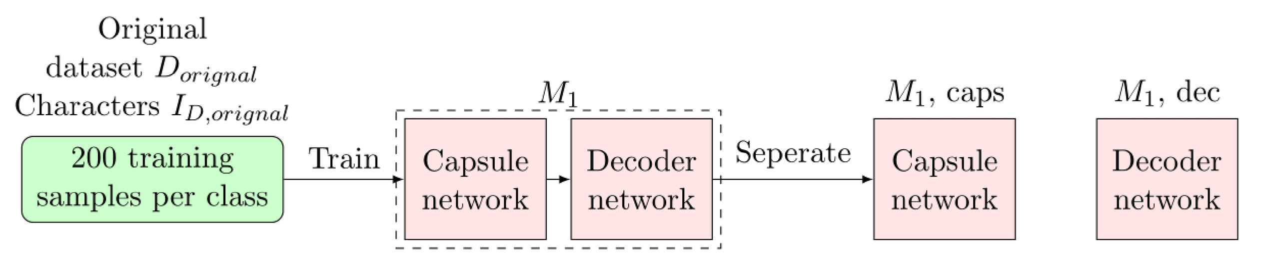
- An inherent problem in deep learning is the necessity of huge datasets for its applications.
- The problem persists in the vision and languages domain as well. Popular languages can reap the benefits of deep learning for multiple tasks including recognition, yet localized languages do not enjoy this privilege due to the lack of huge datasets.
- In this research, we achieved the state-of-the-art performance in classification and reconstruction with very small datasets.

ISSUES CAUSED BY LACK OF DATA

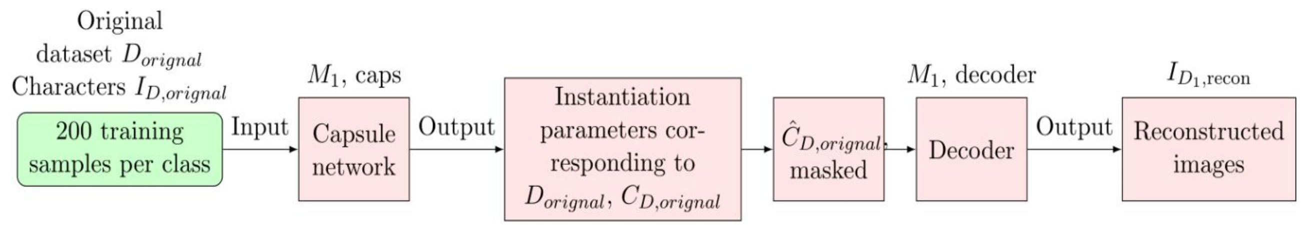
- Reduced versions of well-known datasets, with 200 training samples per class, were used as the small datasets.
- We used Capsule Networks as they excel with limited data and can encode each image with 16 instantiation parameters.



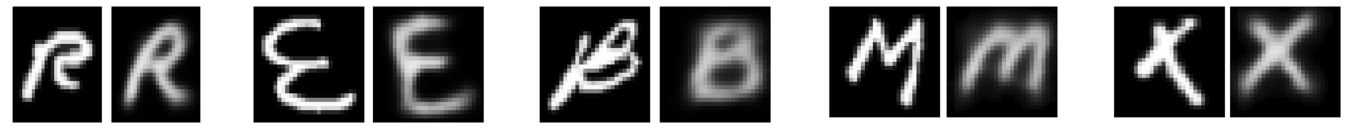
- Training a Capsule Network with very small datasets



- Generating instantiation parameters and reconstructed images



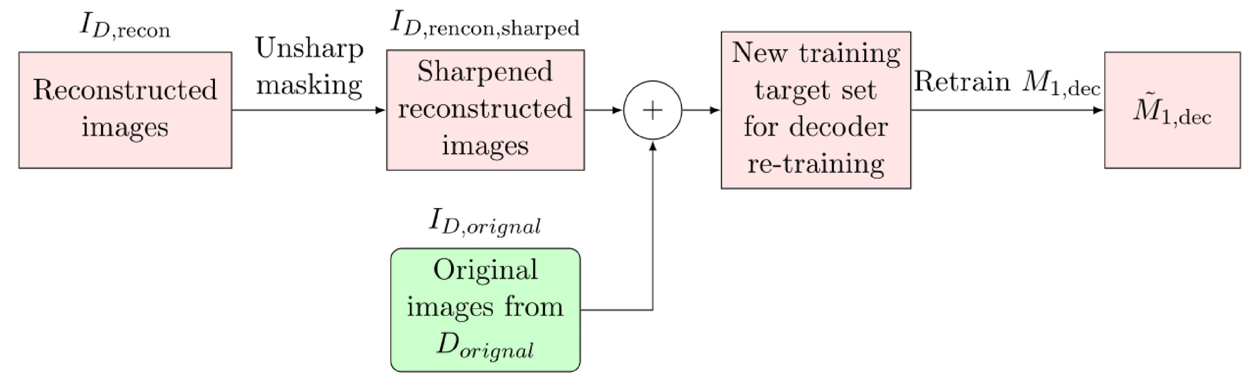
- The decoder network produced the following reconstructions



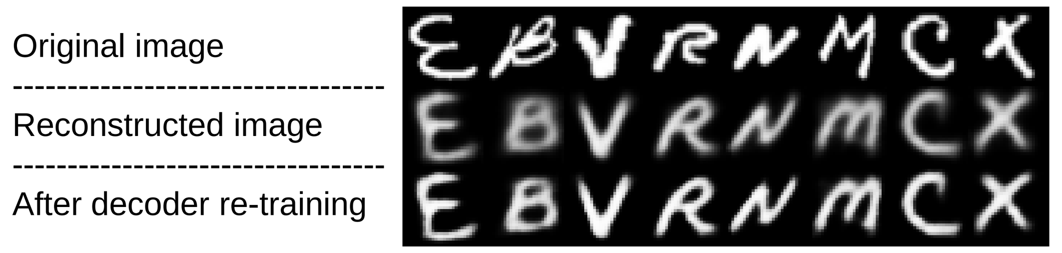
- Two main issues identified
  - Reconstructed images are blurry
  - More importantly, the subtle variations in the characters are not properly captured.

DECODER RETRAINING TECHNIQUE

- We proposed a decoder retraining technique as the solution for reconstructed images being blurry.
- We apply unsharp masking on the reconstructed blurry images, and combine the sharpened images with the original dataset to generate a new training target set for re-training the decoder.



- By using this proposed decoder re-training technique, we were able to give the sharpening ability to the decoder network.



DATA GENERATION INTUITION

- The subtle variations in the characters are not well-captured because the model cannot generalize well with small datasets.
- The most obvious solution is to generate more training data, starting from the existing very small dataset.
- By perturbing the instantiation parameters generated by our model, we can generate human-like variations in characters.

- Uncontrolled perturbations can cause distortions,
  - Visually unrecognizable images

  - Visual class jumps (**H** to **A**)

DATA GENERATION TECHNIQUE

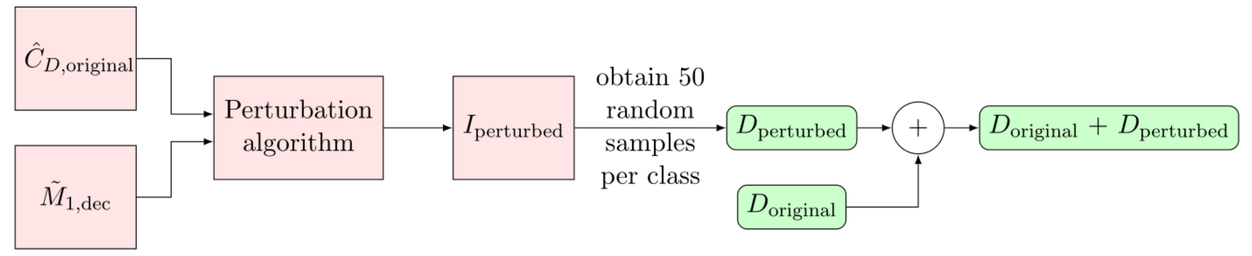
- Our proposed perturbation algorithm automatically generates new data, which avoids distortions without manual inspection.

Algorithm Image data generation using perturbation

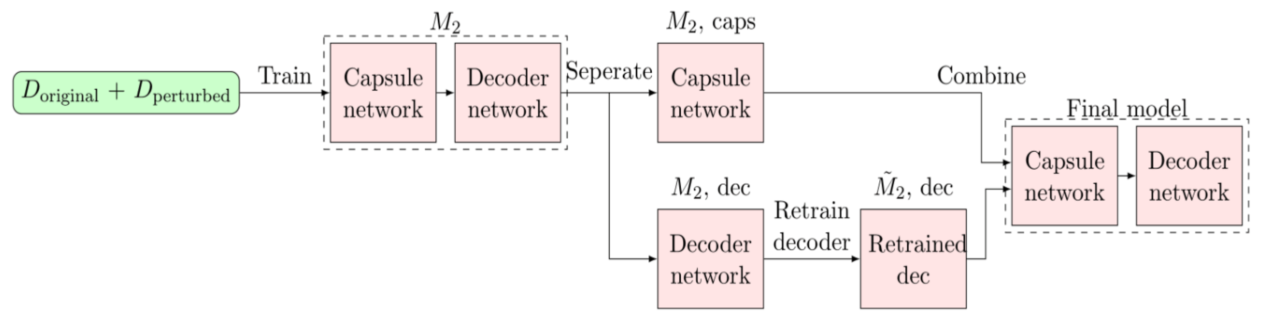
Input: Instantiation parameters $\widehat{C}$, a^{th} highest variance, Decoder Network model ($\widetilde{M}_{dec}$).
Output: Perturbed images $I_{\text{perturbed}}$

Calculate class variance $\sigma_{m,k} = \text{var}_j(\widehat{C}_{m,j,k})$.
Get $\tilde{\sigma}_{m,k'} \leftarrow sort_k(\sigma_{m,k})$ descending.
Get $\hat{k} = k$ corresponding to $k' = a$.
$\tau_{m,k} \leftarrow \frac{\max_j(\widehat{C}_{m,j,k}) - \min_j(\widehat{C}_{m,j,k})}{2}$
get $\tau_k \leftarrow \text{avg}_i(\tau_{m,k})$
for each $\hat{j} \in [j]$ **do**
    if $\widehat{C}_{m,\hat{j},\hat{k}} > 0$ **then**
        $\widehat{C}_{m,\hat{j},\hat{k}} \leftarrow \widehat{C}_{m,\hat{j},\hat{k}} + \min(\tau_{m,\hat{k}}, \tau_{\hat{k}})$
    else
        $\widehat{C}_{m,\hat{j},\hat{k}} \leftarrow \widehat{C}_{m,\hat{j},\hat{k}} - \min(\tau_{m,\hat{k}}, \tau_{\hat{k}})$
$I_{\text{perturbed}} \leftarrow \widetilde{M}_{\text{dec}}(\widehat{C})$

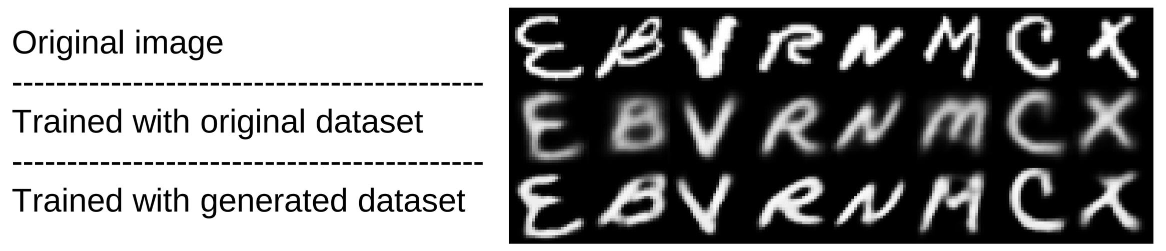
- The newly generated images are combined with the original training set to create a larger dataset.



- Using the new dataset, we can train models to achieve better performance.

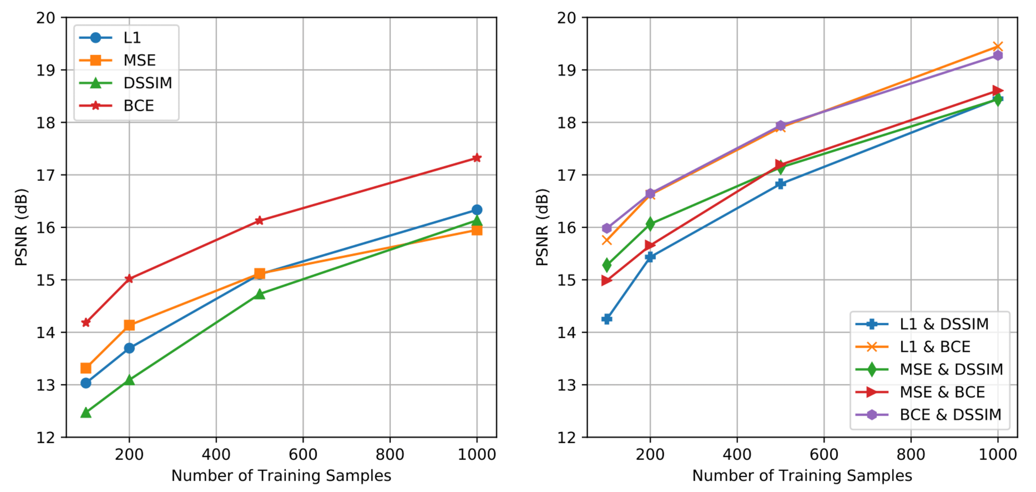


- The reconstructions have significantly improved and the subtle variations in the training images are now well captured.



LOSS FUNCTION ANALYSIS

- When we generate new images using the proposed techniques, the loss functions used in the decoder play an important role.
- A comprehensive analysis on the effect of loss functions on the reconstructed images revealed many interesting observations.



PERFORMANCE COMPARISON

Dataset	Implementation	With full training set	With 200 samp/class
EMNIST Letters	Wiyatno *et al.*	91.27%	-
	TextCaps	**95.36±0.30%**	**92.79±0.30%**
EMNIST Balanced	Dufourq *et al.*	88.30%	-
	TextCaps	**90.46 ± 0.22%**	87.82 ± 0.25%
EMNIST Digits	Dufourq *et al.*	99.30%	-
	TextCaps	**99.79 ± 0.11%**	98.96 ± 0.22%
MNIST	Wan *et al.*	**99.79%**	-
	TextCaps	99.71 ± 0.18%	98.68 ± 0.30%
Fashion MNIST	Zhong *et al.*	**96.35%**	-
	TextCaps	93.71 ± 0.64%	85.36 ± 0.79%

CONCLUSION

- Our data generation technique can produce variations that are closer to human-like variations compared to existing techniques, using only approximately 10% of the data.
- We surpassed the state-of-the-art with full set and achieve the state-of-the-art with the reduced set for multiple datasets.
- Although the proposed techniques were applied on the vision and languages domain in this research, we believe that they are applicable in many other domains as well.